Springer Proceedings in Mathematics & Statistics

Ray Liu
Yi Tsong *Editors*

Pharmaceutical Statistics

MBSW 39, Muncie, Indiana, USA, May 16–18, 2016

Springer

Springer Proceedings in Mathematics & Statistics

Volume 218

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Ray Liu · Yi Tsong
Editors

Pharmaceutical Statistics

MBSW 39, Muncie, Indiana, USA,
May 16–18, 2016

Springer

Editors
Ray Liu
Statistical Innovation
and Consultation Group
Takeda Pharmaceuticals
Cambridge, MA, USA

Yi Tsong
Division of Biometrics VI, CDER
U.S. Food and Drug Administration
Silver Spring, MD, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-67385-1 ISBN 978-3-319-67386-8 (eBook)
https://doi.org/10.1007/978-3-319-67386-8

Library of Congress Control Number: 2019933200

Mathematics Subject Classification (2010): 62-06, 62-07, 62-09, 62H12, 62H15, 62H20, 65C20, 65C60, 92-08, 97M60

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Contributors

Stan Altan Janssen Pharmaceutical R&D, Raritan, NJ, USA

Dean Billheimer Department of Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ, USA

Aili Cheng Pfizer, Pharmaceutical Sciences and Manufacturing Statistics, Andover, MA, USA

Yijie Dong Global Statistics, Bristol-Myers Squibb Co., New Brunswick, NJ, USA

Lingrui Gan Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Feng Gao Biogen, Cambridge, MA, USA

Ping Gao Brightech-International, Somerset, NJ, USA

Pranab Ghosh Cytel Corporation, Cambridge, MA, USA

Cong Han Takeda Pharmaceutical Company Limited, Cambridge, MA, USA

Greg Hather Takeda Pharmaceuticals Inc., Cambridge, MA, USA

Xin Hu The George Washington University, Washington, USA;
ORISE, Oak Ridge, USA

Zhaowei Hua Alnylam Pharmaceuticals, Inc., Cambridge, MA, USA

Ina Jazić Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, Cambridge, MA, USA

Yuan Ji Program of Computational Genomics & Medicine, NorthShore University HealthSystem, Evanston, IL, USA;
Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA

Bonnie LaFleur HTG Molecular Diagnostics, Inc., Tucson, AZ, USA

Dominic LaRoche HTG Molecular Diagnostics, Inc., Tucson, AZ, USA

Fasheng Li Pharmaceutical Science and Manufacturing Statistics, Pfizer Inc., Groton, CT, USA

Qianqiu Li Janssen Research & Development LLC, Spring House, PA, USA

Yunzhi Lin Takeda Pharmaceutical Company Limited, Cambridge, MA, USA

Kevin Liu Janssen Research & Development, Raritan, NJ, USA

Lingyun Liu Cytel Corporation, Cambridge, MA, USA

Ray Liu Takeda Pharmaceuticals Inc., Cambridge, MA, USA

Yi Liu Takeda Pharmaceuticals, Cambridge, MA, USA

Richard Lostritto Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, USA

Areti Manola Janssen Pharmaceutical R&D, Raritan, NJ, USA

Cyrus Mehta Cytel Corporation, Cambridge, MA, USA;
Harvard School of Public Health, Boston, MA, USA

Kurt Michels HTG Molecular Diagnostics, Inc., Tucson, AZ, USA

Peter Müller Department of Mathematics, The University of Texas at Austin, Austin, TX, USA

Yang Ni Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX, USA

Steven Novick Medimmune, Gaithersburg, MD, USA

Gene Pennello Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD, USA

Bill Pikounis Janssen Research & Development LLC, Spring House, PA, USA

Meiyu Shen Division Biometrics VI, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA

Jyh-Ming Shoung Janssen Pharmaceutical R&D, Raritan, NJ, USA

Max Shpak Sarah Cannon Research Institute, Nashville, TN, USA;
Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, TX, USA;
Fresh Pond Research Institute, Cambridge, MA, USA

Jiayang Sun Department of Statistics, Case Western Reserve University, Cleveland, OH, USA

Szu-Yu Tang Ventana Medical Systems, Inc., Tucson, AZ, USA

Neal Thomas Pfizer, Statistical Research and Consulting Center, Groton, CT, USA

Hong Tian Janssen Research & Development, Raritan, NJ, USA

Yi Tsong Division of Biometrics VI, Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA

Saran Vardhanabhuti Takeda Pharmaceuticals, Cambridge, MA, USA

Ke Wang Pharmaceutical Science and Manufacturing Statistics, Pfizer Inc., Groton, CT, USA

Tianhua Wang Office of Biostatistics/Office of Translational Science, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA

Iris Yan Global Statistics, Bristol-Myers Squibb Co., New Brunswick, NJ, USA

Miao Yang Department of Statistics, Oregon State University, Corvallis, OR, USA

Yarong Yang Department of Statistics, North Dakota State University, Fargo, ND, USA

Jingjing Ye Division of Biometrics V, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

Pingye Zhang Merck, Rahway, NJ, USA

Weidong Zhang Pfizer Inc., Cambridge, MA, USA

Wencan Zhang Takeda Develop Center, Deerfield, IL, USA

Zhongfa Zhang Department of Statistics, Case Western Reserve University, Cleveland, OH, USA

Jian Zhu Global Statistics, Takeda Pharmaceutical Company Limited, Tokyo, Japan

Yonghong Zhu Shanghai Henlius Biotech Inc., Shanghai, China

Part I
Specification and Sampling Acceptance Tests

Statistical Considerations in Setting Quality Specification Limits Using Quality Data

Yi Tsong, Tianhua Wang and Xin Hu

Abstract According to ICH Q6A (Specifications: test procedures and acceptance criteria for new drug substances and new drug procedures: chemical substances, (1999) [5]) Guidance, a specification is defined as a list of tests, references to analytical procedures, and appropriate acceptance criteria, which are numerical limits, ranges, or other criteria for the tests described. They are usually proposed by the manufacturers, and subject to the regulatory approval for use. When the acceptance criteria in product specifications cannot be pre-defined based on prior knowledge, the conventional approach is to use data of clinical batches collected during the clinical development phases. This interval may be revised with the accumulated data collected from released batches after drug approval. Dong et al. (J Biopharm Stat 25:317–327, 2015 [1]) discussed the statistical properties of the commonly used intervals and made some recommendations. However, in reviewing the proposed intervals, it is often difficult for the regulatory scientists to understand the difference between the intervals, when some intervals require only pre-specified target proportion of the distribution, and others require confidence level, in addition. Therefore, we propose to use the same confidence level of 95%, and calibrate each interval to the true coverage, under the tolerance interval setting. It is easy to show that the predictive interval and reference interval has the variable true coverage, and increases with the sample size, while tolerance interval covers the fixed true coverage. Based

Y. Tsong—The project is completed as part of the requirement of 2017 OB/ORISE summer intern program.

*Disclaimer: This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

Y. Tsong (✉) · T. Wang
Division of Biometrics VI, Office of Biostatistics, CDER, FDA, 10903 New Hampshire Ave, Silver Spring, MD 20903, USA
e-mail: yi.tsong@fda.hhs.gov

X. Hu
The George Washington University, Washington, USA

ORISE, Oak Ridge, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_1

on our study results, we propose somesome appropriate statistical methods, in setting product specifications, to better ensure the product quality for the regulation purpose.

Keywords Specification · Prediction interval · Reference interval · Tolerance interval · Coverage

1 Introduction

The two key documents on specifications are ICH Q6A Guidance which covers the specifications of chemical products and ICH Q6B Guidance which covers the specification of biological products. In ICH Q6A and ICH Q6B, the specification is defined as:

> A list of tests, references to analytical procedures acceptance criteria, which are numerical limits, ranges, or other criteria for the tests described. It establishes the set of criteria, to which a drug substance or drug product conforms, to be considered acceptable for the intended use. 'Conformance to specification' means that the drug substance and/or drug product, when tested according to the listed analytical procedures, will meet the listed acceptance criteria. Specifications are critical quality standards that are proposed and justified by the manufacturers, and approved by the regulatory authorities as condition of approval.

For a drug product, specifications may be required for potency assay, impurities, pH, dissolution, water content and microbial limits, depending on the dosage form DiFeo, DDIP [6]. Statistical involvement in specifications would be in the determination of the acceptance criteria. In 2015, Dong et al. [1] discussed some statistical methods used or proposed in setting up the criteria based on data collected. In this paper, we extend the discussion about the properties of the methods in pre-marketing determination. We also discuss the cases when the post-marketing revision is applicable.

2 Statistical Methods for Setting Specification Criteria

In 2015, Dong et al. [1] discussed the statistical properties of the intervals (including reference interval, tolerance interval and Min-Max interval) often used or proposed for setting the specification limits. In this article, we extend the discussion and comparison of the interval methods. In addition, we also include predictive interval proposed by Geisser [4] in the discussion. In this article, we will restrict the quality measurement to normally distributed random variable, and consider the two-sided specification limits including both the lower and upper limits. The tolerance interval will be restricted to the intersection of two one-sided intervals. As discussed in Dong et al. [1], Min-Max interval is defined to target on asymptotic 100% of the distribution and is very different from the other three intervals. Thus we will not include it in the discussion of this article.

We use the same symbols and notations as described in Dong et al. [1], such that the testing results $X_1, X_2, \ldots, X_n$ are i.i.d. random samples from a normal distribution with mean μ and variance σ^2. The desired specification is formulated as an interval, covering a certain proportion of the population, say 100p% , with $0 < p < 1$; then the symmetric interval $\left(\mu - Z_{(1+p)/2}\sigma, \mu + Z_{(1+p)/2}\sigma\right)$ with $Z_{(1+p)/2}$ being the $(1 + p)/2$ percentile of the standard normal distribution, is the shortest interval to satisfy the requirement. Since μ and σ are to be estimated from the sample, this interval $\left(\mu - Z_{(1+p)/2}\sigma, \mu + Z_{(1+p)/2}\sigma\right)$ needs to be estimated using the available data. Some commonly used methods are predictive interval (PI), reference interval (RI) and tolerance interval (TI). When the mean and variance are unknown, all three intervals are represented in the form of $\left(\overline{X} - kS, \overline{X} + kS\right)$, with k defined differently as shown below for the three intervals.

- Predictive interval (PI): with unknown mean and variance, it is formulated as

$$\left(\overline{X} - t_{(1+p)/2}.S\sqrt{1 + 1/n}, \overline{X} + t_{(1+p)/2}.S\sqrt{1 + 1/n}\right) \tag{1}$$

with $t_{(1+p)/2}$ being the $100(1 + p)/2$th percentile of t-distribution with n–1 degrees of freedom. In other words, PI is $\left(\overline{X} - kS, \overline{X} + kS\right)$ with $k = k_1 = t_{(1+p)/2}.\sqrt{1 + 1/n}$. It assures with probability p that the next observed value X_{n+1} will fall within the PI. PI gives an interval with the coverage p of the next observed value X_{n+1}. But there is no degrees of assurance how frequently it will happen.

- Reference interval (RI) EMA [2]: it is formulated as

$$\left(\overline{X} - Z_{(1+p)/2}.S, \overline{X} + Z_{(1+p)/2}.S\right) \tag{2}$$

with $k = k_2 = Z_{(1+p)/2}$ in $\left(\overline{X} - k.S, \overline{X} + k.S\right)$ for a pre-specified coverage p. RI is a point estimate of $\left(\mu - Z_{(1+p)/2}.\sigma, \mu + Z_{(1+p)/2}.\sigma\right)$. Therefore, its asymptotic coverage is p. Its actual coverage is a function of sample size and confidence level, as we will discuss in Sect. 4.

- Tolerance Interval (TI), Faulkenberry [3]: tolerance interval actually provides a $1 - \alpha$ confidence that the interval estimated covers a fraction p of the normal distribution. An exact form of the interval can be constructed with two one-sided TIs formulated by

$$\left(\overline{X} - \frac{t_{1-\alpha/2}\left(n - 1, Z_{(1+p)/2}\sqrt{n}\right)}{\sqrt{n}}.S, \overline{X} + \frac{t_{1-\alpha/2}\left(n - 1, Z_{(1+p)/2}\sqrt{n}\right)}{\sqrt{n}}.S\right) \tag{3}$$

with $k = k_3 = \frac{t_{1-\alpha/2}\left(n-1, Z_{(1+p)/2}\sqrt{n}\right)}{\sqrt{n}}$, where $t_{1-\alpha/2}\left(n - 1, Z_{(1+p)/2}\sqrt{n}\right)$ is the $100(1 - \alpha/2)$th percentile of non-central t distribution with the degrees of freedom n–1 and non-central parameter $Z_{(1+p)/2}\sqrt{n}$.

From the three intervals introduced above, if we let $U \sim N(0, 1)$, $V \sim \chi^2_{n-1}$, and let the parameter $W = (U + \delta)/\sqrt{V/(n-1)} \sim t_{n-1}(\delta)$, then it is clear that $V/(n-1) \to 1$ and $U/\sqrt{n} \to 0$ in probability. For $\delta/\sqrt{n} = Z_{(1+p)/2}$, we have $W/\sqrt{n} = 1/\sqrt{V/(n-1)}\left(U/\sqrt{n} + \delta/\sqrt{n}\right) \to Z_{(1+p)/2}$ in probability. Thus, we have both k_1 and k_3 converge to $k_2 = Z_{(1+p)/2}$ as $n \to \infty$.

3 Relationships Between Coverage and k of PI, RI and TI

Dong et al. [1] discussed the statistical properties, in terms of the coverage and width of the resulting interval estimated for a specified coverage p, using the (Min, Max), reference interval, tolerance interval and confidence limit of the percentiles. In general, except (Min, Max), all methods discussed converge to the asymptotic coverage p unbiasedly. Dong et al. [1] also pointed out that both PI and RI are point estimates with a true coverage much smaller than the pre-specified p. However, for any given sample size n and k, we may derive the true coverage p^* with a pre-specified confidence level 1–α. This can be easily done by solving p^* in the equation below

$$k = \frac{t_{1-\alpha/2}\left(n-1, Z_{(1+p^*)/2}\sqrt{n}\right)}{\sqrt{n}} \tag{4}$$

For example, let us consider $k = 3$ used in RI and correspondingly, the asymptotic result of PI and TI is $(\mu - 3\sigma, \mu + 3\sigma)$, with the coverage of 99.73%. For a sample size of 21 observations, with a 95% confidence level, the true coverage p^* of PI can be derived by solving p^* in $t_{(1+0.9973)/2}\sqrt{1 + 1/21} = \frac{t_{0.9975}\left(20, Z_{(1+p^*)/2}\sqrt{21}\right)}{\sqrt{21}}$ and the true coverage p^* of the three standard deviation RI $\left(\overline{X} - 3.S, \overline{X} + 3.S\right)$ can be derived by solving p^* in $\frac{t_{0.9975}\left(20, Z_{(1+p^*)/2}\sqrt{21}\right)}{\sqrt{21}} = 3$.

Table 1 summarizes the TI based coverage p^* (fixed confidence level 95%), calculated from the 3-standard deviation RI $\left(\overline{X} - k.S, \overline{X} + k.S\right)$, and the PI with asymptotic coverage $p = 0.9973$ under different sample sizes. It is clear to see that the TI-based coverage p^*, from the 3-standard deviation RI $\left(\overline{X} - k.S, \overline{X} + k.S\right)$ increases when sample size n increases, and it decreases when sample size n decreases. On the other hand, for the same asymptotic coverage (for example $p = 0.9973$) of the RI, the coefficient k of PI and TI can also be derived for each given sample size n (Fig. 1) when the confidence level of the TI is fixed to be 95%. Similarly, for the same asymptotic coverage 0.9973, the k values of PI, RI and TI, for each sample size n can be calculated. As shown in Fig. 1, k values of PI and TI are large when n is small, but they gradually reduced to 3, asymptotically. To illustrate the relative size of the interval, sixty one data points generated from $X \sim N(100, 4)$ are used to calculate PI, RI and TI at every increase sample size. The intervals are shown in Fig. 2.

Table 1 The TI based Coverage p* calculated using Reference Interval $\left(\overline{X} - 3S, \overline{X} + 3S\right)$ and Prediction Interval with asymptotic coverage p = 0.9973 under different sample sizes

Sample Size N	Equivalent coverage probability p^* in the $(100 \times p^*\%$ Coverage$)/(95\%$ Confidence) Two One-Sided tolerance interval	
	Reference Interval (%)	Prediction Interval (%)
N = 10	86.44	97.51
N = 15	92.51	97.75
N = 20	94.88	98.01
N = 25	96.10	98.22
N = 30	96.82	98.38
N = 100	98.88	99.12
N = 200	99.24	99.33

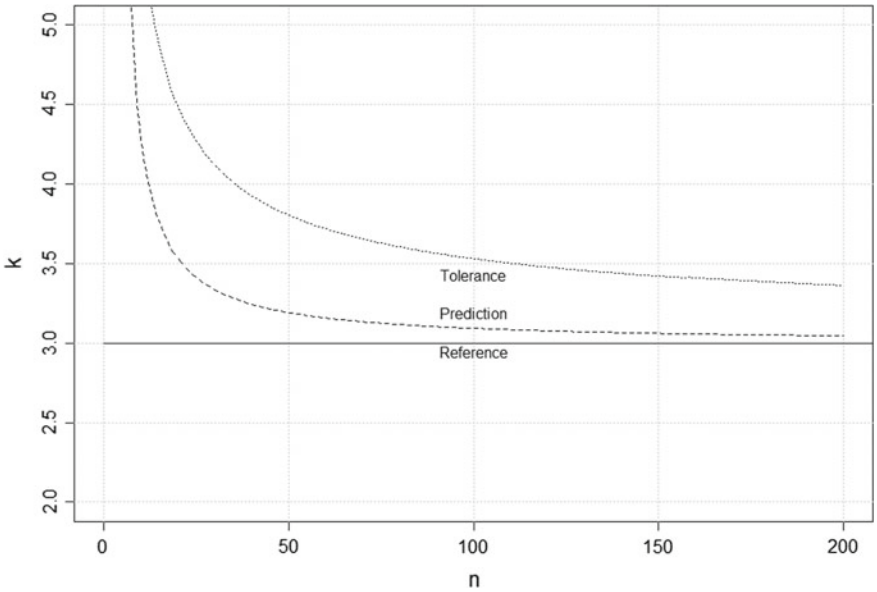


Fig. 1 Coefficient k for PI, RI and TI against sample size n. The asymptotic coverage for the RI, the pre-specified coverage p for the PI, and the coverage for the TI are all 0.9973, the confidence level for the TI is 95%

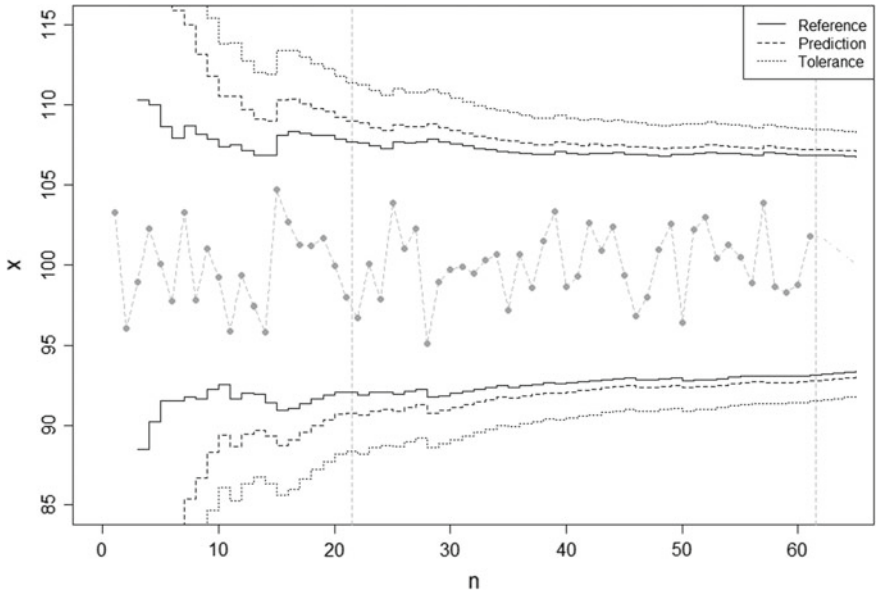


Fig. 2 Simulated Interval of RI, PI and TI against n; data points are simulated from the normal distribution $X \sim N(100, 4)$. The asymptotic coverage for the RI, the pre-specified coverage p for the PI, and the coverage for the TI are all 0.9973, the confidence level for the TI is 95%

Often, the specification limits need to be set with a relatively small sample size before marketing. There is potential that the limits need to be revised when data cumulated post-marketing. In this paper, we will focus on the properties of the specification limits estimated, using the predictive internal (PI), reference interval (RI) and tolerance interval (TI) with the sample size changes.

4 Specification Determined with Pre-marketing Data

Before marketing, the data used for specification determination often consists of only the observed initial and early stability values from the clinical lots subject to phase III clinical evaluation. The sample size consists often of stability values observed at 0, 3, 6, 9, 12, 18 and 24 months of the few stability lots. The specifications limits derived by PI, RI and TI, using 20 and 60 observations are plotted in Fig. 3, and the data simulated is extended to 200 observations here. For the pre-marketing specification determination, it is often assumed that all observations are independent, even though some could be stability data, with no trend. As shown in Fig. 3, the specification limits are determined with the first 20 observations, using PI, RI and TI. The width of TI is much larger than PI and RI. The specification limits are then applied to the observations beyond the first 20 observations, until an updated revision, with an

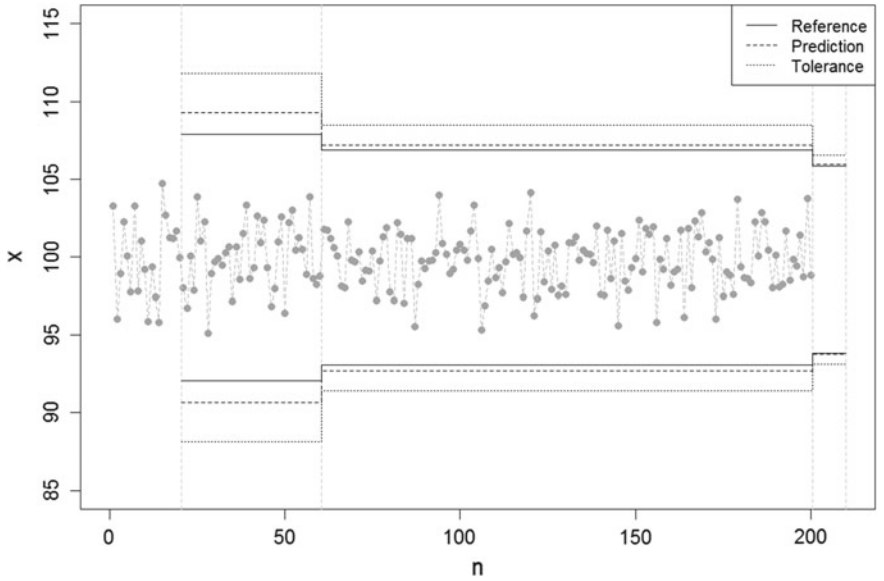


Fig. 3 The lower and upper specification limits determined by TI, RI, and PI when 20 observations are available. The limits are revised when a total of 60 observations or a total of 200 observations are available

additional 40 observations correlated from post-marketing lots manufactured. The limits are updated again, with another 140 observations (a total of 200 observations). As shown in Fig. 3, from the first 20 observations, the specification interval is (93.02, 108.03) by reference interval, (91.69, 109.36) by prediction interval and (89.30, 111.75) by tolerance interval. For the pre-marketing determination, TI gives the lower and upper specification limits, almost two standard deviations below, and above, the corresponding limits determined by RI.

5 Specification Determined Updated with Post-marketing Data

Often times, the scientist may want to revise the specifications using updated data of marketing lots. In Fig. 3, we show an example of the same product with up to 200 lots, after marketing. The qualities of the lots were not changed from the original 20 values. As a contrast, PI, RI and TI are calculated for the data, up to sixty lots. It shows clearly, with larger sample size, RI provides a similar interval, but both PI and TI are narrowed with larger sample size. The change of RI, with the updated data, is due to the sample variance changes.

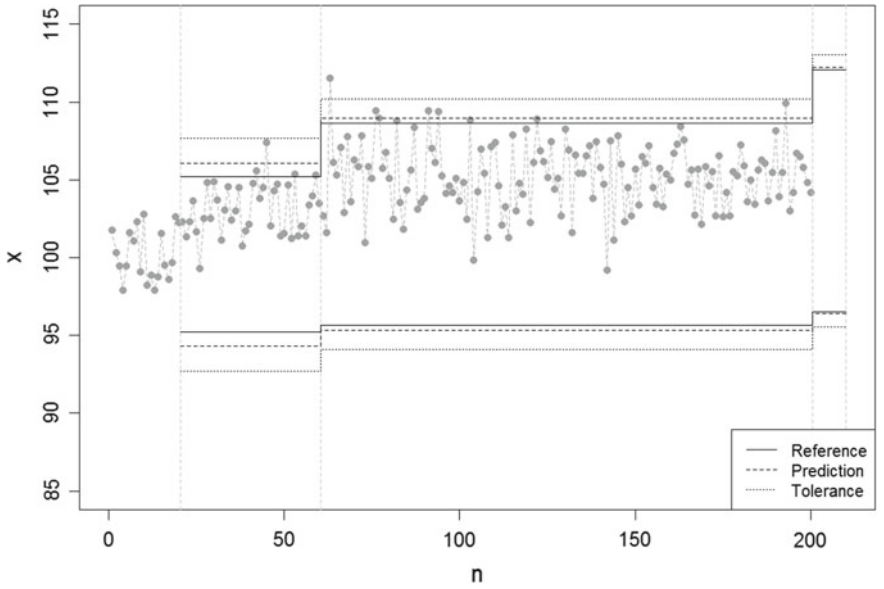


Fig. 4 The lower and upper specification limits determined by TI, RI, and PI when 20 observations are available. The limits are then revised using the data from post-marketing lots information with mean shifting to upper values

On the other hand, the post-marketing data may involve with shift and manufacturing correction, to change the distribution of the data of the attribute. In these cases, the data variability will be a combination of variability of attribute, change and correction. In Fig. 4, the manufactured lots were clearly shifted upward, with most attribute values above the mean line of the control chart. The mean and variance of the attribute were different between pre-marketing and post-marketing. The next forty lots manufactured post-marketing, shifted, crossing the upper specification limit, determined by the first 20 pre-marketing lots. An updated revision, using additional forty post-marketing lots does not help to control the manufacturing process. In Fig. 5, we show another manufacturing process, with both mean shift and short term correction. The products become manufactured asymmetric to the mean. An updated revision of specification will not replace a better quality control adjustment.

6 Conclusions and Discussion

Prediction interval, reference interval and tolerance interval were often used and proposed to be used as the tool to determine the specification limits of a quality attribute. Reference interval is probably the one tool used most frequently by chemists and biologists for its simplicity. It derives from the point estimate of a target asymptotic

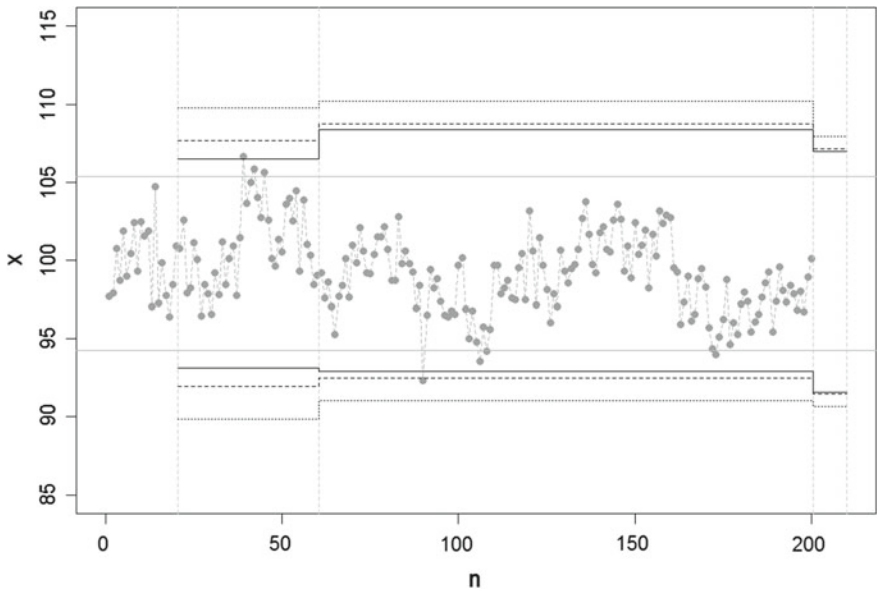


Fig. 5 The lower and upper specification limits determined by TI, RI, and PI when 20 observations are available. The limits are then revised using the data with post-marketing lots information with mean shifting and manufacturing correction

coverage. Prediction interval was proposed for the objective to set the limits of quality specification, by controlling the probability of any single future lot falls, within such limits. The controlled probability is called the coverage for predictive interval. Tolerance interval provides the approach in order to assure the coverage with a level of confidence. For the estimation purpose, statisticians will consider tolerance interval most appropriate. However, the precision only improves with large sample size, and may not be practically useful for setting a specification limits with a small sample size.

It is often difficult to evaluate the three interval approaches by comparing their coverage. In this article, we propose to standardize the confidence level for the three approaches. Through this standardized confidence level, we may recalculate the true coverage, and have a better understanding across the different approaches. With the calibrated true coverage of each approach, we can see that, with the same confidence level, using reference intervals to determine the specification limits will lead to larger population coverage, with a large sample size. This property is probably more appearing for setting specification limits.

In this article, we limited the discussion on normally distributed data. For lognormal distribution, it should be an easy generalization. Other than these distributions, further research will be developed.

References

1. Dong, X., Tsong, Y., Shen, M., Zhong, J.: Using tolerance intervals for assessment of pharmaceutical quality. J. Biopharm. Stat. **25**(2), 317–327 (2015)
2. European Medicines Agency (EMA): Report on the expert workshop on setting specifications for biotech products, London, 9 September 2011
3. Faulkenberry, G.D., Daly, J.C.: Sample size for tolerance limits on a normal distribution. Technometrics **12**(4), 813–821 (1970)
4. Geisser G.: Predictive Inference: An Introduction. Chapman and Hall (1993)
5. ICH Guideline Q6A: Specifications: test procedures and acceptance criteria for new drug substances and new drug procedures: chemical substances (1999)
6. DiFeo, T.J.: Drug product development: a technical review of chemistry, manufacturing, and controls information for the support of pharmaceutical compound licensing activities. Drug Dev. Ind. Pharm. **29**, 939–958 (2003)

Counting Test and Parametric Two One-Sided Tolerance Interval Test for Content Uniformity Using Large Sample Sizes

Meiyu Shen, Yi Tsong and Richard Lostritto

Abstract The purpose of uniformity of dosage unit test is to determine the degree of uniformity in the amount of drug substance among dosage units in a batch. Recently, there are several nonparametric methods including the large sample counting approach proposed in European Pharmacopeia 8.1 (EU Option 2). All nonparametric methods specify a maximum number of tablets, of which the contents fall outside the interval (85%, 115%) of labeling claim (LC) for a given large sample size. The nonparametric method in European Pharmacopeia requires another maximum number of tablets, of which the contents fall outside the interval (75%, 125%) LC. We denote the nonparametric method as the counting test which will be used in the rest of the article. We focus on the comparison of the acceptance probabilities between EU Option 2 and the parametric two one-sided tolerance intervals (PTIT_matchUSP90) test. Obviously, a counting test is less efficient than a parametric test in general. Our simulation study clearly shows that the EU Option 2 is not sensitive to batches with a large variability in contents which follow a normal distribution with an off-target mean, a mixture of two normal distributions, or a mixture of a uniform distribution with small percent of extreme values. The EU Option 2 is not sensitive to the mean shift of the majority population (97%) from 100% LC to 90% LC. In addition, the EU Option 2 is not sensitive to low assay values (about 90% LC). The EU Option 2 is over-sensitive to one extreme case: 97% tablets with 100% LC and 3% tablets with 76% LC.

Keywords Uniformity of dosage units · Content uniformity · Large sample sizes · Tolerance interval · Counting test

This article represents only the authors' opinion and not necessarily the official position of the US Food and Drug Administration.

M. Shen (✉) · Y. Tsong · R. Lostritto
Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, USA
e-mail: meiyu.shen@fda.hhs.gov

M. Shen
Division Biometrics VI, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_2

1 Introduction

The purpose of the uniformity of dosage unit (UDU) test is to quantify the variability in the amount of drug substance between dosage units in a batch and to compare that variability to a corresponding specification [6, 11]. In most cases, the UDU test corresponds to the amount of drug substance contained within the unit (e.g., tablets and capsules). In some cases, UDU corresponds to the amount of drug delivered via a dose metering device component (e.g., metered dose inhalers, or MDIs). In these cases, the test is referred to as uniformity of delivered dose (UDD). In a few instances, both the UDU and UDD tests may apply to the same product. For example, in a dry powder inhaler (DPI) that uses individual capsules or blisters containing the formulation to be inhaled. In such cases, UDU testing is needed to control for the variability in the individual capsules or blisters, and UDD testing is needed to control for the delivered dose ex-device used for inhalation. Content uniformity test is based on the assay of the individual content of drug substance(s) in a number of dosage units. Content uniformity test is required by the regulatory authorities, such as the US Food and Drug Administration, to confirm that the unit dose of a drug product is consistent with the label claim. ICH guideline Q4B, Annex 6 [5] recommends that a Pharmacopeia procedure is used to assess the uniformity of dosage units. For example, the United States Pharmacopoeia (USP) publishes USP <905>, the harmonized content uniformity test using tolerance interval and indifference zone concepts [11]. Although USP <905> is not intended for batch releases as it is clearly declared in the USP pharmacopeia, many pharmaceutical companies propose to implement USP <905> as the batch release specification. USP <905> consists of two stages. Ten dosage units (e.g., tablets) are randomly sampled from a batch at the 1st stage and 20 additional units are randomly sampled from a batch at the 2nd stage if the sample is not accepted at the 1st stage. As a compendia test, the USP <905>, a harmonized content uniformity test, has been developed for a small fixed sample size, and it provides no guidance on how to conduct the test at different sample sizes. There have been several proposals in the literature for content uniformity using large sample sizes. In [8], Sandell, Vukovinsky, Diener, Hofer, Pazdan, and Timmermans proposed a one-tiered nonparametric test which counted the number of tablets with content outside (85%, 115%) LC. The batch complies if no more than c units are outside (85%, 115%) LC. They also provided values of c for a selection of potential sample sizes. This proposed counting test had operating characteristic (OC) curves intersecting with the OC curve of the USP <905> around 45% probability of accepting a batch with the batch mean of 100% LC (see Fig. 3, [8]). This counting test has been proposed as a batch-release test when a large number of dosage units is tested.

In 2010, Bergum and Vukovinsky proposed a modified version of the counting test by Sandell et al. They proposed setting the integer part of $0.03 \times n$ equal to c^*, which was the acceptance limit for the maximum number of tablets outside (85%, 115%) LC, here n was the number of dosage units tested (For example, with $n = 260$, c^* is 7). Thus, a batch complies if the number of tablets outside (85%, 115%) LC is no more than c^*.

In 2012, the Council of Europe published two options for uniformity of dosage units, using large sample sizes in European Pharmacopoeia 7.7. Option 1 is a parametric two-sided tolerance interval based method modified with an indifference zone and counting units outside of (0.75 M, 1.25 M). Here, M is defined by the sample mean of the tested n dosage units, $\bar{X}$, as: $M = 98.5\%$ if $\bar{X} < 98.5\%$, $M = 101.5\%$ if $\bar{X} > 101.5\%$, and $M = \bar{X}$ otherwise. Option 2 is a nonparametric counting method with an additional indifference zone concept. Option 2 in the EP 7.7 was intended to revise the original proposal of large sample UDU test whose OC curve intersects with OC curve of USP <905> around 45% passing probability [7]. However, mathematical and statistical methodologies to derive the revised Option 2 were not discussed in [7]. It would result in an unanswered question how to determine *c1* and *c2* for sample sizes not included in Table 2.9.47.2 Council of Europe, [2].

In April 2014, the Council of Europe officially published the corrected version of EU Option 2 in which the indifference zone is removed for uniformity of dosage units using large sample sizes in European Pharmacopoeia 8.1 (Council of Europe, [3]. With Option 2, a batch is accepted if no more than *c1* units lie outside (0.85T, 1.15T) and no more than *c2* units lie outside (0.75T, 1.25T), where T is defined as 100%. The counts *c1* and *c2* for the selected sample sizes are given in Table 2.9.47.-2 of EP 8.1. The EP 8.1 provides no explanation how the counts were derived. But the counts *c2* are almost identical to Option 1.

Note that the approaches described earlier provide some assurance that a batch passing the release test may comply with USP <905>, the content uniformity sampling test using small number of dosage units.

Shen et al. [10] studied the statistical properties of the large sample tests for content uniformity. They proposed a large sample acceptance sampling method based on parametric two one-sided tolerance intervals. In particular, this proposed method was designed to have its OC curve for any given sample size intersect with the OC curve of the harmonized USP <905> at the acceptance probability of 90% for a batch whose individual tablets follow normality with the mean of 100% LC. They denoted this proposed method as PTIT_matchUSP90 method. They compared the acceptance probabilities of PTIT_matchUSP90 method with those of the two procedures recommended in the European Pharmacopeia 7.7 when the unit dose is assumed to be a continuous random variable, e.g., normal or mixture of normal distributions. They also showed that with two one-sided tolerance intervals criteria, the proposed test could accept or reject a batch based on the percentage outside either 85% or 115% LC regardless the distribution was normal or not. Such statistical property assures high acceptance probability of the same batch when tested again with USP <905>, a small sample size test, since OC curve (acceptance probability versus standard deviation) of PTIT_matchUSP90 for any sample size intersects with OC curve of USP <905> at 90% of acceptance probability when the batch mean is on the target (100% LC). OC curves of others [8, 1] that intersect with OC curve of USP <905> at <90% (e.g., 45%) of acceptance probability provide low assurance that the batch may be accepted when subjected to the USP<905> UDU sampling acceptance plan. In this paper, we provide more details on lack of quality assurance of a batch accepted by the proposed non-parametric plans when subjected to USP<905> UDU plan. We also demonstrate

the results through simulation results of comparisons between PTIT_matchUSP90 and EU Option 2 in European Pharmacopoeia 8.1. Some of the data distributions used in the following comparisons may be extreme in order to make some points.

The structure of this paper is arranged as follows. In Sect. 2, we present the details of EU Option 2 (the counting method) proposed in the European Pharmacopoeia 8.1 and our proposed PTIT_matchUSP90 method. In Sect. 3, we discuss the simulation methods to generate the OC curves. In Sect. 4, we compare PTIT_matchUSP90 method and the EU Option 2 for a normal distribution, a mixture of two normal distributions, and other specific scenarios.

2 Sampling Tests

Although there are several versions of the counting method [8], such as Bergum et al. [1], Council of Europe [2, 7], and Council of Europe [3], we will focus on EU Option 2 in the European Pharmacopoeia 8.1 since all counting methods share certain statistical properties as EU Option 2.

EU Option 2
EU Option 2 is a nonparametric acceptance sampling test defined as follows: With a sample of n ($\geq$100) units, count the number of individual dosage units with a content outside $(1 \pm L1 \times 0.01)100$ LC and the number of individual dosage units with a content outside $(1 \pm L2 \times 0.01)100$ LC. Here $L1 = 15$ and $L2 = 25$. A batch passes the test for the uniformity of dosage units if

1. The number of individual dosage units outside $(1 \pm L\ 1 \times 0.01)100$ LC is less than or equal to $c1$; and
2. The number of individual dosage units outside $(1 \pm L2 \times 0.01)100$ LC is less than or equal to $c2$.

Here, $c1$ and $c2$ are listed in Table 2.9.47.-2 of the European pharmacopoeia 8.1.

PTIT_matchUSP90
Let $\bar{X}$ be the sample mean and S the sample standard deviation of contents of n units. Unlike the USP<905> with a two-tier procedure, PTIT_matchUSP90 is a single tier content uniformity test with large sample size of n. A batch passes the PTIT_matchUSP90 test if $\left(\bar{X} - K(n)S, \bar{X} + K(n)S\right) \in (85\%, 115\%)$LC, where $K(n)$ is the tolerance factor for the PTIT_matchUSP90 sampling test using n units and is derived as follows.

The PTIT_matchUSP90 test consists of two one-sided hypothesis tests which ensure that the proportions of over-filled and under-filled tablets are both under control. Let $p(n)$ (see Tsong et al. [12]) be the desired proportion of the population within the interval (85%, 115%) LC for the PTIT_matchUSP90 sampling test using n units. In general, the acceptance probability of a batch with a given standard deviation when the mean is 100% LC increases with an increase of n for a fixed coverage within the interval (85%, 115%) LC. Since the PTIT_matchUSP90 test is designed to have

90% acceptance probability for any n at the standard deviation at which USP <905> has 90% acceptance probability when the batch mean is 100% LC, then we must increase $p(n)$ within the interval (85%, 115%) LC. The first one-sided test is to ensure that the proportion of over-filled tablets (>115%LC) is less than $(1-p(n))/2$ and the second one-sided test is to ensure the proportion of under-filled tablets (<85% LC) is also less than $(1-p(n))/2$. The first test is carried out by comparing the upper one-sided tolerance interval with 95% confidence level and $(1 + p(n))/2$ coverage with the upper limit 115%LC. More specifically, this one-sided tolerance interval is in the form of $(-\infty, \bar{X} + K(n)S)$, where the tolerance factor $K(n)$ [4] is solved by the following equation for given values of μ and σ:

$$\Pr\left\{\left[\int_{-\infty}^{\bar{X}+K(n)S} n(x:\mu,\sigma)dx\right] \geq \frac{1+p(n)}{2}\right\} = 0.95$$

where $f(x:\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. It can be easily showed that for normally distributed random variable, $K(n) = t_{n-1,1-\alpha}(Z_{(1+p(n))/2}\sqrt{n})/\sqrt{n}$ where $Z_{(1+p(n))/2}$ is the 100 $(1 + p(n))/2$%-th percentile of the standard normal distribution and $t_v(\gamma)$ denotes the non-central t-distribution with degree of freedom v and non-centrality parameter γ. If $\bar{X} + K(n)$S < 115% LC, it ensures that the percentage of over-filled tablets is less than $(1-p(n))/2$. Similarly, if $\bar{X}-K(n)S$ > 85% LC, the second test ensures that the percentage of under-filled tablets (<85% LC) is less than (1- $p(n)$)/2.

The algorithm to obtain $K(n)$ and $p(n)$ is briefly described below under normality and the detail should be referred to the reference [12]. First, $K(n)$ is determined for any n such that the probability of the event of $\left(\bar{X} - K(n)S, \bar{X} + K(n)S\right) \in$ (85%, 115%)LC is 90% at the standard deviation at which the USP <905> has 90% acceptance probability when the batch mean is 100% LC. Second, $p(n)$ is determined by satisfying the equation $K(n) = t_{n-1,1-\alpha}(Z_{(1+p(n))/2}\sqrt{n})/\sqrt{n}$. Once we have $K(n)$, the acceptance probability of a batch with an individual table content as a random variable can be obtained from Monte Carlo simulations described in the following Sect. 3.

3 Monte Carlo Simulation Method

Assume that content values, $X_1, X_2, \ldots, X_n$, are random samples from some specific distributions with known parameters. The acceptance probability of a batch passing any sampling test can be obtained with the following Monte Carlo simulation procedure:

(1) Generate data, $X_1, X_2, \ldots, X_n$ from a specific parametric distribution with the known parameters.
(2) Calculate the sample statistics for the n samples.

(3) Determine if the sample satisfies the acceptance criteria of a given sampling test.
(4) Repeat Steps 1 to 3 for a large number of times, say 100,000 times.
(5) Calculate the average rate of accepting a batch for the sampling test.

4 Comparison of Acceptance Probabilities Between EU Option 2 and PTIT_matchUSP90

Generally speaking, a good sampling plan should be sensitive to variations in content, such as shifting of batch means, deviation from normal distribution, mixture of two populations, and out of boundary limit percentage. In this section, we compare acceptance probabilities between EU Option 2 and PTIT_matchUSP90 for two normal distributions with on target mean and off target mean, a mixture of two normal distributions, and three extreme cases with non-normal distributions.

4.1 Comparison of PTIT_matchUSP90 with EU Option 2 for Normal Variables with 100% LC Mean when n = *1000*

When $n = 1000$, $p(n) = 0.9672$ and $K(n) = 2.2321$. OC curves of PTIT_matchUSP90 (in dashed lines) using 1000 tablets, EU Option 2 (in solid lines) using 1000 tablets, and the USP <905> (in dashed-dotted line) against the standard deviation for a batch with mean of 100% LC under normality are plotted in Fig. 1. It can be seen that the acceptance probability of PTIT_matchUSP90 for a given standard deviation between 6 and 7 is always slightly smaller than that of EU Option 2 for batches with 100% LC. As standard deviation increases, the difference in acceptance probability between PTIT_matchUSP90 and EU Option 2 increases. Overall, the OC curve of PTIT_matchUSP90 for normal variables with the on-target (100% LC) mean is close to that of EU Option 2. OC curve of PTIT_matchUSP90 for normal variables with the on-target (100% LC) mean intersects with that of the USP <905> at 90% passing probability.

4.2 Comparison of PTIT_matchUSP90 with EU Option 2 for Normal Variables with 102% LC Mean when n = *1000*

Figure 2 compares OC curves of PTIT_matchUSP90 (in dashed lines) using 1000 tablets and EU Option 2 (in solid lines) using 1000 tablets against the standard deviation for a batch with mean of 102% LC under normality. Clearly, the acceptance probability of PTIT_matchUSP90 for a given standard deviation between 5.6 and 7

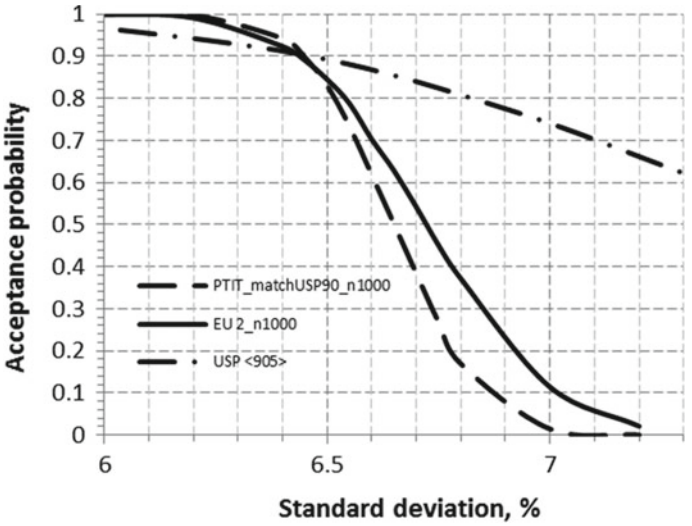


Fig. 1 Comparison of PTIT_matchUSP90 with European Union option 2 for individual dose content distributed as independent and identical normal variable with 100% of the label claim when $n = 1000$

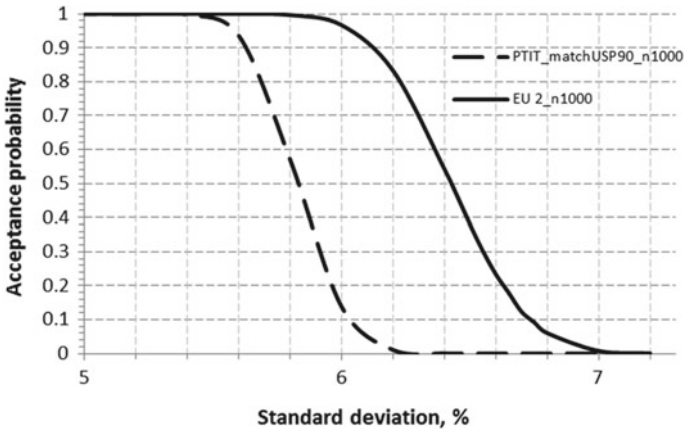


Fig. 2 Comparison of PTIT_matchUSP90 with European Union option 2 for individual dose content distributed as independent and identical normal variable with 102% of the label claim when $n = 1000$

is always much smaller than that of EU Option 2 for batches with 102% LC. The acceptance probability of PTIT_matchUSP90 is 0.1; while that of EU Option 2 is almost 1 when the standard deviation is 6%. Note that OC curve of USP <905> is not added to Figs. 2 and 3 since USP <905> rewards batches with the off-target mean due to indifference zone [9].

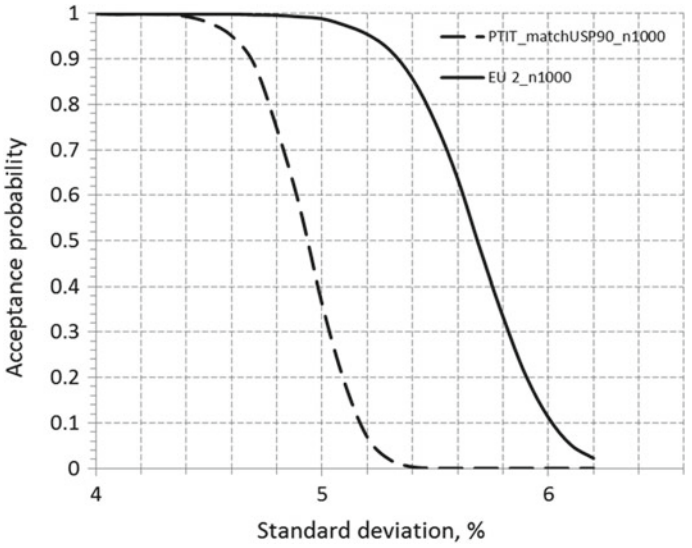


Fig. 3 Comparison of PTIT_matchUSP90 with European Union option 2 for individual dose content distributed as a mixture of two normal variables, the overall mean of the mixture is 104% when $n = 1000$

4.3 Comparison of PTIT_matchUSP90 with EU Option 2 for a Mixture of Two Normal Variables with off Target Mean When n = *1000*

To assess the performance of the two methods for data not normally distributed, we compare the OC curves of a batch having a mixture of two normal distributions with different mean values and different variances. One variable is normally distributed with the mean of 100% LC and the standard deviation of 10%. The probability of being this variable is 0.1. The other variable is also normally distributed with the mean of 104.4% and the standard deviation of σ_2. The overall mean of the mixture variable is 104%LC. σ_2 can be determined for a given standard deviation σ. From Fig. 3, it can be seen that the acceptance probability of PTIT_matchUSP90 is smaller than that of EU Option 2 for any given standard deviation greater than 4.7 when $n = 1{,}000$. For a standard deviation of 5.3%, the acceptance probability of PTIT_matchUSP90 is almost zero while that of EU Option 2 is slightly smaller than 1. It is apparent that EU Option 2 is not sensitive to large variances and a mean shift under a mixture of two normal variables as the PTIT_matchUSP90 is.

4.4 Comparison of PTIT_matchUSP90 with EU Option 2 for Three Extreme Cases

In order to further illustrate the difference between the two approaches, we also consider the following three extreme cases with non-normal distributions. In the first case, the tablet contents are simulated from the mixture of a uniform distribution and a point distribution, such that the individual tablet content is assumed to be an independently and identically distributed random variable which, with probability 97%, is uniformly distributed in the interval (85%, 115%) LC, and with a probability of 3% is equal to 84% LC. The purpose of the first case is to investigate how sensitive these two sampling tests react to a large variation in contents.

In the second case, the tablet contents are simulated from the mixture of two point distributions such that individual tablet contents is assumed to be an independently and identically distributed random variable which is equal to 100% LC and 76% LC with probability 97% and 3%, respectively. The purpose of the second case is to investigate how sensitive these two sampling tests respond to a small percent of extreme observations which leads to a small variability in content although 97% of individual tablets are taken from a batch with 100% LC.

In the third case, individual tablet contents are again simulated from the mixture of two point distributions such that individual tablet content is assumed to be an independently and identically distributed random variable which is equal to 90% LC and 76% LC with probability 97% and 3%, respectively. With this third case, we investigate how sensitive the two sampling tests respond to excessive low contents with low percentage. This kind of batches sometimes can pass the assay criteria if all samples are 90% LC since the assay criteria are (90%, 110%) LC.

The results of the simulation are described as follow.

Case 1: Mixture of a uniform distribution and one single value
For case 1 data, the acceptance probability of the USP <905> is 3.72% for a sample size of 30 units. As shown in Table 1, it shows that the EU option 2 has acceptance probabilities higher than 50% for sample sizes up to 200. The acceptance probability of EU Option 2 goes down to less than 20% only when the sample size is significantly larger than 1000. On the other hand, such a batch has 0% probability of acceptance using the PTIT_macthUSP90 procedure with any sample size greater than 100. It illustrates that the EU option 2 does not take the content variability into consideration and falsely accepts such batch for content uniformity.

Cases 2 and 3: Mixture of two distinct points
Case 2 in Table 2 shows that the EU Option 2 has 64% probability of passing a batch whose individual tablet content is assumed to be an independently and identically random variable which equals to 100% LC with 97% probability and equals to 76% LC with 3% probability when the sample size is 100; the PTIT_matchUSP90 has about 99% probability of passing the batch. When the sample size increases to 150, the EU Option 2 has 53% probability of passing the batch and the PTIT_matchUSP90 has about 99% probability of passing the batch. As the sample size increases, the

Table 1 Acceptance probabilities of EU option 2 method for mixture of a uniform distribution and a point distribution

Sample size, n	Acceptance probability of EU option 2
100	0.6458
150	0.5276
200	0.6047
300	0.455
500	0.3509
1000	0.2075
1500	0.0899

probability of passing EU Option 2 decreases; while the probability of passing PTIT_matchUSP90 is about 99%. This should be the case since a batch is almost ideal batch with 97% of all values on target. Based on the simulation result for Case 3 in Table 2, the EU Option 2 has 65% probability of passing the batch whose individual tablet content is assumed to be an independently and identically random variable which equals to 90% LC with 97% probability and equals to 76% LC with 3% probability when the sample size is 100; the PTIT_matchUSP90 has only 41.7% probability of passing the batch. When the sample size increases to 1000, the EU Option 2 has about 20% probability of passing the batch and the PTIT_matchUSP90 has only about 7% probability of passing the batch. Case 3 illustrates that the EU Option 2 is not sensitive to low assay value (about 90% LC) since the EU Option 2 has more than 50% probability of passing the batch whose 97% of individual tablets have 90% LC when a sample size is 150. Furthermore both Case 2 and Case 3 in Table 2 reveal that the EU Option 2 is over-sensitive to the variation in content from small percent of extreme observations.

Results in Tables 2 show that for a batch in Case 2 and Case 3, the EU Option 2 is not sensitive to the mean shift of the majority population (97%) from 100% LC to 90% LC. For the same sample size, the batch of 97% tablets with 100% LC content will be accepted with the same probability as the batch of 97% tablets with 90% LC using EU Option 2. As also shown in the Tables, EU Option 2 has about 64% acceptance rate for batches of Case 2 or Case 3 with a sample size of 100.

5 Discussion and Conclusion

It is well understood that content values may not be normally distributed and a non-parametric content uniformity sampling acceptance plan may need to be developed. However, when developing a new large sample non-parametric approach based on USP <905>, we need to provide assurance of proper power to satisfy small sample compendia method USP <905> applied to samples any time. Any method developed should not fail the basic requirement that samples on shelf should pass the USP <905> at any time. In this paper, we reviewed the non-parametric approach

Table 2 Comparison of the acceptance probabilities between EU option 2 and PTIT_matchUSP90 for two cases with a mixture of two distinct values

Case	Sample size, n	Acceptance probability	
		EU Option 2	PTIT_matchUSP90
Case 2: X_i is 100 with 97% probability, and 76 with 3% probability	100	0.6485	0.9887
	150	0.5346	0.9972
	200	0.6114	0.9984
	300	0.4693	0.9998
	500	0.3565	1.0
	1000	0.2057	1.0
Case 3: X_i is 90 with 97% probability, and 76 with 3% probability	100	0.6531	0.417
	150	0.5369	0.3414
	200	0.6046	0.2833
	300	0.4477	0.3177
	500	0.3617	0.1848
	1000	0.1978	0.0721

proposed in the European pharmacopoeia 8.1 and in the recent literature. Due to lack of publication of mathematical or statistical derivation of the approach, we can only evaluate the EU Option 2 in the European pharmacopoeia 8.1 through simulation study. PTIT_matchUSP90 is developed to assure 90% acceptance probability at the standard deviation where USP <905> has 90% acceptance probability for batches with 100% LC mean. The following are the summaries of such evaluations.

Under normality, the acceptance probability of EU option 2 is slightly higher than that of PTIT_matchUSP90 for batches with a large variability. However, under normality, EU Option 2 is not sensitive to the off-target mean since the acceptance probability of EU Option 2 is much larger than that of PTIT_matchUSP90 for batches with large variances. Furthermore, EU Option 2 is not sensitive to large variances and a mean shift under a mixture of two normal variables as the PTIT_matchUSP90 is.

Given that 97% of tablets have content values falling within the range from 85% LC to 115% LC, clearly EU Option 2 is not sensitive to a large variability in contents from a uniform distribution; on other hand, EU Option 2 is oversensitive to a small percent of extreme observations. As a result, EU Option 2 provides low assurance to be accepted by USP<905> in the subsequent compendia testing. The EU Option 2 is not sensitive to the mean shift of the majority population (97%) from 100% LC to 90% LC. In addition, the EU Option 2 is not sensitive to low assay values (about 90% LC). Together, it shows that it may be hard to declare the EU Option 2 is a good sampling test in the sense of providing high probability assurance of acceptance by the subsequent USP<905> compendia test. Certainly, the examples simulated are extremes for any batch with consistent good quality. However, they are useful to illustrate the potential problems of interpreting a variable acceptance sampling test

(e.g., tolerance interval method) with an acceptance sampling by attribute test (e.g., counting test).

In conclusion, the counting method such as EU Option 2 performs insensitively for a batch with a large variability and an off target mean even under normality and a mixture of two normal variables, and performs over-sensitively for a batch with small percent of low extreme values. For content uniformity assessment, information such as variability can be lost when EU Option 2 is used.

Acknowledgements The authors would like to thank two anonymous reviewers for their comments.

References

1. Bergum, J., Vukovinsky, K.E.: A proposed content-uniformity test for large sample sizes. Pharm. Technol. **34**(11), 72–79 (2010)
2. Council of Europe, Uniformity of dosage units using large sample sizes. Chapter 2.9.47 of European Pharmacopoeia 7.7. 5142–5145, Renouf Pub Co Ltd; PhEur 7th edn (16 Oct 2012)
3. Council of Europe, Uniformity of dosage units using large sample sizes. Chapter 2.9.47 of European Pharmacopoeia 8.1. 3669–3671, Renouf Pub Co Ltd; PhEur 8th edn (April 2014)
4. Faulkenberry, G.D., Daley, J.C.: Sample size for tolerance limits on a normal distribution. Technometrics **12**(4), 813–821 (1970)
5. Food and Drug Administration, Guidance for Industry, ICH Q4B Evaluation and Recommendation of Pharmacopoeial Texts for Use in the ICH Regions, Annex 6 Uniformity of Dosage Units General Chapter, June 2014
6. Food and Drug Administration, Guidance for Industry, Nasal Spray and Inhalation Solution, Suspension, and Spray Drug Products—Chemistry, Manufacturing, and Controls Documentation (2002)
7. Holte, Ø., Horvat, M.: Uniformity of dosage units using large sample sizes. Pharma. Sci. Technol. **36**(10), 118–122 (2012)
8. Sandell, D., Vukovinsky, K., Diener, M., Hofer, J., Pazdan, J., Timmermans, J.: Development of a content uniformity test suitable for large sample sizes. Drug Inf. J. **40**(3), 337–344 (2006)
9. Shen, M., Tsong, Y.: Bias of the USP harmonized test for dose content uniformity. Stimuli. Revis. Process. **37** (2011)
10. Shen, M., Tsong, Y., Dong, X.: Statistical properties of large sample tests for dose content uniformity. Therapeutic Innov. Regul. Sci. **48**(5), 613–622 (2014)
11. The United States Pharmacopeia Convention, General Chapter <905> Uniformity of dosage units, United States Pharmacopeia (USP) **38**, 675–679 (2015). Rockville, MD
12. Tsong, Y., Dong, X., Shen, M., Lostritto, R.T.: Quality assurance test of delivered dose uniformity of multiple-dose inhaler and dry powder inhaler drug products. J. Biopharm. Stat. **25**(2), 328–338 (2015)

Part II
Analytical Biosimilar and Process Validation

Sample Size Consideration for Equivalent Test of Tier-1 Quality Attributes for Analytical Biosimilarity Assessment

Tianhua Wang, Yi Tsong and Meiyu Shen

Abstract FDA recommends a stepwise approach for obtaining the totality-of-the-evidence for assessing biosimilarity between a proposed biosimilar product and its corresponding reference biologic product being considered (US Food and Drug Administration.: Guidance for industry: scientific considerations in demonstrating biosimilarity to a reference product. US Food and Drug Administration, Silver Spring, 2015 [6]). The stepwise approach starts with analytical studies for assessing similarity in critical quality attributes (CQAs), which are relevant to clinical outcomes. For critical quality attributes that are most relevant to clinical outcomes (Tier 1 CQAs), FDA requires equivalence testing to be performed for similarity assessment, based on an equivalence acceptance criteria. In practice, the number of Tier 1 CQAs might be greater than one, and should be no more than four. The number of biosimilar lots is often recommended to be no less than 10, and the ratio between the reference product sample size and biosimilar product sample size is recommended within the range from 2/3 to 3/2 (US Food and Drug Administration.: Guidance for industry: Statistical Approaches to Evaluate Analytical Similarity. US Food and Drug Administration, Silver Spring, 2017 [7]). Accordingly, we derive the formulas for the power calculation for the sample size for analytical similarity assessment based on the equivalence testing currently used in analytical biosimilar assessment (Tsong et al. J Biopharm Stat **27:**197–205, (2017)[10]).

Keywords Analytical similarity · Equivalence testing · Sample size · Analytical power function · Correlation coefficient · Satterthwaite approximation · Sparse grid

Disclaimer: This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

T. Wang (✉) · Y. Tsong · M. Shen
Office of Biostatistics/Office of Translational Science,
Center for Drug Evaluation and Research, U.S. Food and Drug Administration,
Room 4667, Bldg. 21, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA
e-mail: tianhua.wang@fda.hhs.gov; tianhuawang2009@gmail.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_3

1 Introduction

To evaluate the analytical similarity between the proposed biosimilar product and the US-licensed reference product, the equivalence testing approach for the critical quantitative quality attributes (CQAs), assigned to Tier 1 in the tiered approach was developed by [8]. In the equivalence testing, the proposed biosimilar product is concluded to be highly similar to the US-licensed reference product regarding a specific Tier 1 quality attribute, if the following two one-sided null hypotheses are rejected at a nominal level α (e.g., 0.05).

$$H_0 : \mu_T - \mu_R \leq -1.5\sigma_R \quad \text{or} \quad \mu_T - \mu_R \geq 1.5\sigma_R \tag{1}$$

$$H_a : -1.5\sigma_R < \mu_T - \mu_R < 1.5\sigma_R$$

where μ_T and μ_R are respectively the population means of the biosimilar product and reference product, and σ_R is the standard deviation of the reference product for a given quality attribute. In practice, the reference variability σ_R is usually unknown, and needs to be estimated from the reference sample. In this paper, the margin $1.5\sigma_R$ is estimated from the values of the reference product lots generated by the applicant, and is established, assuming it is known and a constant. The hypothesis of (1) is rejected at a significance level of α when the test statistics $T_1 > t_{df^*,1-\alpha}$ and $T_2 \leq -t_{df^*,1-\alpha}$ where

$$T_1 = \frac{\left(\overline{X}_T - \overline{X}_R\right) + 1.5\sigma_R}{\sqrt{\frac{S_T^2}{n_T^*} + \frac{S_R^2}{n_R^*}}}; \quad T_2 = \frac{\left(\overline{X}_T - \overline{X}_R\right) - 1.5\sigma_R}{\sqrt{\frac{S_T^2}{n_T^*} + \frac{S_R^2}{n_R^*}}} \tag{2}$$

$\overline{X}_T$ and $\overline{X}_R$ are respectively the biosimilar and reference sample means; S_T^2 and S_R^2 are respectively the biosimilar and reference sample variances; $n_R^* = \min(n_R, 1.5n_T)$ and $n_T^* = \min(n_T, 1.5n_R)$ are respectively the adjusted reference product and biosimilar product sample sizes based on the original reference product and biosimilar product sample sizes n_R and n_T. The $t_{df^*,1-\alpha}$ is the 100α-th percentile of the t-distribution with the adjusted degrees of freedom df^*. In practice, both the variances and sample sizes of the test and reference products are different, therefore the sampling distribution needs to be approximated, using the degrees of freedon df*, determined by Satterthwaite (1964) approximation as $df^* = \left(\frac{\sigma_T^2}{n_T^*} + \frac{\sigma_R^2}{n_R^*}\right)^2 / \left(\frac{\sigma_T^4}{(n_T^*)^2(n_T-1)} + \frac{\sigma_R^4}{(n_R^*)^2(n_R-1)}\right)$. The test above is the same as requiring the $(1-2\alpha)100\,\%$ two-sided confidence interval of the mean difference $\left(\overline{X}_T - \overline{X}_R\right) \pm t_{df^*,1-\alpha}\sqrt{S_T^2/n_T^* + S_R^2/n_R^*}$ to be completely covered by the equivalence margin $(-1.5\sigma_R, 1.5\sigma_R)$.

Chow et al. [3] discussed the sample size requirement in analytical studies for similarity assessment. In that paper, they discussed the sample size determination approach, which was a function of (i) overall significant level α, (ii) type II error rate β or power $1-\beta$, (iii) clinically or scientifically meaningful difference $\mu_T - \mu_R$, and

(iv) the variability associated with the reference product, assuming that $\sigma_T = \sigma_R$. Their sample size was determined by

$$n_T = f(\alpha, \beta, \mu_T - \mu_R, k, \sigma_R) = \frac{\left(Z_\alpha + Z_{\beta/2}\right)^2 \sigma_R^2 (1 + 1/k)}{(\delta - |\mu_T - \mu_R|)^2} \tag{3}$$

where $k = n_T/n_R$ and $\delta = 1.5\sigma_R$, for a given reference sample size n_R, and selective appropriate k for achieving a desired power of $1 - \beta$ for detecting a meaningful difference of $\mu_T - \mu_R$ at a prespecificed level of significance α,assuming the true variability is σ_R.

Although the sample size adjustment approach proposed by Dong et al. [4] could be applied, the ratio between the reference product sample size and biosimilar product sample size n_R/n_T is usually proposed from 2/3 to 1.5, so that the information from one product sample won't dominate the information from the other product sample, and there is no loss of sample information, since no sample size adjustment will be needed. In practice, the number of Tier 1 CQAs is limited to be less than four and the number of biosimilar product lots recommended is often no less than 10. When there are more than one Tier 1 quality attributes, these Tier 1 quality attributes are usually correlated with each other, the power of passing multiple equivalence testings for multiple Tier 1 quality attributes becomes very complicated. The method proposed by Chow et al. was an approximation approach for the case when there is only one Tier 1 quality attribute. Furthermore, their approach may lead to very large, unbalanced sample sizes with less than 10 biosimilar lots. In Sect. 2, we present the explicit mathematical formula for the power function in the context of the two one-sided tests procedure, and the sample size requirement of the equivalence testing for a single Tier 1 quality attribute. In Sect. 3, we present the explicit mathematical formula for the power function of passing the equivalence testing, and the sample size requirement for two correlated quality attributes. The conclusion and discussion will be presented in Sect. 4.

2 Sample Size Requirement for the Equivalence Testing for a Single Tier 1 Quality Attribute

When there is only a single Tier 1 quality attribute to be considered, the power function of the hypothesis testing in (1) will be

$$P_1 = P\left(T_1 > t_{df^*,1-\alpha}, T_2 < -t_{df^*,1-\alpha} | \mu_T - \mu_R = \theta \in (-1.5\sigma_R, +1.5\sigma_R), \sigma_T, \sigma_R\right)$$

We will present two mathematical formulas below to calculate the power function P_1. The first one is the exact mathematical derivation of the power function as shown in Theorem 1, and the second one is an approximation method to estimate the power

function, and it is shown in Theorem 2. The proofs are presented in the Appendix 1 and Appendix 2, respectively.

Theorem 1 Given the two one-sided tests in (1) and test statistics in (2), let $A_1 = \frac{1.5\sigma_R-\theta}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} - t_{df^*,1-\alpha}\sqrt{\frac{\frac{x_1\sigma_T^2}{n_T^*(n_T-1)}+\frac{x_2\sigma_R^2}{n_R^*(n_R-1)}}{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}}$ and $B_1 = \frac{-1.5\sigma_R-\theta}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} + t_{df^*,1-\alpha}\sqrt{\frac{\frac{x_1\sigma_T^2}{n_T^*(n_T-1)}+\frac{x_2\sigma_R^2}{n_R^*(n_R-1)}}{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}}$, one explicit mathematical formula for the power function (denoted by P_1) is

$$P_1 = \int_0^{+\infty}\int_0^{+\infty}\{[\Phi(A_1)-\Phi(B_1)]\times f(x_1,x_2)\}\times \mathbf{I}\left\{\frac{x_1\sigma_T^2}{n_T^*(n_T-1)}+\frac{x_2\sigma_R^2}{n_R^*(n_R-1)} \le \frac{9\sigma_R^2}{4t^2_{df^*,1-\alpha}}\right\}dx_1dx_2 \tag{4}$$

where $f(x_1,x_2) = f(x_1)f(x_2)$ are the joint density of two independent chi-square distributions with $x_1 \sim \chi^2_{n_T-1}$ and $x_2 \sim \chi^2_{n_R-1}$. $\mathbf{I}\{.\}$ is the indication function to restrict the triangle area formed by x_1 and x_2.

Proof: See Appendix 1.

The above two-dimensional intergration over a triangle area, is the exact mathematical formula of the power function for the test (1), based on the test statistics (2). We can also simplify it to a univariate integral over a bounded interval, based on the Satterthwaite (1964) approximation summarized in theorem 2 below.

Theorem 2 Given the two one-sided tests in (1) and test statistics in (2), let $A_1' = \frac{1.5\sigma_R-\theta}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} - t_{df^*,1-\alpha}\sqrt{\frac{x}{df^*}}\sqrt{\frac{\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}}{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}}$ and $B_1' = \frac{-1.5\sigma_R-\theta}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} + t_{df^*,1-\alpha}\sqrt{\frac{x}{df^*}}\sqrt{\frac{\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}}{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}}$, let the integration limit $L = \frac{9\sigma_R^2 df^*}{4t^2_{df^*,1-\alpha}\left(\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}\right)}$, one approximated mathematical formula for the power function (denoted by P_1') is

$$P_1' = \int_0^{L}\left\{\left[\Phi\left(A_1'\right)-\Phi\left(B_1'\right)\right]\times\frac{1}{2^{\frac{df^*}{2}}\Gamma\left(\frac{df^*}{2}\right)}x^{\frac{df^*}{2}-1}e^{-\frac{x}{2}}\right\}dx \tag{5}$$

Proof: See Appendix 2.

Note that both P_1 and P_1' are functions of (i) overall significance level α, (ii) clinically or scientifically meaningful difference $\mu_T-\mu_R=\theta$, and (iii) the ratio of the reference and biosimilar variability σ_T/σ_R, and (iv) sample sizes n_R and n_T. We compare the calculated power values, based on the two mathematical formulas (4) and (5), by using R packages "**pracma**" for the two dimensional integration and R function "**integrate**" for the univariate integration separately, with the relative accuracy requested being no less than 10^{-10} in the results. Table 1 shows the numerical results,

Table 1 The comparison between the values of P_1 and $P_1^{'}$, assuming the fixed biosimilar sample size $n_T = 10$, $\alpha = 0.05$, $\theta = 1/8\sigma_R$. Three senarios of the variance ratio σ_T/σ_R were investigated

n_R	$\sigma_T/\sigma_R = 1$		$\sigma_T/\sigma_R = 1/\sqrt{2}$		$\sigma_T/\sigma_R = \sqrt{2}$	
	P_1	$P_1^{'}$	P_1	$P_1^{'}$	P_1	$P_1^{'}$
7	0.7705007	0.7704235	0.8607125	0.8609354	0.5659709	0.5665486
8	0.8164339	0.8164113	0.9047541	0.9046626	0.6060362	0.6062558
9	0.8489424	0.8489381	0.9324436	0.9326455	0.6362877	0.6363838
10	0.8727395	0.8727387	0.9509593	0.9511171	0.6598435	0.6598981
11	0.8906682	0.8906633	0.9636144	0.9636593	0.6786235	0.6787000
12	0.9045066	0.9044975	0.9723765	0.9723951	0.6939049	0.6940430
13	0.9154084	0.9154021	0.9786049	0.9786212	0.7065559	0.7067804
14	0.9241406	0.9241557	0.9831431	0.9831517	0.7173005	0.7175105
15	0.9312208	0.9312954	0.9865090	0.9865112	0.7264445	0.7266641

using the two mathematical formulas. With the assumption of fixed biosimilar sample size $n_T = 10$, significant level $\alpha = 0.05$, and the assumed difference $\theta = 1/8\sigma_R$, as recommended by FDA guidance [10], three senarios of the variability ratio between reference and biosimilar product were investigated. As expected, the power values are significantly impacted by the variability ratio σ_T/σ_R and the reference sample size n_R. Larger variability of the reference product or larger size of reference product would lead to higher power values and lower variability of reference product or lower size of reference product would lead to lower power values. From Table 1, the differences between the power values by P_1 and $P_1^{'}$ are not exceeding 0.00025, which are very small. In the following part of the paper, we will simply use $P_1^{'}$ as the power function for the equivalence testing for a single Tier 1 quality attribute.

Using the mathematical formulas, we can find out the minimal reference product sample size required (denote by n_R^{min}) for achieving the target power of passing the equivalence testing. The iterative algorithm is summarized as (i) set the target power value (for example 80%), (ii) fix the sample size n_T, the number of biosimilar product lots available, (iii) set the other parameters σ_T/σ_R, α, and θ, (iv) set the ratio between the reference product sample size and biosimilar product sample size n_R/n_T between 2/3 and 1.5 so that no sample size adjustment will be needed, (v) search the calculated power values starting from $n_R = 2n_T/3$ until the target power is achieved when $n_R = n_R^{min} \in [2n_T/3, 1.5n_T]$ or until $n_R > 1.5n_T$ which is the case that no n_R^{min} is found in the range $[2n_T/3, 1.5n_T]$. Using the above proposed algorithm, we can obtain the minimal reference sample size, n_R^{min} required to achieve the target power, 80% with the constraint that $2n_T/3 \leq n_R \leq 1.5n_T$,$\alpha = 0.05$, $\theta = 1/8\sigma_R$, and the actual power under different combinations of n_T and σ_T/σ_R. The results are shown in Table 2. From Table 2, it can be seen that n_R^{min} is smaller than, or equal to, n_T, for each fixed n_T, when reference product variability is larger than or equal to the

Table 2 Values of n_R^{min}, the minimal sample size required to achieve at least 80% power with the constraint of $2n_T/3 \leq n_R \leq 1.5n_T$,$\alpha = 0.05$, and $\theta = 1/8\sigma_R$ under different combinations of n_T and σ_T/σ_R

	(n_R^{min}, actual power)					
	$n_T = 10$	$n_T = 12$	$n_T = 15$	$n_T = 20$	$n_T = 25$	$n_T = 30$
$\sigma_T/\sigma_R = 1/\sqrt{2}$	(7, 0.861)	(8, 0.917)	(10, 0.969)	(14, 0.996)	(17, 0.999)	(20, >0.999)
$\sigma_T/\sigma_R = 1$	(8, 0.816)	(8, 0.850)	(10, 0.931)	(14, 0.984)	(17, 0.996)	(20, 0.999)
$\sigma_T/\sigma_R = \sqrt{2}$	$n_R^{min} > 1.5n_T$	(15, 0.809)	(10, 0.815)	(14, 0.932)	(17, 0.972)	(20, 0.989)

biosimilar product variability, and n_R^{min} is larger than or equal to n_T for each fixed n_T when biosimilar product variability is larger than reference product variability. It is noted that n_R^{min} may not be available within the range from $2n_T/3$to $1.5n_T$ for achieving the target power when biosimilar product variability is larger than reference product variability.

3 Sample Size Requirement for the Equivalence Testings for Two Correlated Tier 1 Quality Attributes

Most of times, there are more than one Tier 1 quality attributes in the biologics license application (BLA) submitted by the pharmaceutical industry. These attributes are usually correlated with each other. In the case of two Tier 1 quality attributes, we assume the two quality attributes in the reference product $X_{1,R}$ and $X_{2,R}$ have the same correlation coefficient value ρ as for the 2 quality attributes in the test product $X_{1,T}$ and $X_{2,T}$. which could be written as

$$\begin{bmatrix} X_{1,R} \\ X_{2,R} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu_{1,R} \\ \mu_{2,R} \end{bmatrix}, \begin{bmatrix} \sigma_{1,R}^2 & \rho\sigma_{1,R}\sigma_{2,R} \\ \rho\sigma_{1,R}\sigma_{2,R} & \sigma_{2,R}^2 \end{bmatrix}\right)$$
$$\begin{bmatrix} X_{1,T} \\ X_{2,T} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu_{1,T} \\ \mu_{2,T} \end{bmatrix}, \begin{bmatrix} \sigma_{1,T}^2 & \rho\sigma_{1,T}\sigma_{2,T} \\ \rho\sigma_{1,T}\sigma_{2,T} & \sigma_{2,T}^2 \end{bmatrix}\right)$$

Let $i = 1, 2$ represent the first ($i = 1$) or the second ($i = 2$) Tier 1 quality attribute. We further assume the sample size of either quality attribute is the same in the reference product or in the test product for the purpose of correlation, that is $n_{1,R} = n_{2,R} = n_R$ and $n_{1,T} = n_{2,T} = n_T$. The two equivalence testings are:

$$H_{0i} : \mu_{i.T} - \mu_{i.R} \leq -1.5\sigma_{i.R} \text{ or } \mu_{i.T} - \mu_{i.R} \geq +1.5\sigma_{i.R} \quad (6)$$

$$H_{ai} : -1.5\sigma_{i.R} < \mu_{i.T} - \mu_{i.R} < +1.5\sigma_{i,R}$$

where $\mu_{i.T}$ and $\mu_{i.R}$ are the population means for the ith quality attribute for the biosimilar and reference products respectively, $\sigma_{i.R}$ is the standard deviation of the reference product for the ith quality attribute. Both tests from (6) are rejected (each at a significant level α) when the test statistics $T_{i1} > t_{df_i^*,1-\alpha}$ and $T_{i2} \leq -t_{df_i^*,1-\alpha}$ where

$$T_{i1} = \frac{\left(\overline{X}_{i.T} - \overline{X}_{i.R}\right) + 1.5\sigma_{i.R}}{\sqrt{\frac{S_{i.T}^2}{n_T^*} + \frac{S_{i.R}^2}{n_R^*}}}; \quad T_{i2} = \frac{\left(\overline{X}_{i.T} - \overline{X}_{i.R}\right) - 1.5\sigma_{i.R}}{\sqrt{\frac{S_{i.T}^2}{n_T^*} + \frac{S_{i.R}^2}{n_R^*}}} \tag{7}$$

$\overline{X}_{i.T}$ and $\overline{X}_{i.R}$ are the biosimilar and reference sample means for the ith quality attribute respectively; $S_{i.T}^2$ and $S_{i.R}^2$ are the biosimilar and reference sample variances for the ith quality attribute respectively; $t_{df_i^*,1-\alpha}$ is the 100α-th percentile of the t-distribution with degrees of freedom df_i^*, which can be approximated by Satterthwaite (1964) approximation as $df_i^* = \left(\frac{\sigma_{i.T}^2}{n_T^*} + \frac{\sigma_{i.R}^2}{n_R^*}\right)^2 / \left(\frac{\sigma_{i.T}^4}{(n_T^*)^2(n_T-1)} + \frac{\sigma_{i.R}^4}{(n_R^*)^2(n_R-1)}\right)$. When there are two quality attributes, the probability of simultaneously rejecting 2 hypothesis tests in (6) is

$$P_2 = P_{\bigcap_{i=1,2}}\left(T_{i1} > t_{df_i^*,1-\alpha}, T_{i2} < -t_{df_i^*,1-\alpha} | \mu_{i.T} - \mu_{i.R} = \theta_i \in (-1.5\sigma_{i.R}, +1.5\sigma_{i.R}), \sigma_{i.T}, \sigma_{i.R}\right)$$

Note that in the test statistics T_{i1} and T_{i2}, $S_{1,T}^2$ and $S_{2,T}^2$ are not independent and neither are $S_{1,R}^2$ and $S_{2,R}^2$ due to the existence of correlation between the two quality attributes. The joint distribution for some transformations of $S_{1,T}^2$ and $S_{2,T}^2$ or $S_{1,R}^2$ and $S_{2,R}^2$ is a distribution called a bivariate Chi-square distribution which was also discussed in [6]. We also presented the bivariate Chi-square distribution from [6] here. Given $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, they are two dimensional independent random vectors where

$$\mathbf{X}_j = \begin{bmatrix} X_{1j} \\ X_{2j} \end{bmatrix} \sim \text{MVN}\left(\left[\theta, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right]\right)$$

Let the sample variances of the 2 variables are $S_1^2 = \sum_{j=1}^{N}(X_{1j} - \overline{X}_1)^2/(N-1)$ and $S_2^2 = \sum_{j=1}^{N}(X_{2j} - \overline{X}_2)^2/(N-1)$. The joint distribution of $u = (N-1)S_1^2/\sigma_1^2$ and $v = (N-1)S_2^2/\sigma_2^2$ is a bivariate Chi-square distribution with parameters ρ and $m = N-1$, and the density function is

$$f_2(u, v|\rho, m) = \frac{\left(1-\rho^2\right)^{-\frac{m}{2}}(uv)^{\frac{m-2}{2}}}{2^{m+1}\sqrt{\pi}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{m-1}{2}\right)}\exp\left\{-\frac{(u+v)}{2\left(1-\rho^2\right)}\right\} \times \sum_{k=0}^{\infty}\left\{\frac{1}{k!}\left(\frac{\rho\sqrt{uv}}{1-\rho^2}\right)^k\left(1+(-1)^k\right)\frac{\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{m-1}{2}\right)}{\Gamma\left(\frac{k+m}{2}\right)}\right\}$$

Theorem 3 let $u_T = \frac{(n_T-1)S_{1,T}^2}{\sigma_{1,T}^2} \sim \chi^2_{n_T-1}$, $v_T = \frac{(n_T-1)S_{2,T}^2}{\sigma_{2,T}^2} \sim \chi^2_{n_T-1}$ and $u_R = \frac{(n_R-1)S_{1,R}^2}{\sigma_{1,R}^2} \sim \chi^2_{n_R-1}$, $v_R = \frac{(n_R-1)S_{2,R}^2}{\sigma_{2,R}^2} \sim \chi^2_{n_R-1}$, then (u_T, v_T) is a bivariate Chi-square distribution with parameters ρ and $m = n_T - 1$; and (u_R, v_R) is a bivariate Chi-square distribution with parameters ρ and $m = n_R - 1$.

The proof of Theorem 3 is very straight forward by using the bivariate Chi-square distribution introduced above. Using the property of Theorem 3, we will present a mathematical formula below to calculate the power function P_2.

Theorem 4 Given the hypotheses in (6) and test statistics in (7), let $\tau_1 = \sqrt{\frac{\sigma_{1,T}^2}{n_T} + \frac{\sigma_{1,R}^2}{n_R}}$; $\tau_2 = \sqrt{\frac{\sigma_{2,T}^2}{n_T} + \frac{\sigma_{2,R}^2}{n_R}}$; $\rho^* = \left(\frac{\rho\sigma_{1,T}\sigma_{2,T}}{n_T} + \frac{\rho\sigma_{1,R}\sigma_{2,R}}{n_R}\right)/\sqrt{\left(\frac{\sigma_{1,T}^2}{n_T} + \frac{\sigma_{1,R}^2}{n_R}\right)\left(\frac{\sigma_{2,T}^2}{n_T} + \frac{\sigma_{2,R}^2}{n_R}\right)}$ and $f_1(x_1, x_2) = \frac{1}{2\pi\tau_1\tau_2\sqrt{1-\rho^{*2}}}\exp\left\{-\frac{x_1^2}{2\tau_1^2\left(1-\rho^{*2}\right)} + \frac{\rho^* x_1 x_2}{\tau_1\tau_2\left(1-\rho^{*2}\right)} - \frac{x_2^2}{2\tau_2^2\left(1-\rho^{*2}\right)}\right\}$ be the joint distribution of the bi-variate normal distribution with mean zero and variance-covariance matrix $\begin{bmatrix}\tau_1^2 & \rho^*\tau_1\tau_2 \\ \rho^*\tau_1\tau_2 & \tau_2^2\end{bmatrix}$. Let $U = \frac{\sigma_{1,T}^2 u_T}{n_T^*(n_T-1)} + \frac{\sigma_{1,R}^2 u_R}{n_R^*(n_R-1)} = \lambda_1 u_T + \mu_1 u_R$ and $V = \frac{\sigma_{2,T}^2 v_T}{n_T^*(n_T-1)} + \frac{\sigma_{2,R}^2 v_R}{n_R^*(n_R-1)} = \lambda_2 v_T + \mu_2\ v_R$, the limits $LL_1 = -1.5\sigma_{1,R} - \theta_1 + t_{df_1^*,1-\alpha}\sqrt{U}$, $LL_2 = -1.5\sigma_{2,R} - \theta_2 + t_{df_2^*,1-\alpha}\sqrt{V}$ and $UL_1 = 1.5\sigma_{1,R} - \theta_1 - t_{df_1^*,1-\alpha}\sqrt{U}$, $UL_2 = 1.5\sigma_{2,R} - \theta_2 - t_{df_2^*,1-\alpha}\sqrt{V}$; the integration limits $L_U = \frac{9\sigma_{1,R}^2}{4t^2_{df_1^*,1-\alpha}}$ and $L_V = \frac{9\sigma_{2,R}^2}{4t^2_{df_2^*,1-\alpha}}$. The mathematical formula for the power P_2 is

$$P_2 = \int_0^{L_V}\int_0^{L_U}\left\{\int_{LL_1(U)}^{UL_1(U)}\int_{LL_2(V)}^{UL_2(V)} f_1(x_1, x_2)dx_1dx_2\right\} \times \left\{\int_0^{\frac{V}{\lambda_2}}\int_0^{\frac{U}{\mu_1}} FF(U, u_R, v_T, V)du_R dv_T\right\}dUdV$$

In the integration, $F(U, u_R, v_T, V)$ is the joint distribution with two bivariate Chi-square distributions together inside as

$$FF(U, u_R, v_T, V) = f_2\left(\frac{U-\mu_1 u_R}{\lambda_1}, v_T|\rho, m = n_T - 1\right) \times f_2\left(u_R, \frac{V-\lambda_2 v_T}{\mu_2}|\rho, m = n_R - 1\right)\frac{1}{\lambda_1\mu_2}$$

Proof: See Appendix 3.

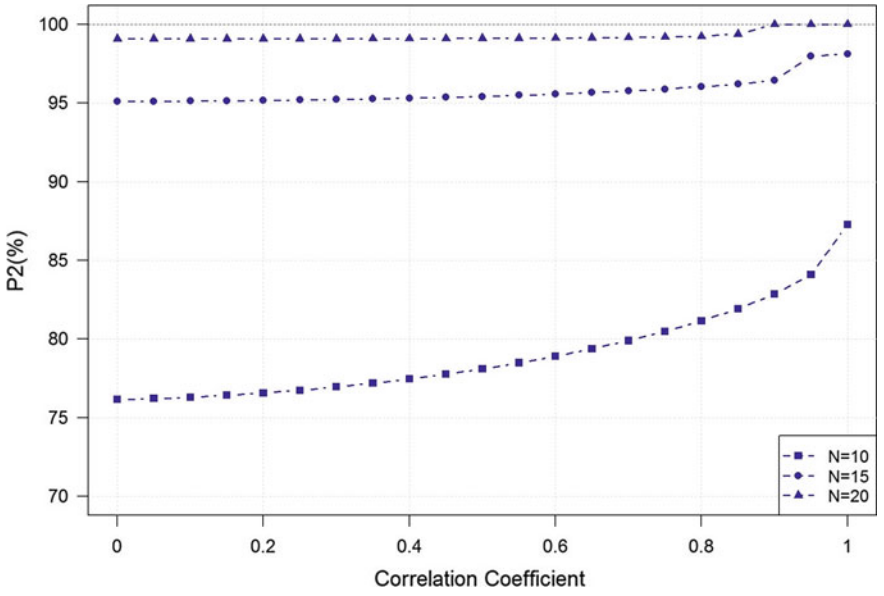


Fig. 1 The relationship between P_2 and the correlation coefficient ρ between two Tier 1 quality attributes. $\theta_1 = \sigma_{1.R}/8$, $\theta_2 = \sigma_{2.R}/8$, $\sigma_{1.T}/\sigma_{1.R} = \sigma_{2.T}/\sigma_{2.R} = 1$, $\alpha = 0.05$

The power P_2 is actually the probability of passing both equivalence testings when there are two Tier 1 quality attributes. It is a function of multiple parameters, including biosimilar and reference sample sizes, n_T and n_R, correlation coefficient between the two attributes ρ, the variability ratios between the biosimilar and reference products $\sigma_{1.T}/\sigma_{1.R}$ and $\sigma_{2.T}/\sigma_{2.R}$, the clinically or scientifically meaningful differences $\theta_1 = \mu_{1.T} - \mu_{1.R}$ and $\theta_2 = \mu_{2.T} - \mu_{2.R}$, and the significance level α for each individual hypothesis test. Based on theorem 4, the calculation of P_2 is a multi-dimensional intergration. Using the package "**mvtnorm**"and the "**Sparse Grid**" (Smolyak, [7] computational method, one could develop some R programming functions to calculate P_2, when parameter values are given. Figure 1 below illustrates the relationship between P_2 and ρ for different sample sizes $N = n_T = n_R$ under the assumption of $\theta_1 = \sigma_{1R}/8$, $\theta_2 = \sigma_{2R}/8$, the variability ratios between the biosimilar and reference products $\sigma_{1.T}/\sigma_{1.R} = \sigma_{2.T}/\sigma_{2.R} = 1$, and $\alpha = 0.05$. From Fig. 1, it can be seen that (i) P_2 increases from 76% to 88%, as ρ increases from 0 to 1 when $N = 10$. P_2 approaches the multiplication of the powers of two independent equivalence testing when $\rho \to 0$. On the other hand, when $\rho \to 1$, the two attributes are close to be unique and $P_2 \to P_1^{'}$. (ii) when sample size N is larger ($N = 15, 20$), the power of each individual equivalence testing $P_1^{'}$ is already large enough (> 0.95), the impact of ρ on P_2 becomes less profound.

By using the mathematical formulas in theorem 5, we can find out the minimal sample size (n_R^{min}) required for achieving the target power of passing both equivalence testings. For two correlated quality attributes, the iterative algorithm of

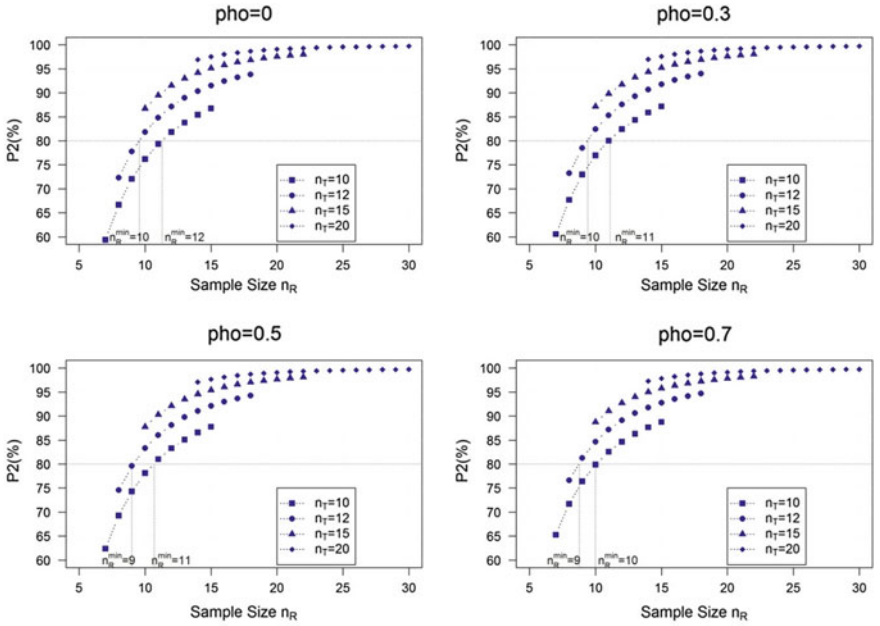


Fig. 2 The minimum reference sample size n_R^{min} required to at least 80% power at different correlation coefficients. $\theta_1 = \sigma_{1.R}/8$, $\theta_2 = \sigma_{2.R}/8$, $\sigma_{1.T}/\sigma_{1.R} = \sigma_{2.T}/\sigma_{2.R} = 1$, $\alpha = 0.05$

obtaining the minimum sample size n_R^{min} required is summarized as below: (i) set the target power value (for example 80%), (ii) fix the sample size n_T, the number of biosimilar product lots available, (iii) set the other parameters $\sigma_{1.T}/\sigma_{1.R}$, $\sigma_{2.T}/\sigma_{2.R}$, α, θ_1 and θ_2, (iv) set the ratio between the reference product sample size and biosimilar product sample size n_R/n_T between 2/3 and 1.5 so that no sample size adjustment will be needed, (v) search the calculated power values starting from $n_R = 2n_T/3$, until the target power is achieved when $n_R = n_R^{min} \in [2n_T/3, 1.5n_T]$, or until $n_R > 1.5n_T$, which is the case that no n_R^{min}, is to be found in the range $[2n_T/3, 1.5n_T]$. Figure 2 demonstrates the minimal reference sample size, n_R^{min}, required to achieve target power, 80%, at different senarios of biosimilar sample sizes n_T. It clearly shows that with smaller biosimilar sample sizes $n_T = 10, 12$, it requires larger reference sample size, n_R, when correlation coefficient values ρ is smaller; with larger biosimilar sample sizes $n_T = 15, 20$, there is no impact on P_2 to achieve the target power, 80%. This is also shown in the summarized results in Table 3.

Table 3 Values of n_R^{min}, the minimal sample size required to achieve at least 80% power with the constraint that $2n_T/3 \leq n_R \leq 1.5n_T$,$\alpha = 0.05$, $\theta_1 = \sigma_{1.R}/8$, $\theta_2 = \sigma_{2.R}/8$ and $\sigma_{1.T}/\sigma_{1.R} = \sigma_{2.T}/\sigma_{2.R} = 1$ under different combinations of ρ and n_T

	(n_R^{min}, actual power)					
	$n_T = 10$	$n_T = 12$	$n_T = 15$	$n_T = 20$	$n_T = 25$	$n_T = 30$
$\rho = 0$	(12, 0.82)	(10, 0.82)	(10, 0.87)	(14, 0.97)	(17, 0.99)	(20, >0.99)
$\rho = 0.10$	(12, 0.82)	(10, 0.82)	(10, 0.87)	(14, 0.97)	(17, 0.99)	(20, >0.99)
$\rho = 0.30$	(11, 0.80)	(10, 0.82)	(10, 0.87)	(14, 0.97)	(17, 0.99)	(20, >0.99)
$\rho = 0.50$	(11, 0.81)	(9, 0.80)	(10, 0.88)	(14, 0.97)	(17, 0.99)	(20, >0.99)
$\rho = 0.70$	(10, 0.80)	(9, 0.81)	(10, 0.89)	(14, 0.97)	(17, 0.99)	(20, >0.99)

4 Discussion and Comments

In this paper, we introduce the sparse grid method for the computation of passing two equivalence testings when two correlated Tier 1 quality attributes are presented. The sparse grid method is a general numerical discretization technique for multivariate problems. This approach, first introduced by the Russian mathematician Smolyak [7], constructs a multidimensional multilevel basis by a special truncation of the tensor product expansion of a one-dimensional multilevel basis. Figure 3 presents a regular two-dimensional sparse grid, which is different from Monte Carlo simulation. The simulation points at each dimension was not random, and each point has a different weight at averaging the final value. Please see the exact statistical techniques in Smolyak [7].

Another question is how we should estimate the value of the correlations during the computation. It is recommended that the correlation coefficient should be estimated from the historical test values or from other biosimilar products which have the same reference products. On the other hand, as we have seen from this paper, when the reference and biosimilar sample sizes get larger, the correlation coefficient, ρ, between the attributes would only slightly impact the power of passing multiple equivalence testing under the given assumed parameters, as discussed above. It is always recommended that the sponsors get enough batches of reference and biosimilar products for satisfying the target power. When in some cases, there are three or more Tier 1 quality attributes, the mathematical formulas and computation become even more complicated.

In this paper, we theoretically derived the analytical power calculation formula $\left(P_1^{'}\text{and } P_2\right)$. The R programing functions could also be introduced to get the power values efficiently when needed parameters are given. By keeping the biosimilar sample size $n_T \geq 10$, one can find the reference sample size n_R required to achieve target power (say 80% or 90%) to pass one or both of the equivalence testing result(s), using our derived formulas.

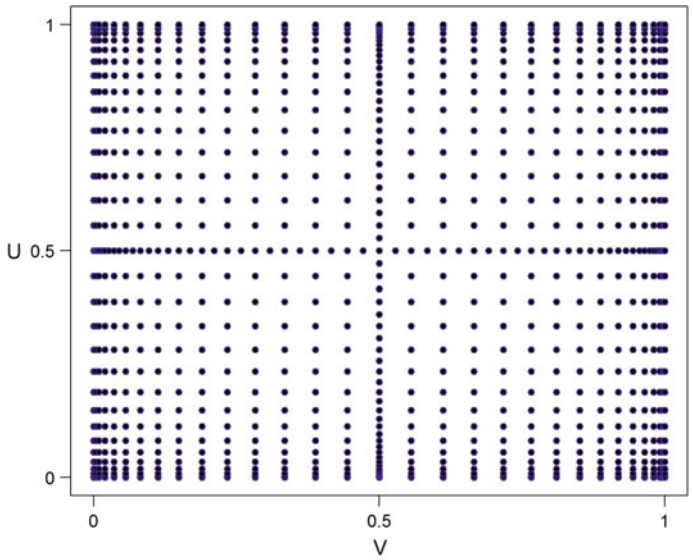


Fig. 3 A regular two-dimensional sparse grid

The current practice of assuming the equivalence are constant despite being a function of the standard deviation, determined by the study data, without taking the estimation error into account. The biasness of the current practice were discussed in Burdict et al. [1], Dong et al. [5], and [2]. Statistical approaches on the parameter margin of the statistical equivalence testing have been proposed and disused in Burdict et al. [1], Dong et al. [5], [2] and Weng et al. [11]. Further sample size development of the proposed approaches will be presented in a separate paper.

Appendix

1. **The proof of Theorem 1.**

Proof: Because the power function is

$$P_1 = P\left(T_1 = \frac{(\overline{X}_T - \overline{X}_R) + 1.5\sigma_R}{\sqrt{\frac{S_T^2}{n_T^*} + \frac{S_R^2}{n_R^*}}} > t_{df^*,1-\alpha}, T_2 = \frac{(\overline{X}_T - \overline{X}_R) - 1.5\sigma_R}{\sqrt{\frac{S_T^2}{n_T^*} + \frac{S_R^2}{n_R^*}}} < -t_{df^*,1-\alpha} \right) \tag{A.1}$$

given $\mu_T - \mu_R = \theta \in (-1.5\sigma_R, +1.5\sigma_R)$, σ_T, σ_R. The inequailities in (A.1) is equivalent to

$$-1.5\sigma_R + t_{df^*,1-\alpha}\sqrt{\frac{S_T^2}{n_T^*} + \frac{S_R^2}{n_R^*}} < \left(\overline{X}_T - \overline{X}_R\right) < 1.5\sigma_R - t_{df^*,1-\alpha}\sqrt{\frac{S_T^2}{n_T^*} + \frac{S_R^2}{n_R^*}} \quad (A.2)$$

Because $Z = \frac{(\overline{X}_T - \overline{X}_R - \theta)}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} \sim N(0, 1)$, the inequailities in (A.2) is equivalent to

$$\frac{(-1.5\sigma_R - \theta)}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} + \frac{t_{df^*,1-\alpha}\sqrt{\frac{S_T^2}{n_T^*} + \frac{S_R^2}{n_R^*}}}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} < Z < \frac{(1.5\sigma_R - \theta)}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} - \frac{t_{df^*,1-\alpha}\sqrt{\frac{S_T^2}{n_T^*} + \frac{S_R^2}{n_R^*}}}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} \quad (A.3)$$

Let $x_1 = \frac{(n_T-1)S_T^2}{\sigma_T^2} \sim \chi^2_{n_T-1}$ and $x_2 = \frac{(n_R-1)S_R^2}{\sigma_R^2} \sim \chi^2_{n_R-1}$, then $S_T^2 = \frac{x_1\sigma_T^2}{(n_T-1)}$ and $S_R^2 = \frac{x_2\sigma_R^2}{(n_R-1)}$, the inequalities in (A.3) is

$$\frac{(-1.5\sigma_R - \boldsymbol{\theta})}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} + t_{df^*,1-\alpha}\frac{\sqrt{\frac{x_1\sigma_T^2}{n_T^*(n_T-1)} + \frac{x_2\sigma_R^2}{n_R^*(n_R-1)}}}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} < Z < \frac{(1.5\sigma_R - \theta)}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} - t_{df^*,1-\alpha}\frac{\sqrt{\frac{x_1\sigma_T^2}{n_T^*(n_T-1)} + \frac{x_2\sigma_R^2}{n_R^*(n_R-1)}}}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} \quad (A.4)$$

Let $A_1 = \frac{1.5\sigma_R - \theta}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} - t_{df^*,1-\alpha}\sqrt{\frac{\frac{x_1\sigma_T^2}{n_T^*(n_T-1)} + \frac{x_2\sigma_R^2}{n_R^*(n_R-1)}}{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}}$ and $B_1 = \frac{-1.5\sigma_R - \theta}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} + t_{df^*,1-\alpha}\sqrt{\frac{\frac{x_1\sigma_T^2}{n_T^*(n_T-1)} + \frac{x_2\sigma_R^2}{n_R^*(n_R-1)}}{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}}$,

From (A.4) since $B_1 < A_1$, we can get that $\frac{x_1\sigma_T^2}{n_T^*(n_T-1)} + \frac{x_2\sigma_R^2}{n_R^*(n_R-1)} < \frac{9\sigma_R^2}{4t^2_{df^*,1-\alpha}}$, then the power function (A.1) is

$$E_{x_1,x_2}\{P(B_1 < Z < A_1)\}$$

Given $x_1 \sim \chi^2_{n_T-1}$ and $x_2 \sim \chi^2_{n_R-1}$, the power function is

$$P_1 = \int_0^{+\infty}\int_0^{+\infty}\{[\Phi(A_1) - \Phi(B_1)] \times f(x_1, x_2)\} \times \mathbf{I}\left\{\frac{x_1\sigma_T^2}{n_T^*(n_T-1)} + \frac{x_2\sigma_R^2}{n_R^*(n_R-1)} \le \frac{9\sigma_R^2}{4t^2_{df^*,1-\alpha}}\right\}dx_1dx_2$$

where $f(x_1, x_2) = f(x_1)f(x_2)$ is the density function of two independent Chi-square distributions with $x_1 \sim \chi^2_{n_T-1}$ and $x_2 \sim \chi^2_{n_R-1}$. $\mathbf{I}\{.\}$ is the indication function to restrict the triangle area formed by x_1 and x_2.

2. **The proof of Theorem 2.**

Proof: By using the Satterthwaite approximation

$$\frac{df^*\left(\frac{S_T^2}{n_T^*}+\frac{S_R^2}{n_R^*}\right)}{\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}} \sim \chi^2_{df^*} \tag{A.5}$$

Where $df^* = \left(\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}\right)^2 / \left(\frac{\sigma_T^4}{(n_T^*)^2(n_T-1)}+\frac{\sigma_R^4}{(n_R^*)^2(n_R-1)}\right)$, let $S^2 = \frac{S_T^2}{n_T^*}+\frac{S_R^2}{n_R^*}$ and $x = \frac{df^* S^2}{\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}} \sim \chi^2_{df^*}$, then the inequalities (A.3) is

$$\frac{(-1.5\sigma_R-\theta)}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} + t_{df^*,1-\alpha}\sqrt{\frac{x}{df^*}}\frac{\sqrt{\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}}}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} < Z < \frac{(1.5\sigma_R-\theta)}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} - t_{df^*,1-\alpha}\sqrt{\frac{x}{df^*}}\frac{\sqrt{\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}}}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} \tag{A.6}$$

Let $A_1' = \frac{1.5\sigma_R-\theta}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} - t_{df^*,1-\alpha}\sqrt{\frac{x}{df^*}}\sqrt{\frac{\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}}{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}}$ and $B_1' = \frac{-1.5\sigma_R-\theta}{\sqrt{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}} + t_{df^*,1-\alpha}\sqrt{\frac{x}{df^*}}\sqrt{\frac{\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}}{\frac{\sigma_T^2}{n_T}+\frac{\sigma_R^2}{n_R}}}$, using $B_1' < A_1'$, we can get $x < L = \frac{9\sigma_R^2 df^*}{4t^2_{df^*,1-\alpha}\left(\frac{\sigma_T^2}{n_T^*}+\frac{\sigma_R^2}{n_R^*}\right)}$, the power function (A.1) is $E_x\left\{P(B_1' < Z < A_1')\right\}$, which is equivalent to

$$P_1' = \int_0^L \left\{\left[\Phi\left(A_1'\right)-\Phi\left(B_1'\right)\right]\times\frac{1}{2^{\frac{df^*}{2}}\Gamma\left(\frac{df^*}{2}\right)}x^{\frac{df^*}{2}-1}e^{-\frac{x}{2}}\right\}dx$$

3. **The proof of Theorem 4.**

Proof: If there are 2 quality attributes that are correlated, assume the 2 quality attributes in the reference product $X_{1,R}$ and $X_{2,R}$ have the correlation value ρ; the same as the 2 quality attributes in the test product $X_{1,T}$ and $X_{2,T}$

$$\begin{bmatrix} X_{1,R} \\ X_{2,R} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_{1,R} \\ \mu_{2,R} \end{bmatrix}, \begin{bmatrix} \sigma_{1,R}^2 & \rho\sigma_{1,R}\sigma_{2,R} \\ \rho\sigma_{1,R}\sigma_{2,R} & \sigma_{2,R}^2 \end{bmatrix}\right)$$
$$\begin{bmatrix} X_{1,T} \\ X_{2,T} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_{1,T} \\ \mu_{2,T} \end{bmatrix}, \begin{bmatrix} \sigma_{1,T}^2 & \rho\sigma_{1,T}\sigma_{2,T} \\ \rho\sigma_{1,T}\sigma_{2,T} & \sigma_{2,T}^2 \end{bmatrix}\right)$$

We further assume the sample size of either quality attribute is the same in the reference product or in the test product, that is $n_{1,R} = n_{2,R} = n_R$ and $n_{1,T} = n_{2,T} = n_T$, given the hypotheses in (6) and test statistics in (7), the power of passing both equivalence testings is equivalent to calculate

$$P\left(\begin{bmatrix} -1.5\sigma_{1,R}+t_{df_1^*,1-\alpha}\sqrt{\frac{S_{1,T}^2}{n_T^*}+\frac{S_{1,R}^2}{n_R^*}} \\ -1.5\sigma_{2,R}+t_{df_2^*,1-\alpha}\sqrt{\frac{S_{2,T}^2}{n_T^*}+\frac{S_{2,R}^2}{n_R^*}} \end{bmatrix} \le \begin{bmatrix} \overline{X}_{1.T}-\overline{X}_{1.R} \\ \overline{X}_{2.T}-\overline{X}_{2.R} \end{bmatrix} \le \begin{bmatrix} +1.5\sigma_{1,R}-t_{df_1^*,1-\alpha}\sqrt{\frac{S_{1,T}^2}{n_T^*}+\frac{S_{1,R}^2}{n_R^*}} \\ +1.5\sigma_{2,R}-t_{df_2^*,1-\alpha}\sqrt{\frac{S_{2,T}^2}{n_T^*}+\frac{S_{2,R}^2}{n_R^*}} \end{bmatrix}\right) \tag{A.9}$$

The degrees of freedom $df_1^* = \left(\frac{\sigma_{1,T}^2}{n_T^*} + \frac{\sigma_{1,R}^2}{n_R^*}\right)^2 / \left(\frac{\sigma_{1,T}^4}{(n_T^*)^2(n_T-1)} + \frac{\sigma_{1,R}^4}{(n_R^*)^2(n_R-1)}\right)$ and $df_2^* = \left(\frac{\sigma_{2,T}^2}{n_T^*} + \frac{\sigma_{2,R}^2}{n_R^*}\right)^2 / \left(\frac{\sigma_{1,T}^4}{(n_T^*)^2(n_T-1)} + \frac{\sigma_{1,R}^4}{(n_R^*)^2(n_R-1)}\right)$. Let $u_T = \frac{(n_T-1)S_{1,T}^2}{\sigma_{1,T}^2} \sim \chi^2_{n_T-1}$, $v_T = \frac{(n_T-1)S_{2,T}^2}{\sigma_{2,T}^2} \sim \chi^2_{n_T-1}$ and $u_R = \frac{(n_R-1)S_{1,R}^2}{\sigma_{1,R}^2} \sim \chi^2_{n_R-1}$, $v_R = \frac{(n_R-1)S_{2,R}^2}{\sigma_{2,R}^2} \sim \chi^2_{n_R-1}$, and $\theta_1 = \mu_{1,T} - \mu_{1,R}$, $\theta_2 = \mu_{2,T} - \mu_{2,R}$. The inequailities in (A.9) is equivalent to

$$\begin{bmatrix} -1.5\sigma_{1,R} - \theta_1 + t_{df_1^*,1-\alpha}\sqrt{\frac{\sigma_{1,T}^2 u_T}{n_T^*(n_T-1)} + \frac{\sigma_{1,R}^2 u_R}{n_R^*(n_R-1)}} \\ -1.5\sigma_{2,R} - \theta_2 + t_{df_2^*,1-\alpha}\sqrt{\frac{\sigma_{2,T}^2 v_T}{n_T^*(n_T-1)} + \frac{\sigma_{2,R}^2 v_R}{n_R^*(n_R-1)}} \end{bmatrix} \le \mathbf{B}$$

$$\le \begin{bmatrix} +1.5\sigma_{1,R} - \theta_1 - t_{df_1^*,1-\alpha}\sqrt{\frac{\sigma_{1,T}^2 u_T}{n_T^*(n_T-1)} + \frac{\sigma_{1,R}^2 u_R}{n_R^*(n_R-1)}} \\ +1.5\sigma_{2,R} - \theta_2 - t_{df_2^*,1-\alpha}\sqrt{\frac{\sigma_{2,T}^2 v_T}{n_T^*(n_T-1)} + \frac{\sigma_{2,R}^2 v_R}{n_R^*(n_R-1)}} \end{bmatrix} \tag{A.10}$$

where **B** is a bi-variate normal distribution with mean 0 and variance-covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \frac{\sigma_{1,T}^2}{n_T} + \frac{\sigma_{1,R}^2}{n_R} & \frac{\rho\sigma_{1,T}\sigma_{2,T}}{n_T} + \frac{\rho\sigma_{1,R}\sigma_{2,R}}{n_R} \\ \frac{\rho\sigma_{1,T}\sigma_{2,T}}{n_T} + \frac{\rho\sigma_{1,R}\sigma_{2,R}}{n_R} & \frac{\sigma_{2,T}^2}{n_T} + \frac{\sigma_{2,R}^2}{n_R} \end{bmatrix}$$

let $\tau_1 = \sqrt{\frac{\sigma_{1,T}^2}{n_T} + \frac{\sigma_{1,R}^2}{n_R}}$; $\tau_2 = \sqrt{\frac{\sigma_{2,T}^2}{n_T} + \frac{\sigma_{2,R}^2}{n_R}}$; $\rho^* = \frac{\rho\sigma_{1,T}\sigma_{2,T}}{n_T} + \frac{\rho\sigma_{1,R}\sigma_{2,R}}{n_R} / \sqrt{\left(\frac{\sigma_{1,T}^2}{n_T} + \frac{\sigma_{1,R}^2}{n_R}\right)\left(\frac{\sigma_{2,T}^2}{n_T} + \frac{\sigma_{2,R}^2}{n_R}\right)}$, then **B** is a bi-variate normal distribution with mean 0 and variance-covariance matrix $\begin{bmatrix} \tau_1^2 & \rho^*\tau_1\tau_2 \\ \rho^*\tau_1\tau_2 & \tau_2^2 \end{bmatrix}$. The power function would be

$$P_2 = E_{u_T, v_T, u_R, v_R}\{P(\textbf{Inequality}(A.10)\,\text{Holds} \mid u_T, v_T, u_R, v_R)\}$$

Let $U = \frac{\sigma_{1,T}^2 u_T}{n_T^*(n_T-1)} + \frac{\sigma_{1,R}^2 u_R}{n_R^*(n_R-1)} \le \frac{9\sigma_{1,R}^2}{4t^2_{df_1^*,1-\alpha}}$ and $V = \frac{\sigma_{2,T}^2 v_T}{n_T^*(n_T-1)} + \frac{\sigma_{2,R}^2 v_R}{n_R^*(n_R-1)} \le \frac{9\sigma_{2,R}^2}{4t^2_{df_2^*,1-\alpha}}$ by the inequalities (A.10), the limits $LL_1 = -1.5\sigma_{1,R} - \theta_1 + t_{df_1^*,1-\alpha}\sqrt{U}$, $LL_2 = -1.5\sigma_{2,R} - \theta_2 + t_{df_2^*,1-\alpha}\sqrt{V}$ and $UL_1 = 1.5\sigma_{1,R} - \theta_1 - t_{df_1^*,1-\alpha}\sqrt{U}$, $UL_2 = 1.5\sigma_{2,R} - \theta_2 - t_{df_2^*,1-\alpha}\sqrt{V}$; then

$$P_2 = E_{U,V}\left\{\int_{LL_1}^{UL_1}\int_{LL_2}^{UL_2} \frac{1}{2\pi\tau_1\tau_2\sqrt{1-\rho^{*2}}}\exp\left\{-\frac{x_1^2}{2\tau_1^2(1-\rho^{*2})} + \frac{\rho^* x_1 x_2}{\tau_1\tau_2(1-\rho^{*2})} - \frac{x_2^2}{2\tau_2^2(1-\rho^{*2})}\right\}dx_1 dx_2\right\}$$

The problem remains to find the joint distribution of U and V. Please note that from theorem 4, $F(u_T, u_R, v_T, v_R) = f_2(u_T, v_T) * f_2(u_R, v_R)$ where f_2 is the joint density

of a bivariate chi-square distribution. Let $\frac{\sigma_{1,T}^2 u_T}{n_T^*(n_T-1)} + \frac{\sigma_{1,R}^2 u_R}{n_R^*(n_R-1)} = \lambda_1 u_T + \mu_1 u_R = U$ and $\frac{\sigma_{2,T}^2 v_T}{n_T^*(n_T-1)} + \frac{\sigma_{2,R}^2 v_R}{n_R^*(n_R-1)} = \lambda_2 v_T + \mu_2 v_R = V$, and we want to get the joint distribution of U and V. The Jacobian transformation matrix is

$$\text{Jacobian}\left[\left(u_T = \frac{U-\mu_1 u_R}{\lambda_1}, u_R, v_T, v_R = \frac{V-\lambda_2 v_T}{\mu_2}\right) \to (U, u_R, v_T, V)\right] = \frac{1}{\lambda_1\mu_2}$$

Thus

$$\begin{aligned} FF(U, u_R, v_T, V) &= F\left(\frac{U-\mu_1 u_R}{\lambda_1}, u_R, v_T, \frac{V-\lambda_2 v_T}{\mu_2}\right)\frac{1}{\lambda_1\mu_2} \\ &= f_2\left(\frac{U-\mu_1 u_R}{\lambda_1}, v_T | m = n_T - 1\right) * f_2\left(u_R, \frac{V-\lambda_2 v_T}{\mu_2} | m = n_R - 1\right)\frac{1}{\lambda_1\mu_2} \end{aligned}$$

We then integrate out u_R, v_T, we can get the joint distribution of U and V

$$F_2(U, V) = \int_0^{\frac{V}{\lambda_2}} \int_0^{\frac{U}{\mu_1}} FF(U, u_R, v_T, V) du_R dv_T$$

We can get the power function as

$$P_2 = \int_0^{\frac{9\sigma_{2,R}^2}{4t^2_{df_2^*,1-\alpha}}} \int_0^{\frac{9\sigma_{1,R}^2}{4t^2_{df_1^*,1-\alpha}}} \left\{ \int_{LL_1(U)}^{UL_1(U)} \int_{LL_2(V)}^{UL_2(V)} f_1(x_1, x_2) dx_1 dx_2 \right\} \times F_2(U, V) dU dV$$

where $f_1(x_1, x_2) = \frac{1}{2\pi\tau_1\tau_2\sqrt{1-\rho^{*2}}} \exp\left\{ -\frac{x_1^2}{2\tau_1^2(1-\rho^{*2})} + \frac{\rho^* x_1 x_2}{\tau_1\tau_2(1-\rho^{*2})} - \frac{x_2^2}{2\tau_2^2(1-\rho^{*2})} \right\}$ be the joint distribution of the bi-variate normal distribution with mean zero and variance-covariance matrix $\begin{bmatrix} \tau_1^2 & \rho^*\tau_1\tau_2 \\ \rho^*\tau_1\tau_2 & \tau_2^2 \end{bmatrix}$.

References

1. Burdick, R.K., Thomas, N., Cheng, A.: Statistical considerations in demonstrating cmc analytical similarity for a biosimilar product. Stat. Biopharm. Res. **9**(3), 249–257 (2017)
2. Chen, Y.M., Weng, Y.T., Dong, X., Tsong, Y.: Wald tests for variance-adjusted equivalence assessment with normal endpoints. J. Biopharm. Stat. **27**(2), 308–316 (2017)
3. Chow, S.C., Song, F., Bai, H.: Sample size requirement in analytical studies for similarity assessment. J. Biopharm. Stat. **27**(2), 233–238 (2017)
4. Dong, X., Bian, Y., Tsong, Y., Wang, T.: Exact test-based approach for equivalence test with parameter margin. J. Biopharm. Stat. **27**(2), 317–330 (2017)
5. Dong, X., Weng, Y.T., Tsong, Y.: Adjustment for unbalanced sample size for analytical biosimilar equivalence assessment. J. Biopharm. Stat. **27**(2), 220–232 (2017)
6. Joarder, A.H.: Moments of the product and ratio of two correlated chi-square variables. Stat. Pap. **50**(3), 581–592 (2009)

7. Smolyak, S.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Soviet Math. Dokl. **4**, 240–243. American Mathematical Society (1963)
8. Tsong, Y., Dong, X., Shen, M.: Development of statistical methods for analytical similarity assessment. J. Biopharm. Stat. **27**(2), 197–205 (2017)
9. US Food and Drug Administration.: Guidance for industry: scientific considerations in demonstrating biosimilarity to a reference product. Silver Spring, MD., US Food and Drug Administration (2015)
10. US Food and Drug Administration.: Guidance for industry: Statistical Approaches to Evaluate Analytical Similarity. Silver Spring, MD., US Food and Drug Administration (2017)
11. Weng, Y.T., Tsong, Y., Shen, M., and Wang, C.: Improved Wald test for equivalence assessment of analytical biosimilarity. Int. J. of Clinical Biostat. And Biometrics. (2018) Submitted

A Probability Based Equivalence Test of NIR Versus HPLC Analytical Methods in a Continuous Manufacturing Process Validation Study

Areti Manola, Steven Novick, Jyh-Ming Shoung and Stan Altan

Abstract Continuous manufacturing processes rely on Process Analytical Technology (PAT) and chemometric Near Infrared (NIR) technologies to carry out real time release testing (RTRt). A critical requirement for this purpose is to establish the equivalence between the NIR analytical method with the gold standard analytical method, say an HPLC method. We propose a variance components model that acknowledges the inherent blocking across individual dosage units through a paired comparison. Variance terms corresponding to dosage unit, location effects due to a stratified sampling plan and heterogeneous residual terms provide estimates of the total measurement uncertainty in both methods free of dosage unit effects. Bayesian posterior parameter estimates and the posterior predictive distribution are used to assess the performance of the NIR method in relation to the HPLC gold standard method as a measure of equivalence, referred to as a Relative Performance Index (Rel_Pfm). An acceptably high probability of a Rel_Pfm of 1 (or greater) is proposed as the essential requirement for establishing equivalence (or superiority).

Keywords Continuous manufacture · Bayesian mixed model · Equivalence test · Comparison of analytical methods

1 Introduction

Traditionally, pharmaceutical manufacturing has been carried out using the standard batch process. This involves multiple steps as well as downstream end-product quality testing. Product is produced in discrete, fixed sizes, irrespective of market demand. It is a method of manufacture that has changed little over the past half century or longer. In contrast, continuous manufacturing (CM) works in a virtu-

A. Manola (✉) · J.-M. Shoung · S. Altan
Janssen Pharmaceutical R&D, Raritan, NJ 08869, USA
e-mail: amanola@its.jnj.com

S. Novick
Medimmune, Gaithersburg, MD, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_4

ally uninterrupted, sequential process with little delay between operation steps, no downtime with product sitting in queue and real-time quality testing throughout the process [2]. Work in process-inventory control is essentially eliminated. CM also allows quality testing to be performed during the production process and real-time decisions to be made based on the resulting data. This means that virtually all end-product testing can be eliminated, and the benefits accrued to the manufacturer, and ultimately, the consumer, are enormous. To achieve the goal of Real Time Release testing (RTRt), Process Analytical Testing (PAT) technologies, based on chemometric near infrared (NIR) methods, must be validated, and shown sufficiently accurate, and precise. Our objective here is to discuss the connections between experimental design, modeling, and inferential methodology in order to verify the comparability between an NIR analytical method, with the gold standard HPLC analytical method, as part of Stage 2 Process Validation. Equivalence of the two methods is related to the probability of falling within a desirable interval of the true, but unknown value of the analyte, accounting for incipient tablet-to-tablet variation.

2 HPLC—NIR Calibration During Process Design

It is outside the scope of this paper to discuss the details of how an NIR calibration model is developed during the Process Design stage. The current design, with 3 concentration levels at target tablet weight, is sufficient to establish the calibration curve, using a Partial Least Squares (PLS) methodology. The 3-point design would be the minimal design for establishing linearity and testing for curvature. Multivariate performance measures of the PLS model are useful for tracking the robustness of the model over the lifecycle of the product. Following the establishment of the calibration model, a Gage R&R design can be used to assess analytical performance and comparability with the HPLC method during the early stages of development. Any experimental design used needs to acknowledge the blocking at the individual tablet level, hence leading to a paired comparison where the natural pairing of NIR versus HPLC at the tablet level is the block. This is a consequence of the NIR methodology, which allows nondestructive testing. The natural pairing and blocking at the tablet level permits a straightforward comparison, and makes possible a major paradigm shift, from comparisons at the method mean level, to the level of the tablet. In subsequent sections, we will propose a ***Relative Performance Index,*** which exploits the natural pairing of NIR to HPLC at the tablet level, by eliminating the Tablet variance component. The paired design permits probability assessment of an individual analytical determination (i.e. tablet content), falling within a pre-specified limit of its true value for both NIR and HPLC methods. The probability assessment is carried out through a Bayesian posterior predictive calculation. It is important for the statistical model to acknowledge HPLC analytical run design. How best to include the effect of multiple HPLC runs in the comparison with the NIR method which has no multiple run effects, is an open research question.

The proposed method will be compared with a common equivalence measure at the method–mean level, the Two One-Sided Test [7] or TOST. The TOST test proposes the following null and alternative hypotheses:

H_0: $\delta < \Delta_L$ or $\delta > \Delta_U$

versus

H_1: $\Delta_L \leq \delta \leq \Delta_U$

where Δ_L and Δ_U are the lower and upper bounds establishing the criterion for equivalence. Typically, $|\Delta_L| = |\Delta_U|$ and δ represents the difference between the method means. One can understand δ as an estimate of the bias if a gold standard method is being used. The equivalence test is conducted by performing the two following tests separately:

$H_{0,1}$: $\delta < \Delta_L$

$H_{1,1}$: $\delta \geq \Delta_L$

and

$H_{0,2}$: $\delta > \Delta_U$

$H_{1,2}$: $\delta \leq \Delta_U$.

If p_1 and p_2 are the p-values for the two tests, the overall p-value is $\max(p_1, p_2)$. Equivalence is claimed if the $100(1-2\alpha)\%$ confidence interval for δ is completely contained within the interval (Δ_L, Δ_U).

3 Relative Performance Index

We assume the HPLC method is the gold standard method and express the probability of a single analytical determination y from the HPLC (or NIR) method falling within some interval Δ of the true value μ as follows:

$$\Pr\_H = P(|y-\mu|) \leq \Delta|\text{HPLC}) = \Phi\left(\frac{\Delta}{\sigma_{HPLC}}\right) - \Phi\left(\frac{-\Delta}{\sigma_{HPLC}}\right)$$

$$\Pr\_H = P(|y-\mu|) \leq \Delta|\text{NIR}) = \Phi\left(\frac{\Delta - bias}{\sigma_{NIR}}\right) - \Phi\left(\frac{-\Delta - bias}{\sigma_{NIR}}\right)$$

where $\Phi(\bullet)$ is the CDF of standard normal distribution and *bias* is the difference between the gold standard HPLC method and the NIR method in relation to their expected values, $E(y_{NIR}) = \mu + bias$.

We define a ***Relative Performance Index (Rel_Pfm)*** as follows:

$$\text{Rel\_Pfm} = \Pr\_N / \Pr\_H$$

The practical interpretation of the relative performance index arises from its construction. It is the ratio of two probabilities. The numerator is the probability of a random analytical determination, falling within Δ units of its true value by the NIR method. The denominator is the corresponding probability by the HPLC method. A

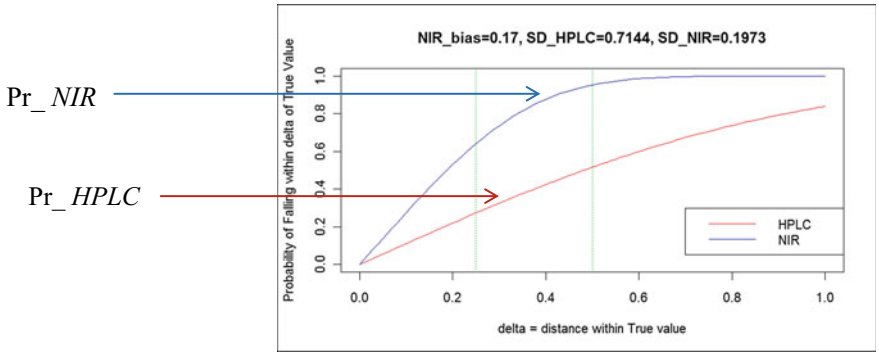


Fig. 1 OC curves of the probability of falling within delta

value of 1 or higher implies that the NIR method is comparable to, or superior to, the HPLC method.

4 Method Comparison Using the Relative Performance Index

Given our definition for the relative performance index, and values for ***delta***, ***bias*** and ***method variability,*** it is straightforward to calculate OC curves of the probability of falling within delta units of the true value for both methods. An example of such an OC curve is given in Fig. 1.

Another interesting feature is that it is possible to plot the relative performance index across increasing values of delta as shown in Fig. 2.

We propose that the criterion for equivalence be defined as $\Pr(\text{Rel\_Pfm} \geq 1) \geq P_C$, where P_C is a desired probability level. The choice of P_C is a scientific judgment call. We propose that a $P_C = 80\%$ is a reasonable criterion, based on common power considerations in designing experiments. In the following section, we present a case study providing details on the practical utility of the proposed equivalence criterion.

5 Case Study

5.1 Data Description

The case study consists of a single CM batch sampled at 20 locations chosen equi-spaced throughout the CM run. At each location, 3 tablets were sampled and tested

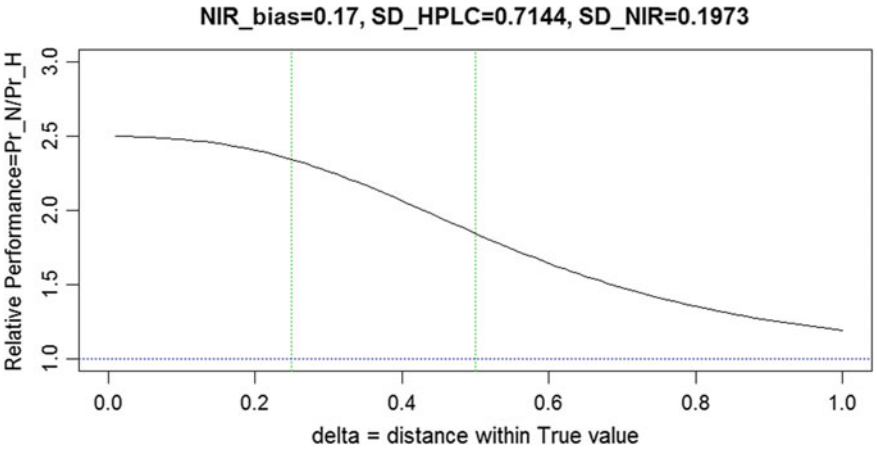


Fig. 2 Relative performance index across increasing values of delta

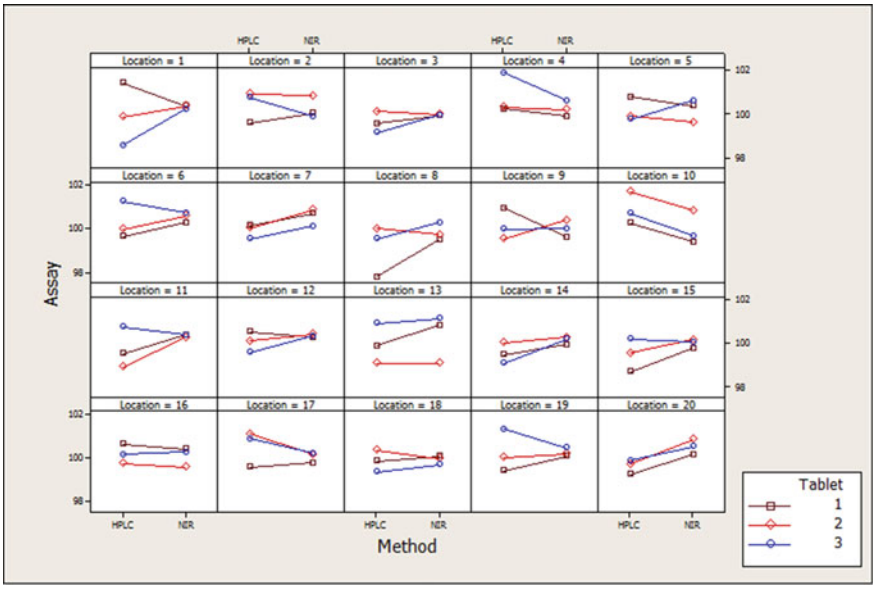


Fig. 3 Case study–NIR and HPLC methods results across 20 locations

by both NIR and HPLC methods. Figure 3 displays the results of the two methods for each tablet, connected by a line.

Table 1 REML parameter estimates and confidence intervals

Effect	Parameter	Estimate (se)	95% confidence interval	
			Lower	Upper
Fixed	HPLC	100.01 (0.10)	99.81	100.32
	NIR	100.18 (0.05)	100.08	100.29
	Bias[a] (NIR-HPLC)	**0.17 (0.09)**	**−0.02**	**0.36**
Random (SD)	Tablet	0.35	0.26	0.53
	Residual (HPLC)	0.70	0.58	0.86
	Residual (NIR)	0.20	0.11	1.04

[a]90% confidence limit for bias = (0.01, 0.33)

5.2 Statistical Model and Results

The following variance components model was used to describe the data:

$$y_{j(i),k} = M_k + L_i + T_{j(i)} + \varepsilon_{j(i),k}$$

where $y_{j(i),k}$ = assay of jth tablet (j = 1, 2, 3) from ith location (i = 1, 2, …, 20) from kth method (k = 1, 2 for HPLC, NIR),

M_k = overall mean from kth method, M_{NIR}, M_{HPLC},

L_i = random effect of ith location: ~ N(0, σ_L^2),

$T_{j(i)}$ = random effect of jth tablet from ith location: ~ N(0, σ_T^2),

$\varepsilon_{j(i),k}$ = residual error from kth method: ~ N(0, $\sigma_{\varepsilon k}^2$).

A preliminary exploratory analysis showed no location effects, therefore the random effect of location was dropped from the final model. The results of the analysis are summarized in Table 1, where the REML parameter estimates and their 95% confidence intervals are given.

A Bayesian simulation of the posterior parameters distribution based on the previous hierarchical model, was carried out, using JAGS [4, 5] with the following vague priors:

M_{HPLC}, M_{NIR} ~ N(μ = 100, σ = 10)

σ_{Tablet} ~ U(0, 5)

σ_{HPLC}, σ_{NIR} ~ U(0, 5).

For the simulation run, 60,000 posterior samples were generated, with 3 chains, following a burn-in of 20,000 simulations and a thinning rate of 25. The results of the Bayesian simulation are given in Table 2. Other prior distributions appropriate for sigma are the half-Cauchy and log normal [3].

The normal Density plots of HPLC and NIR methods centered on their true means are shown in Fig. 4, given the estimated mean bias and median of sigmas from the Bayesian posterior predictive distributions.

OC Curves of probability of falling within Δ units of the true value given the estimated mean bias and median of sigmas for HPLC and NIR methods are shown

Table 2 Bayesian parameter estimates and credible limits

Effect	Parameter	Mean (Median)	95% credible interval	
			Lower	Upper
Fixed	HPLC	100.02 (100.02)	99.81	100.22
	NIR	100.19 (100.19)	100.08	100.29
	Bias[a] (NIR-HPLC)	**0.17 (0.17)**	**−0.02**	**0.36**
Random (SD scale)	Tablet	0.36 (0.36)	0.21	0.47
	Residual (HPLC)	0.72 (0.71)	0.59	0.89
	Residual (NIR)	0.19 (0.20)	0.02	0.37

[a]90% credible interval for bias = (0.011, 0.33)

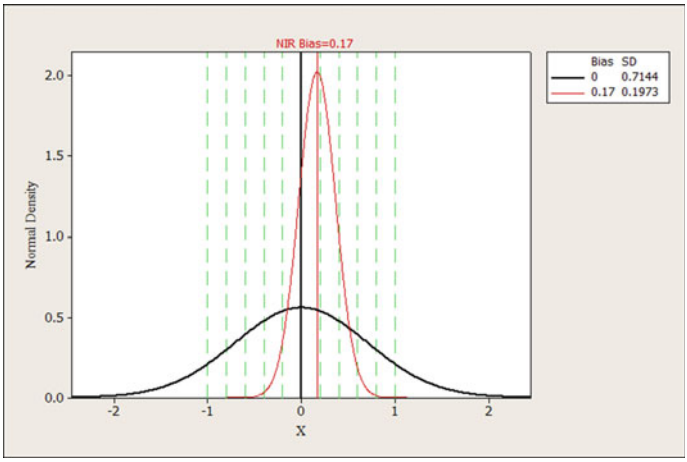


Fig. 4 Normal density plots of HPLC and NIR methods

in Fig. 5. Relative Performance Index across delta values given estimated mean bias and median of sigmas for HPLC and NIR methods are shown in Fig. 6.

Summary Table 3 lists the posterior distribution of *Relative Performance Index* with various values of Δ.

The comparison of the TOST and *Relative Performance Index (Rel_Pfm)* for assessing method equivalence is given in Table 4.

It's clear from the above table that the Rel_Pfm can lead to different conclusions, with respect to the TOST. It's easy to see how this can happen, since the two tests are fundamentally different in their definition of equivalency. The TOST test places a confidence bound on the closeness of M_{HPLC} and M_{NIR}, and the REL_PFM is measuring the closeness of y_{HPLC} to mu and y_{NIR} to the true value. In practical terms, the REL_PFM is a more meaningful test to the practitioner, since it compares directly,

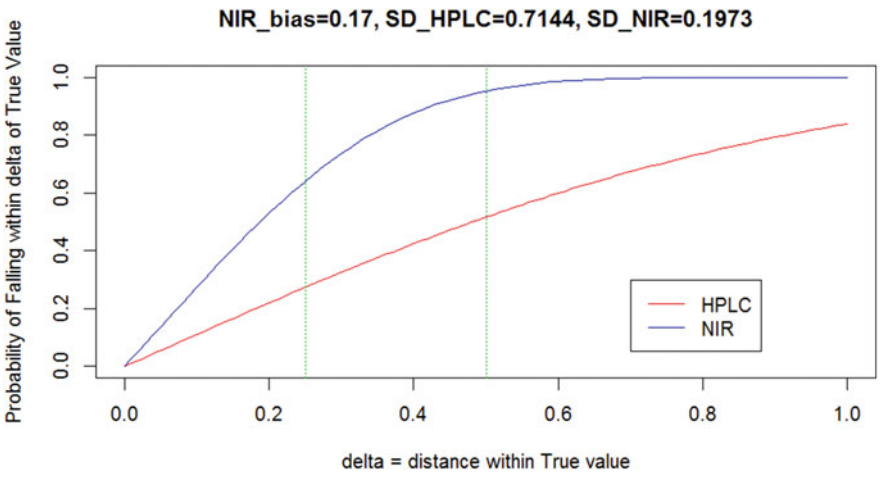


Fig. 5 Probability of falling within Δ units of the true value for HPLC and NIR methods

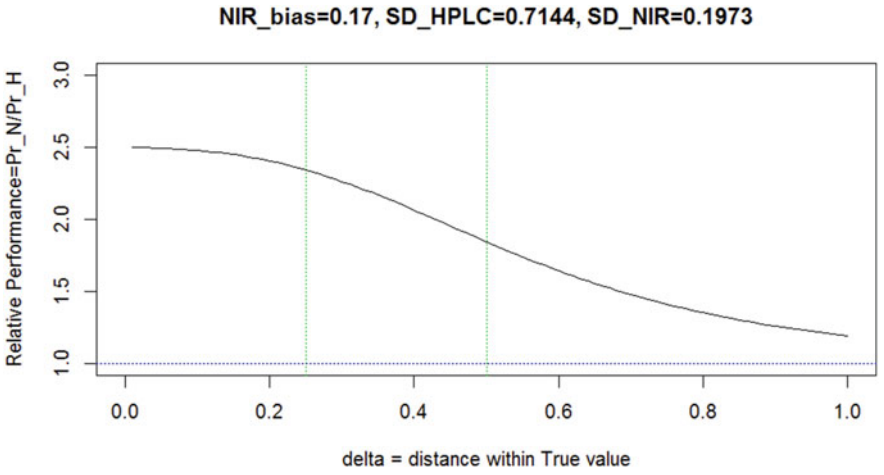


Fig. 6 Relative performance index across delta values for HPLC and NIR methods

the performance of random analytical determinations and their relative outcomes, rather than a mean value, repeated over a long series of trials.

In the clinical literature, the concepts of average bioequivalence (ABE) and individual bioequivalence (IBE) [1] are distinguished in relation to "prescribability" and "switchability", respectively. These concepts describe the physician's ability to interchange patient dosing between two drug products. Corresponding distinctions have not been formulated with respect to analytical method comparisons or bridging studies [8], but we believe that the method and statistic presented in this article can be understood as the analytical method equivalent of individual bioequivalence with

Table 3 Relative performance index versus Δ

Δ	Mean	Median	Maximum	Minimum	Pr(Rel_Pfm ≥ 1)
0.05	2.21	2.01	24.06	0.00	0.819
0.10	2.25	2.05	12.13	0.00	0.849
0.15	2.29	2.11	8.46	0.00	0.890
0.20	2.31	2.15	6.64	0.00	0.929
0.25	2.28	2.17	5.78	0.00	0.959
0.30	2.21	2.15	4.91	0.00	0.980
0.35	2.11	2.10	4.24	0.00	0.991
0.40	2.01	2.02	3.72	0.00	0.996
0.45	1.90	1.91	3.33	0.00	0.999
0.50	1.79	1.80	3.01	0.00	0.999
0.55	1.70	1.70	2.75	0.86	>0.999
0.60	1.61	1.61	2.54	0.89	>0.999
0.65	1.53	1.53	2.36	0.92	>0.999
0.70	1.47	1.46	2.21	0.95	>0.999
0.75	1.41	1.40	2.08	0.97	>0.999
0.80	1.35	1.34	1.97	0.99	>0.999
0.85	1.31	1.30	1.87	1.01	1.000
0.90	1.26	1.26	1.79	1.02	1.000
0.95	1.23	1.22	1.71	1.02	1.000
1.00	1.20	1.19	1.64	1.02	1.000

Table 4 Comparison of TOST with the relative performance index

Test	Criterion	$\Delta = 0.25$	$\Delta = 0.50$
TOST	90% credible interval of bias	90%CI = (0.01, 0.33)	90%CI = (0.01, 0.33)
Decision		Fail	Pass
Relative performance index	$\Pr(\text{Rel\_Pfm} \geq 1) \geq 0.80$	$\Pr(\text{Rel\_Pfm} \geq 1) = 0.96$	$\Pr(\text{Rel\_Pfm} \geq 1) = 1.0$
Decision		Pass	Pass

respect to a random sample preparation. This would be particularly important in the context of stability studies when a method change is being contemplated in the middle of the study. This is a topic that should attract greater discussion by nonclinical statisticians engaged in bridging studies and analytical method comparisons.

6 Summary

Continuous Manufacture is being actively encouraged by the FDA. Pharmaceutical companies are now engaged in weighing its costs/benefits, and several products have received FDA approval to market a product with an associated continuous manufacturing process. Continuous Manufacture offers many scientific and business advantages, but requires PAT analytical methodologies that have been properly validated. We have shown one example of how to establish equivalence of an NIR method to the gold standard HPLC method, through a *Relative Performance Index,* evaluated through Bayesian calculations. The approach is made possible because tablet dispersion can be removed orthogonally, given the paired comparison design. The proposed relative performance index provides a natural interpretation of method performance. We would expect that the REL_PFM would perform well over a wide range of concentrations, as indicated by linearity studies of both analytical methods.

References

1. Chow, S.C.: Bioavailability and bioequivalence in drug development. Wiley Interdiscip. Rev. Comput. Stat. **6**(4), 304–312 (2014). https://doi.org/10.1002/wics.1310
2. Collins, J., et al.: A continuous improvement metric for pharmaceutical manufacturing. Pharm. Technol. **41**(8), 40–42. http://www.pharmtech.com/continuous-improvement-metric-pharmaceutical-manufacturing-0 (2017). Accessed 31 Jan 2018
3. Gelman, A., et al.: Bayesian Data Analysis. CRC Press Bova Raton, FL (2013)
4. JAGS, GNU General Public License version 2
5. Plummer, M.: JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria (2003)
6. SAS 9.4 by SAS Institute Inc., Cary, NC, USA
7. Schuirmann, D.J.: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J. Pharmacokinet. Biopharm. **15**(6), 657–680 (1987)
8. Yang, H., Schofield, T.: statistical considerations for design and analysis of bridging studies. J. Valid. Technol. **20** (2017)

A Further Look at the Current Equivalence Test for Analytical Similarity Assessment

Neal Thomas and Aili Cheng

Abstract Establishing analytical similarity is the foundation of biosimilar product development. Although there is no guidance on how to evaluate analytical data for similarity, the US Food and Drug Administration (FDA) recently suggested an equivalence test on the mean difference between innovator and the biosimilar product as the statistical similarity assessment for Tier 1 quality attributes (QAs), defined as the QAs that are directly related to the mechanism of action. However, the mathematical derivation and simulation work presented in this paper shows that the type I error is typically increased in most realistic settings when an estimate of sigma is used for the equivalence margin. This error cannot be improved by increasing sample size. The impacts of the constant c on type I error and sample size adjustment in the imbalanced situation are discussed, as well.

Keywords Equivalence testing · Type I error rate · Sample size adjustment

1 Introduction

Biosimilar development has recently become a fast growing area. As of June 2017, Europe has 31 approved biosimilars [1], and US has 5 approved [2]. The FDA defines biosimilar as the biological product that 'is highly similar to the reference product notwithstanding minor differences in clinically inactive components,' and that 'there are no clinically meaningful differences between the biological product and the reference product in terms of the safety, purity, and potency of the product.'" [3]. The

N. Thomas
Pfizer, Statistical Research and Consulting Center, 445 Eastern Point Road MS 8260-2270, Groton, CT 06340, USA
e-mail: neal.thomas@pfizer.com

A. Cheng (✉)
Pfizer, Pharmaceutical Sciences and Manufacturing Statistics, 1 Burtt Road, Andover, MA 01810, USA
e-mail: Aili.Cheng@pfizer.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_5

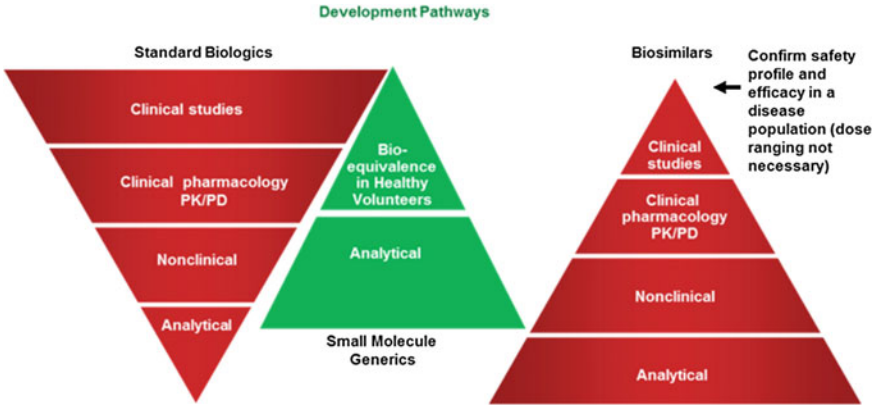


Fig. 1 The comparison of the development pathways among standard biologics, small molecule generics and biosimilars

European Medicine Agency (EMA) defines biosimilars as "*a biopharmaceutical that contains a version of the active substance of an already authorized biopharmaceutical*" [4]. In spite of slightly different definitions, both require similarity to be established based on quality, biological activity, safety, and efficacy. The key is to prove the biosimilar product has no meaningful differences from the already authorized reference product in safety, purity, and efficacy.

It is important to know that for biosimilars the goal is not to reestablish the safety and efficacy. Instead, it is to demonstrate similarity. Therefore, the development pathway is different from the standard biological product and the small molecule generic drug in two aspects [5–8]: (1) the clinical portion of the development is relatively small; (2) the analytical testing becomes the foundation of the similarity development. The FDA not only requires extensive analytical testing, but also recommends formal statistical assessment (Fig. 1).

In the following two sections, the statistical approach recommended by the FDA will be described followed by a detailed assessment on the equivalence test that was recommended by the FDA for the Tier 1 Quality Attributes (QAs).

2 FDA's Tiered Approach for Analytical Similarity Assessment

Biosimilarity development requires extensive analytical testing which often yields 20 to 30 QAs per batch. However, these QAs are not equally related to safety and efficacy. The FDA recommended tiered approach [9–12] applies different levels of statistical rigors to these QAs depending upon their criticality risk ranking, which is determined by their potential impact on activity, PK/PD, safety, and immunogenicity. Tier 1 is for

QAs with the highest risk ranking that usually reflect clinically relevant mechanism(s) of actions. Statistical equivalence test for the mean difference is recommended for Tier 1 QAs. Tier 2 is for QAs with relatively lower risk ranking. A lesser stringent approach, the quality range, is recommended for this tier. Quality range takes the form of $\overline{x} \pm ks$, where $\overline{x}$ and s are the sample mean and sample standard deviation of n_R reference product lots. Biosimilarity for a Tier 2 QA is claimed if at least 90% of the biosimilar product lots fall within the quality range. QAs with lowest risk ranking or important QAs that are not amenable to formal statistical tests or quantitative evaluation are categorized into Tier 3 [13], where only raw data and graphical comparisons are recommended.

3 Assessment of the FDA Recommended Equivalence Test for Tier 1 QAs

Compared to Tier 2, the equivalence test for Tier 1 is not only more rigorous, but also harder to implement. The first and most challenging part is the equivalence acceptance criterion (EAC). Due to the practical limitations, it is very challenging to determine an EAC that is universal for all the Tier 1 QAs across different therapeutic areas. Tsong et al. [10] proposed using $c\sigma_R$ as the universal criterion for all the Tier 1 QAs, where σ_R is the standard deviation of the reference product and c is a constant and is recommended to take the value of 1.5. Since this EAC is a function of σ_R, the performance of the equivalence test is less affected by the assay variability, which is often different for different biosimilar sponsors even targeting for the same reference product. However, the true value of σ_R is usually unknown and has to be estimated from the sample standard deviation, S_R. In other words, cS_R becomes the EAC in practice. However, this apparent minor change in EAC leads to significant changes in the test properties of the equivalence test as cS_R is a variable. The impact of estimating S_R is illustrated in detail below.

3.1 The Impact of S_R Estimate on the Type I Error and the Power

The equivalence test for the mean difference for Tier 1 QAs can be expressed as:

$$\begin{aligned} H_0 &: |\mu_B - \mu_R| \geq \delta \\ H_1 &: |\mu_B - \mu_R| < \delta \end{aligned} \tag{1}$$

where μ_B and μ_R are the true mean responses of the proposed biosimilar and reference product lots, respectively. The δ is the preset equivalence acceptance criterion. If the significance level of the above hypothesis test is set at $\alpha = 5\%$ (i.e. 90% con-

fidence interval of the mean difference is compared against the equivalence margin), the expected type I error rate is 5% if the equivalence margin, δ, is set independently from the data used for equivalence test. Based on the recent recommendation by the FDA, $\delta = 1.5\sigma_R$, where σ_R is the variability of the tested reference product lots which includes both assay and the process variability [9]. The decision rule is to reject the null hypothesis of non-equivalence and claim analytical similarity for the QA if the $(1-2\alpha)$ 100% two-sided confidence interval of the mean difference between the biosimilar and reference product is within $(-1.5\sigma_R, 1.5\sigma_R)$. Unless otherwise specified, α is set at 0.05.

For the rest of this article, the data are denoted by X_{iR}, $i = 1, \ldots, n_R$ and $X_{iB}, i = 1, \ldots n_B$. X_{iR} is the reportable value of the ith reference lot. X_{iB} is the reportable value of the ith biosimilar lot. It is assumed that $X_{iB} \sim N(\mu_B, \sigma_B)$ and $X_{iR} \sim N(\mu_R, \sigma_R)$. Each reportable value X_{iB} or X_{iR} is generally an average of multiple individual test values. The n_R and n_B are the total number of reference and biosimilar lots. The sample means of the reportable values are denoted by $\bar{X}_R$ and $\bar{X}_B$.

In the common variance case, i.e. $\sigma_B^2 = \sigma_R^2 = \sigma^2$, the common variance σ^2, (which includes both within-lot variability and between-lot variability) is estimated by the pooled sample variance, S_p^2. Then, the $(1-2\alpha)$ 100% two-sided confidence interval of the mean difference between the biosimilar and reference product is

$$\left(\bar{X}_B - \bar{X}_R\right) \pm t_{\gamma,(1-\alpha)} S_p \sqrt{\left(\frac{1}{n_R} + \frac{1}{n_B}\right)}, \tag{2}$$

where $S_p = \sqrt{(n_R + n_B - 2)^{-1}\left[\sum_{i=1}^{n_R}\left(X_{iR} - \bar{X}_R\right)^2 + \sum_{i=1}^{n_B}\left(X_{iB} - \bar{X}_B\right)^2\right]}$

$t_{\gamma,(1-\alpha)}$ is the $(1-\alpha)\times 100$ percentile of the t-distribution with γ degrees of freedom, where $\gamma = n_R + n_B - 2$.

If the variances for the biosimilar lots and reference product lots are not equal, the confidence interval of the mean difference is

$$\left(\bar{X}_B - \bar{X}_R\right) \pm t_{\gamma,(1-\alpha)} \sqrt{\left(\frac{S_R^2}{n_R} + \frac{S_B^2}{n_B}\right)} \tag{3}$$

where S_B^2 *and* S_R^2 represent the sample variances for the biosimilar and reference product, respectively. γ is estimated by the Satterthwaite approximation method.

Consider the equal variance case first. Under the null hypothesis: $\mu_B - \mu_R = c\sigma$ and for the alternative hypothesis $\mu_B - \mu_R < c\sigma$, the type I error can be expressed as:

$$P\left[\left(\bar{X}_R - \bar{X}_B\right) + t_{\gamma,(1-\alpha)} S_p \sqrt{\left(\frac{1}{n_R} + \frac{1}{n_B}\right)} < c\sigma\right] \tag{4}$$

Table 1 Comparison of asymptotic and simulated type I error (%) (Nominal level is 5%)

	$\rho = 1$		$\rho = 2$	
n_B	Margin = 1.5 S_R (%)	Margin = 2.5 S_R (%)	Margin = 1.5 S_R (%)	Margin = 2.5 S_R (%)
Asymptotic	9.4	15.2	8.0	12.5
1000	9.1	14.5	7.8	12.3
100	8.3	14.0	7.2	11.5
10	6.2	10.8	5.7	9.2

Equation (3) is equal to the preset significance level, α, as intended. However, when $c\sigma$ is replaced by cS_p, the actual type I error becomes

$$P\left[\left(\bar{X}_R - \bar{X}_B\right) + t_{\gamma,(1-\alpha)} S_p \sqrt{\left(\frac{1}{n_R} + \frac{1}{n_B}\right)} < cS_p\right] \tag{5}$$

Note that the right side of the inequality in (5) is a variable. The distribution of the quality inside the $P(\cdot)$ is not a simple central t distribution anymore. With a relatively small sample size (e.g. <20), the estimation errors in S_p on both sides of the inequality above approximately cancel, and the impact due to the estimation uncertainty of S_p is less. With a large sample size, the estimation error in S_p on the left side of the inequality contributes less. The estimation error in S_p on the right side of the inequality remains, which leads to a relatively larger inflation of the type I error. Furthermore, it can be shown that the asymptotic distribution of the quality inside the $P(\cdot)$ in (5) is normal. Equation (5) can be approximated by $\Phi\left(\frac{z_\alpha}{\sqrt{\lambda}}\right)$, where $\lambda = 1 + \frac{Rc^2}{2(1+R)^2}$, $R = \frac{n_R}{n_B}$ and $\Phi(\cdot)$ is the normal distribution probability function (See Appendix 1). Similarly, when 1.5 S_R is used as the EAC, the asymptotic estimate of the actual type I error can be shown to be approximated by $\Phi\left(\frac{z_\alpha}{\sqrt{\lambda^*}}\right)$, where $\lambda^* = 1 + \frac{c^2}{2(1+R)}$. The asymptotic type I error can be as high as 9.4% when R = 1 (i.e. $n_R = n_B$) (See Appendix 1 Table 1).

To further understand the impact of the estimation uncertainty of EAC, the type I error and the power were calculated via simulation from realistically small to extremely large sample sizes for both equal and unequal variance cases with different EACs (i.e c ranges from 1.5 to 2.5). Assuming normal distribution for both biosimilar and reference products, n_B biosimilar lot values and n_R reference lot values are simulated in each simulation cycle with true mean difference of either $\sigma_R/8$ or 1.5 σ_R. The simulated sample statistics were used to compute 90% confidence intervals on the difference of means, which is then used to compare with different EACs. The number of simulation cycles is 100,000. The type I error rate is estimated as the proportion passing the acceptance criteria when the true mean difference is 1.5 σ_R. The power is estimated as the proportion passing the acceptance criteria when

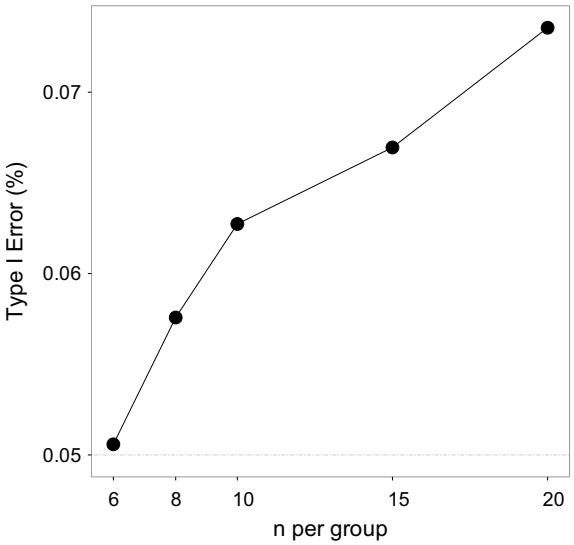


Fig. 2 The type I error for sample size 6 to 20 assuming equal variance case and true mean difference is $1.5*\sigma_R$ (Desired value is 0.05)

the true mean difference is $\sigma_R / 8$. The simulation shows the following when $c\sigma$ is replaced by cS_p or cS_R:

(1) Type I error (i.e. the probability of falsely claiming equivalence, or the patient's risk) is greater than the theoretical value for most of the realistic sample sizes (Fig. 2).
(2) The inflation of Type I error problem cannot be corrected by increasing sample size. On the contrary, the Type I error further increases with an increasing sample size in the balanced case, which is consistent with the derived asymptotic type I error rate (Fig. 3).
(3) The power of the test is decreased compared to the case with known σ, but it still increases with increasing sample sizes as expected. When the sample size increases to 20 per group, the power difference is almost ignorable with a small true mean difference of $\sigma/8$ assuming equal variances (Figs. 4).
(4) The bigger margin (i.e. bigger multiplier c) inflates the impact of the SD estimation errors, which leads to more inflated type I error(s) (Fig. 3).
(5) Similarly, the asymptotic type I error can be derived as well (See Appendix 1) and similar results have been observed for unequal variance case(s) (Fig. 6).

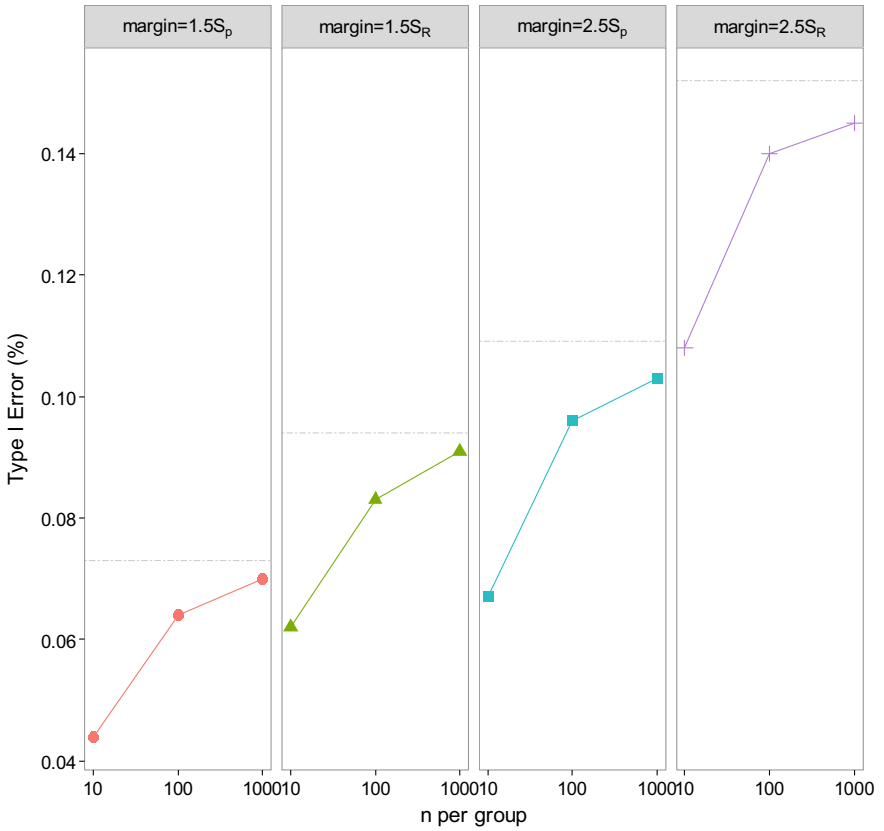


Fig. 3 Type I error assuming equal variance at mean difference of 1.5 $\sigma_R = 1.5\sigma$ (Desired value is 0.05. The grey dash line represents the derived asymptotic type I error rate. Based on the results in Table 1)

3.2 The Unbalanced Sample Size

An unbalanced sample size often occurs and is often due to cost as the expense of purchasing reference product lots is often less than that of manufacturing biosimilar product lots. Theoretically speaking, statistical power is gained by making $n_R / n_B > 1$ without invalidating the testing procedure, which is also shown in the simulation (Figs. 5 and 7). Therefore, the equivalence test result could be dominated by the group with bigger sample size. One way to limit the power increase in this situation is to split the data so that equal sample size can be achieved. However, random data splitting leads to potentially inconsistent results, further inflated Type I error, and decreased power. Another approach to limiting the increase of power due to unbalanced sample size is to establish an upper bound on the bigger sample size. In other words, define the adjusted sample sizes n_R^* and n_B^* such that

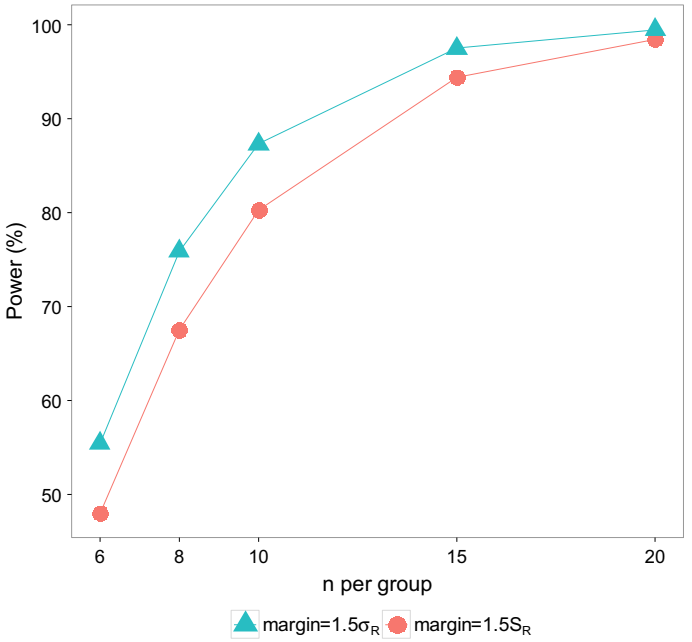


Fig. 4 The power of the equivalence test at true mean difference of $\sigma_R/8$ assuming equal variance and preset type I error rate at 0.05

$$n_R^* = \begin{cases} n_R, & if\ n_R \le cn_B \\ cn_B, & o.w. \end{cases}$$

$$n_B^* = \begin{cases} n_B, & if\ n_B \le cn_R \\ cn_R, & o.w. \end{cases}$$

where c is a contant. It can take a bigger value like 3 or 5 if less adjustment is preferred. However, the FDA statisticians proposed 1.5 for the adjustment in the Tier 1 equivalence test [14]. Therefore, 1.5 is used in this paper.

Then n_R^* and n_B^* can be used in the confidence interval calculation in the place of n_R and n_B as shown in Table 2. Two versions of adjustments are shown in Table 2. In version 1, the degrees of freedom (df) for S_R^2 and S_B^2 are chosen not to be adjusted. So $(n_R - 1)$ and $(n_B - 1)$ are used for the df of the two sample variances S_R^2 and S_B^2. In version 2, adjusted sample sizes, n_R^* and n_B^*, are used to completely replace n_R and n_B. The actual difference between version 1 and 2 is minimal. But for the unequal variance case, version 2 is a slightly more stringent version, leading to a slightly wider confidence interval than version 1. The results shown in Figs. 6 and 7 are based on Version 1, which is also the version used by Dong et al. [14].

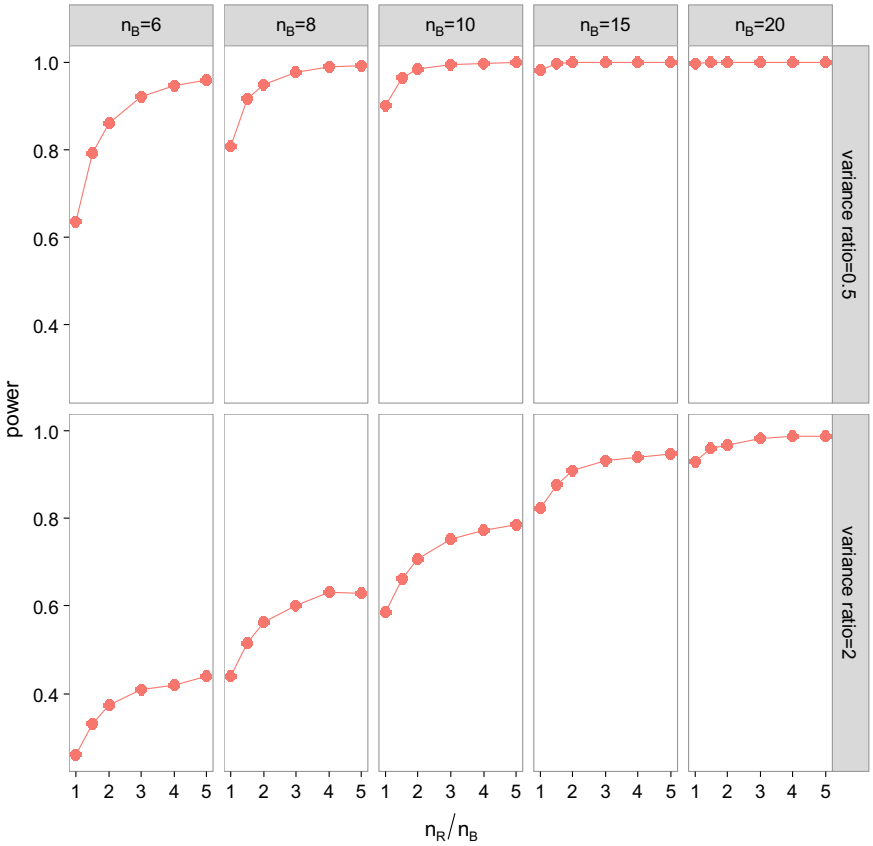


Fig. 5 Power assuming unequal variances at mean difference of $\sigma_R/8$

Table 2 The confidence interval after sample size adjustment

	Equal variance case	Unequal variance case
Version 1	$(\bar{X}_B - \bar{X}_R) \pm t_{\gamma,(1-\alpha)}\sqrt{\left(\frac{S_p^2}{n_R^*} + \frac{S_p^2}{n_B^*}\right)}$ $S_p^2 = \frac{(n_R-1)S_R^2+(n_B-1)S_B^2}{(n_R+n_B-2)}$	$(\bar{X}_B - \bar{X}_R) \pm t_{\gamma,(1-\alpha)}\sqrt{\left(\frac{S_R^2}{n_R^*} + \frac{S_B^2}{n_B^*}\right)}$
Version 2	$(\bar{X}_B - \bar{X}_R) \pm t_{\gamma,(1-\alpha)}\sqrt{\left(\frac{S_p^2}{n_R^*} + \frac{S_p^2}{n_B^*}\right)}$ $S_p^2 = \frac{(n_R^*-1)S_R^2+(n_B^*-1)S_B^2}{(n_R^*+n_B^*-2)}$	$(\bar{X}_B - \bar{X}_R) \pm t_{\gamma,(1-\alpha)}\sqrt{\left(\frac{S_R^2}{n_R^*} + \frac{S_B^2}{n_B^*}\right)}$

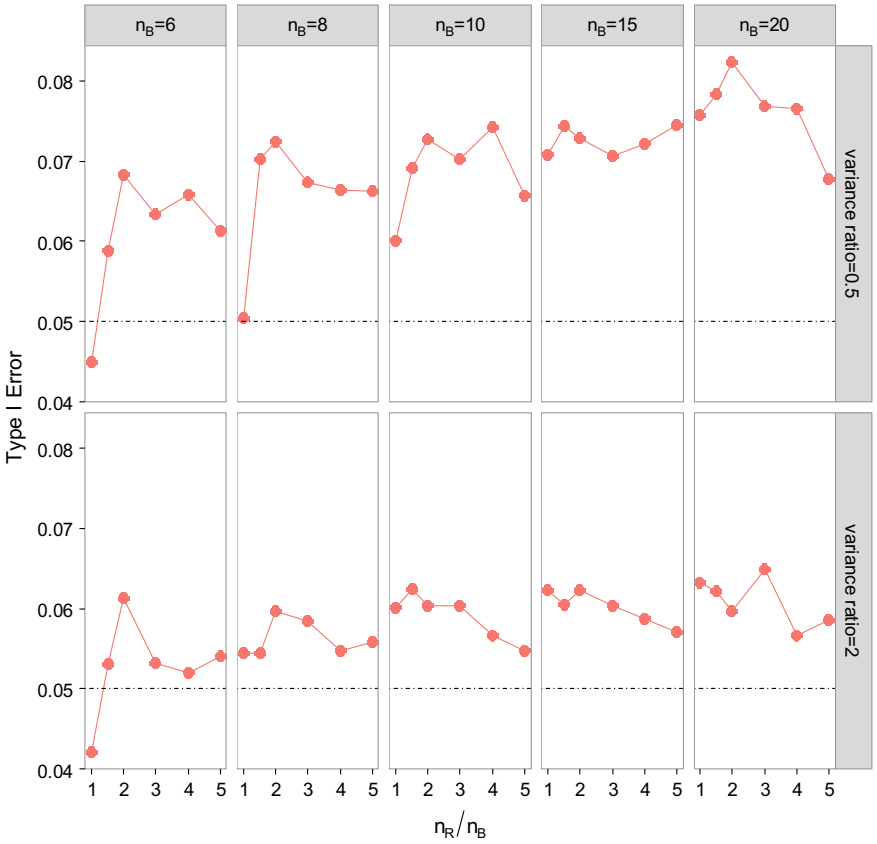


Fig. 6 Type I error assuming unequal variance at mean difference of $1.5\sigma_R$ (Desired value is 0.05)

After sample size adjustment, the power is reduced, especially when the smaller group has sample size of ≤ 7 (Fig. 7). Although it is not the intention of the sample size adjustment approach, the type I error happens to be reduced, compared to the unadjusted version (Fig. 8). In cases where the $n_R / n_B \geq 3$, the type I error drops even below 0.05, the nominal type I error level.

4 Summary

The tiered approach as proposed by the FDA recognizes the different levels of relevance of QAs in the analytical similarity. This approach allows different levels of statistical rigors be applied to different QAs. However, the current equivalence test as

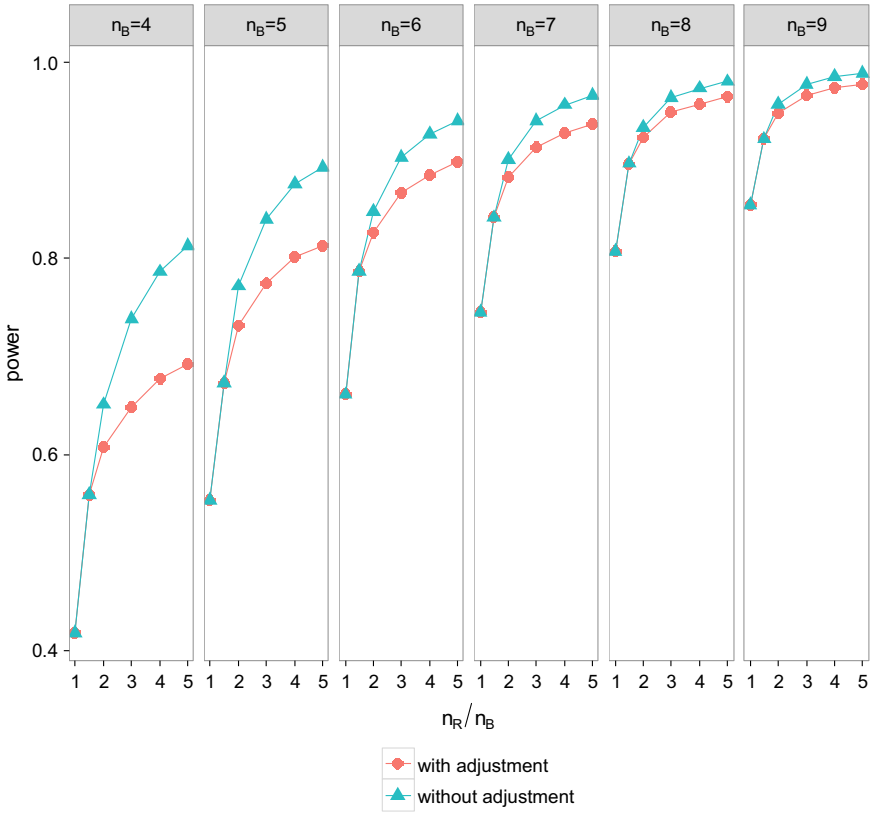


Fig. 7 The power of the equivalence test at true mean difference of $\sigma_R/8$ with and without sample size adjustment assuming equal variance and preset significance level of 0.05

proposed for Tier 1 QAs is not ideal in that the type I error rate cannot be controlled to the nominal level due to the formulation of the equivalence acceptance criterion. Unlike power, which can be improved with increasing sample sizes, the type I error, not only is inflated with realistic sample sizes, but also increases more with increasing sample sizes in the balanced situation. In addition, the type I error rate increases with wider equivalence margin. In the unbalanced case, the introduction of the sample size adjustment approach does help to control the unwanted power increase due to unbalanced sample size. However, the sample size adjustment tends to over-correct the Type I error issue, which could lead to an actual type I error, much lower than the nominal level, especially in cases where the sample size ratio $\geq$3.

Given the above issues, a new exact test would be needed to ensure proper control of the power and the type I error. The equivalence test based on effect size (i.e ratio of the mean difference to the reference product standard deviation) [15] is one such test that could be used for similarity assessment.

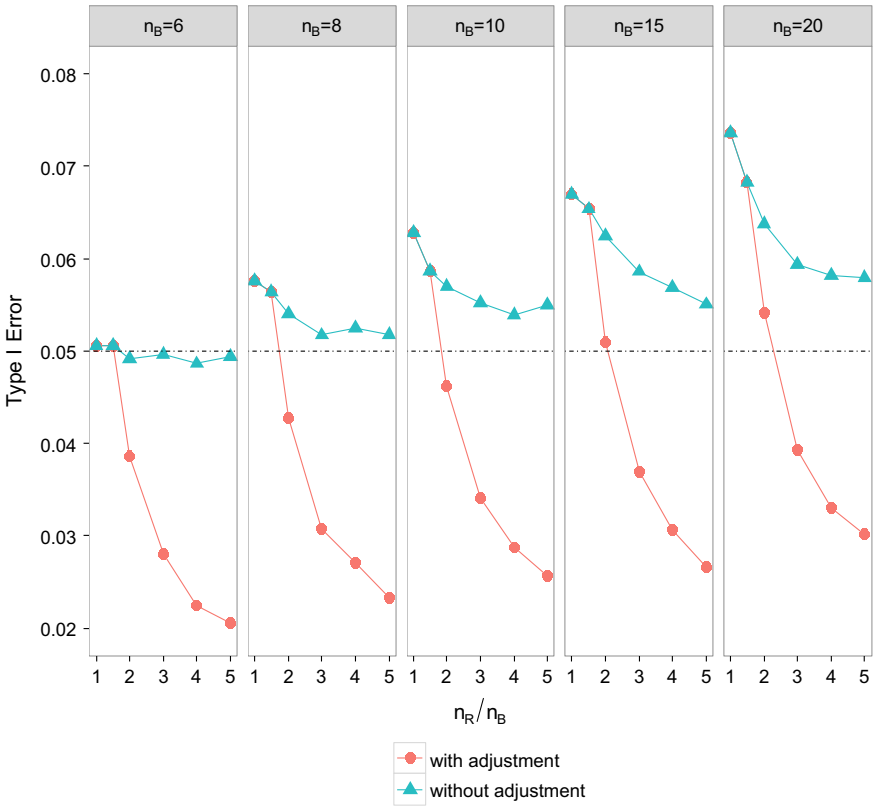


Fig. 8 The simulated type I error of the equivalence test at true mean difference of $1.5\sigma_R$, with and without sample size adjustment, assuming equal variance and preset significance level of 0.05

Acknowledgements The authors thank Ira Jacobs for providing the background information for Fig. 1.

Appendix 1 The Asymptotic Approximation to the Type I Error Rate for the Statistical Equivalence Test

A1.1 Equal Variance Case: $\sigma_R = \sigma_B = \sigma$

Let $\theta = (\mu_R, \mu_B, \sigma^2)$ and $\rho = n_R / n_B$, where $n_B \to \infty$. Let α be the significance level of the equivalence test (type I error rate), and t be the $(1-\alpha) \times 100$th percentile of the t-distribution with $n_R + n_B - 2$ degrees of freedom. All probability calculations that follow are conditional on the parameters in θ.

The probability in (5) can be expressed as

$$\Pr\left[\bar{d} + tS_p\sqrt{\left(\frac{1}{n_R}+\frac{1}{n_B}\right)} < cS_R\right] - \Pr\left[\bar{d} - tS_p\sqrt{\left(\frac{1}{n_R}+\frac{1}{n_B}\right)} < -cS_R \quad \text{and} \quad \bar{d} + tS_p\sqrt{\left(\frac{1}{n_R}+\frac{1}{n_B}\right)} < cS_R\right] \tag{A.6}$$

With small biosimilar sample sizes (e.g., n_B = 5-20), the estimation error in S_R in (A.6) is correlated with the estimation error in S_p, because the same reference-product lots are included in both estimators. This results in substantial cancellation of estimation errors that improves both type I error and power.

As the biosimilar sample sizes become large, i.e, $n_B \to \infty$, the estimation error from S_p is divided by a large sample size, and thus the contribution of the estimation error in S_p rapidly decreases with sample size. In contrast, the estimation error in S_R is not divided by the sample size, and so it decreases at the same rate as the estimation error in $\bar{d}$, and thus it is no longer approximately cancelled in settings with very large sample sizes. This results in an unanticipated increase in type I error as sample size increases. Also, larger values of the multiplier c inflate the impact of the estimation error on the test boundary that results in less cancellation and larger type I error. A formal derivation of the asymptotic results that quantifies this heuristic discussion follows.

With $(\mu_R - \mu_B) = c\sigma$, the first probability in (A.6) can be re-expressed as

$$\Pr\left[\frac{\bar{d} - (\mu_R - \mu_B)}{S_p\sqrt{\frac{1}{n_R}+\frac{1}{n_B}}} + t - c\left(\frac{S_R - \sigma}{S_p\sqrt{\frac{1}{n_R}+\frac{1}{n_B}}}\right) < 0\right] \tag{A.7}$$

which equals

$$P_\theta\left[\frac{\bar{d} - (\mu_R - \mu_B)}{S_p\sqrt{\frac{1}{n_R}+\frac{1}{n_B}}} - c\sqrt{\frac{\rho}{1+\rho}}\left(\frac{\sqrt{n_B}(S_R - \sigma)}{S_p}\right) < -t\right] \tag{A.8}$$

As $n_B \to \infty$, $-t \to z_\alpha$, which is the lower $\alpha\times$ 100th percentile of the normal distribution.

The S_p in the denominators of (A.8) converge in probability to σ, so by Slutksy's Theorem (Billingsley, p. 294, exercise 25.7), the probability in (A.8) converges to the same limit as

$$P_\theta\left[\frac{\bar{d} - (\mu_R - \mu_B)}{\sigma\sqrt{\frac{1}{n_R}+\frac{1}{n_B}}} - c\sqrt{\frac{\rho}{1+\rho}}\left(\frac{\sqrt{n_B}(S_R - \sigma)}{\sigma}\right) < z_\alpha\right]. \tag{A.9}$$

The first term in (A.9) has a standard normal distribution. From Kendall and Stuart, Volume 1, p. 245, Eq. (10.9), $\sqrt{\rho n_B}\left(S_R^2 - \sigma^2\right) \Rightarrow N\left(0,\ 2\sigma^4\right)$, so by the delta method,

$$\sqrt{\rho n_B}(S_R - \sigma) \Rightarrow N\left(0,\ \frac{\sigma^2}{2}\right) \tag{A.10}$$

It follows that the asymptotic variance of the right summand on the left side of the inequality in (A.9) is $\frac{c^2}{2(1+\rho)}$. For normal data, the sample means and variances are independent, so the difference in the two terms in (A.9) is asymptotically normal with mean 0, and variance $V_A = 1 + \frac{c^2}{2(1+\rho)}$, which provides an analytic asymptotic approximation to the type I error of the test:

$$\Pr\left[\bar{d} + tS_p\sqrt{\left(\frac{1}{n_R} + \frac{1}{n_B}\right)} < cS_R\right] \approx \Phi\left(\frac{z_\alpha}{\sqrt{V_A}}\right) \tag{A.11}$$

Note that the asymptotic level of the test does not depend on σ. As $n_B \to \infty$, the second probability in Eq. (A.6) is less than or equal to

$$\Pr\left[\bar{d} - tS_p\sqrt{\left(\frac{1}{n_R} + \frac{1}{n_B}\right)} < -cS_R\right] \tag{A.12}$$

The probability in (A.12) approaches 0 as n_B and n_R increase. Thus, there is no reduction in the asymptotic type I error of the test. When $(\mu_R - \mu_B) = -c\sigma$, the first probability in (A.6) becomes

$$\Pr\left[\bar{d} - tS_p\sqrt{\left(\frac{1}{n_R} + \frac{1}{n_B}\right)} > -cS_R\right], \tag{A.13}$$

and the second probability is changed accordingly. The derivation is unchanged except for some sign changes that cancel to produce the same approximation for the lower bound in (A.13).

Other estimators for σ_R on the right side of the inequality in (A.6) have been proposed, including S_p and the sample variance from the reference product obtained by randomly splitting the reference data. The analytic methods used here can be easily modified to obtain asymptotic approximations for the resulting type 1 errors using these estimators, also yielding similar formulas.

A1.2 Unequal Variance Case: $\sigma_R \neq \sigma_B$

Let $q = \frac{\sigma_B^2}{\sigma_R^2}$. The type I error of the test formed using the Satterthwaite approximation when S_R is used in the test boundary is

$$\Pr\left[\bar{d} + t\sqrt{\left(\frac{S_R^2}{n_R} + \frac{S_B^2}{n_B}\right)} < cS_R\right] \tag{A.14}$$

where the degrees of freedom for t are now obtained by Satterthwaite's approximation (Satterthwaite, 1946). As $n_B \to \infty$, $t \to z_\alpha$, and (A.14) becomes

$$\Pr\left[\frac{\bar{d} - (\mu_R - \mu_B)}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{q\sigma_R^2}{n_B}}} - c\left(\frac{S_R - \sigma_R}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{q\sigma_R^2}{n_B}}}\right) < z_\alpha\right] \tag{A.15}$$

The first term in inequality in (A.15) follows a standard normal, and the second term follows a normal distribution asymptotically as well with mean zero and variance $\frac{c^2}{2(1+q\rho)}$, so

$$\Pr\left[\bar{d} + t\sqrt{\left(\frac{S_R^2}{n_R} + \frac{S_B^2}{n_B}\right)} < cS_R\right] \approx \Phi\left(\frac{z_\alpha}{\sqrt{V_A}}\right) \tag{A.16}$$

where $V_A = 1 + \frac{c^2}{2(1+q\rho)}$.

A1.3 Comparing Asymptotic Approximations to Simulation Results

The asymptotic approximations to the type I error were compared to simulation results with different sample sizes ($n_B = 10,\ 100,\ 1000$; $\rho = 1,\ 2$), and different test boundaries (margin $= 1.5\ S_R,\ 2.5\ S_R$). Table 1 presents a comparison of the simulated and asymptotic type I error rates. As expected, the simulated type I errors approach the asymptotic predictions, but the convergence is slow. The simulation results confirm that the inflation of type I error caused by the estimation error in the test boundary cannot be fixed by increasing n_B. Increasing the ratio ρ; however, decreases the inflation in the type 1 error because this reduces the estimation error on the right side of the inequality more rapidly than the error on the left side where estimation in the mean difference depends on n_B. As predicted, larger values of c inflate the impact of the estimation errors in the test boundary, leading to a higher type I error. Results with unequal variances were qualitatively similar.

References

1. European Medicines Agency: European public assessment reports (2017)
2. Wikipedia (2017) https://en.wikipedia.org/wiki/Biosimilar
3. US Food and Drug Administration: Guidance for Industry: Scientific Considerations in Demonstrating Biosimilarity to a Reference Product. Rockville, MD (2015)
4. European Medicines Agency: Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: Non-clinical and clinical issues (2014). Accessed August 2015
5. Berghout, A.: Clinical programs in the development of similar biotherapeutic products: Rationale and general principles. Biologicals **39**, 293–296 (2011). https://doi.org/10.1016/j.biologicals.2011.06.024
6. McCamish, M.: EBG's perspective on the draft guideline on the non-clinical/clinical issues. EMA Workshop on Biosimilars, London (2013)
7. Schneider, C.K., et al.: Setting the stage for biosimilar monoclonal antibodies. Nat. Biotechnol. **30**, 1179–1185 (2012)
8. US Food and Drug Adminstration: Abbreviated New Drug Applications (ANDA): Generics. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/AbbreviatedNewDrugApplicationANDAGeneri Accessed January 3, 2016
9. Chow, S.C.: On assessment of analytical similarity in biosimilar studies. Drug Des. **3**, e124 (2014). https://doi.org/10.4172/2169-0138.1000e124
10. Tsong, Y., Shen, M., Dong, X.: Equivalence margin determination for analytical biosimilarity assessment. In: IABS Workshop at USP Headquarters, Rockville, MD (2015)
11. Tsong, Y., Shen, M., Dong, X.: Development of statistical approaches for analytical biosimilarity evaluation. In: 2015 DIA/FDA Statistical Forum, Rockville, MD (2015)
12. Tsong, Y., Shen, M., Dong, X.: Development of statistical approaches for analytical biosimilarity evaluation. In: 2015 ISBS-DIA Joint Symposium on Biopharmaceutical Statistics, Beijing, China (2015)
13. U.S. Food and Drug Administration: Statistical approaches to evaluate analytical similarity guidance for industry. Draft guidance (2017)
14. Dong, X (Cassie)., Weng, Y.T., Tsong, Y: Adjustment for unbalanced sample size for analytical biosimilar equivalence assessment. J. Biopharm. Stat. (2017). https://doi.org/10.1080/10543406.2016.1265544
15. Burdick, R.K., Thomas, N., Cheng, A.: Statistical considerations in demonstrating CMC analytical similarity for a biosimilar product. Stat. Biopharm. Res. (2017). https://doi.org/10.1080/19466315.2017.1280412

Shiny Tools for Sample Size Calculation in Process Performance Qualification of Large Molecules

Qianqiu Li and Bill Pikounis

Abstract The regulatory guidance documents on process validation have been recently revised to emphasize the three-stage lifecycle approach throughout validation. As an important milestone within Stage 2: process qualification, the process performance qualification (PPQ) requires taking adequate samples to provide sufficient statistical confidence of quality both within a batch and between batches. To help meet the PPQ requirements and to further support continued process verification for large molecules, for continuous critical quality attributes, Shiny tools have been developed to calculate the minimum numbers of samples within batches to control the batch-specific beta-content tolerance intervals within prespecified acceptance ranges. The tolerance intervals at attribute level are also displayed to assure the suitability of the predefined number of PPQ batches. In addition, another Shiny application for creation and evaluation of the sampling plans for binary attributes will be illustrated in terms of failure rates of future batches and consumer's and producer's risk probabilities. The tools for both continuous and binary attributes allow to adjust the sampling plans based on historical data, and are designed with interactive features including dynamic inputs, outputs and visualization.

Keywords Process performance qualification · R Shiny · Sample size calculation · Tolerance intervals · Sampling plans · Variance component analysis · Normal and Binary attributes

1 Introduction

In pharmaceutical regulation, process validation is a mandatory task consisting of the collection and evaluation of data from process design phase through commercial

Q. Li (✉) · B. Pikounis
Janssen Research & Development LLC, Spring House, PA, USA
e-mail: qli16@its.jnj.com

B. Pikounis
e-mail: bpikouni@its.jnj.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_6

production. The ultimate goal of process validation is to establish scientific evidence that a process is capable of consistently delivering quality products. To achieve this goal, besides proper process design and risk-based control strategy at stage 1 of process validation, preparation for process performance qualification (PPQ) at stage 2 after equipment/utility/facility qualification is also vital. PPQ must be performed under GMP guidance and according to the established protocol and an enhanced sampling plan. The PPQ sampling plan defines the number of PPQ batches and the number of samples within batches for each evaluated critical quality attribute (CQA). The regulatory guidance documents on process validation do not explicitly specify the number of batches and the numbers of samples within batches. However, the 2011 FDA process validation guidance recommends sampling plans with adequate assurance or statistical confidence. The European Medicines Agency (EMA) PV guidance also states that the number of batches should be based on the variability of the process, the complexity of the process and product and the amount of historical data and available process knowledge. In the paper, we describe three R Shiny applications to facilitate the task of developing PPQ sampling plans. The Shiny applications are the web-based interfaces that provides users with customized inputs and outputs using R [13]. Specifically, for continuous CQAs with acceptance limits, the Shiny application SSNormTI calculates tolerance intervals across prespecified numbers of samples within batches under one-way random effects models. These tolerance intervals are then compared with the acceptance limits to help choose the number of samples within batches and to evaluate the proposed number of batches. The Shiny application RiskBinom implements AQL and/or RQL sampling plans for binary CQAs. VarCompLM, as a supplemental tool, can be used to produce estimates and confidence limits for the overall and/or group means, and the between-batch and within-batch variances. The estimates or the confidence limits can then be used as the inputs of SSNormTI for sample size calculation via normal tolerance intervals. Statistical methods are detailed in next section.

2 Sample Size Calculation Methods

2.1 For Continuous CQAs with Acceptance Limits

For continuous CQAs with acceptance limit(s), the Shiny application SSNORMTI determines the number of samples within batches using batch-specific tolerance intervals for a single population that is normally distributed. The batch-specific tolerance intervals are beta-content intervals that with a prespecified confidence level (e.g., 95%), will contain at least a specified proportion (e.g., 99%) of the data population. The confidence level and the coverage probability of the tolerance intervals are defined by prior knowledge for the CQA, for example, using criticality levels of the CQAs evaluated prior to PPQ. The two-sided batch-specific tolerance intervals are calculated using the R function K.factor() in the "tolerance" package via either

close-to-exact or approximation approaches. Specifically, the following approaches were implemented for both two-sided batch-specific and process-specific tolerance intervals, respectively:

- Howe [6] (with method = "H" in K.factor) and Hoffman and Kringle [5];
- Owen (1964) (with method = "OCT" in K.factor) and Mee [10];
- Krishnamoorth and Mathew [9] (with method = "KM" in K.factor) and Krishnamoorthy and Lian ([7]; modified large sample approximation and generalized pivotal quantity approaches)

The degrees of freedom required by K.factor are either not specified or are equal to the degrees of freedom associated with the residual variance for a one-way random effects model, given the minimum number of batches specifically required for each approach to control the process-specific tolerance interval within the acceptance limits.

Besides the above two-sided batch-specific and process-specific tolerance intervals, for CQAs with only a lower or an upper acceptance limit, SSNormTI also calculates the exact one-sided batch-specific tolerance interval via K.factor (i.e., method = "EXACT"; with the degrees of freedom defined as above), and the one-sided process-specific tolerance intervals of Mee and Owen [11], Hoffman [4], and Krishnamoorthy and Mathew [8] via modified large-sample approximation and a generalized pivotal quantity approach.

The rationale of constructing the process-specific tolerance intervals resides in the requirement of specifying the number of PPQ batches in a PPQ sampling plan. Though the number of PPQ batches is not stated in any regulatory process validation guidance document and it is generally based on historical information about the process risk and feasible considerations such as limitation of available resources, or the timing of the PPQ with respect to a regulatory submission, it is still a regulatory suggestion for a successful PPQ to demonstrate an appropriate level of between-batch variability. Therefore, we use the process-specific tolerance intervals to examine the proposed number of PPQ batches. The process-specific tolerance intervals are constructed under the below one-way random effects model with batch as a random effect and thus reflect the level of between-batch variability.

Specifically, for a CQA, let $X_{i1}, X_{i2}, \ldots, X_{iJ}$ be the (original or transformed) measurements randomly sampled from batch $i (i = 1, \ldots, I)$. It is assumed that the randomly sampled PPQ data satisfy a random effect model,

$$Y_{ij} = \mu + B_i + E_{ij} \; with \; j = 1, \ldots, J$$

where Y_{ij} denotes measurement j from batch i., μ is the overall mean, and $\mu + B_i$ is the mean for batch i. B_i for all $i = 1, \ldots, I$ are independent and follow the same normal distribution $N\left(0, \sigma_B^2\right)$ with zero mean and variance σ_B^2. Residuals E_{ij} for all i and j are independently and identically distributed as normal variables with mean 0 and variance σ^2 for all batches, that is, $E_{ij} \sim N\left(0, \sigma_E^2\right)$ $for \; any \; i$ and j. $B_i \; and \; E_{ij}$ are assumed to be mutually independent.

In addition, without between-batch variance and overall mean, SSNormTI calculates the number of samples by controlling batch-specific tolerance intervals within a prespecified acceptance range via the R function norm.ss() in the package "tolerance". The R function adopts three methods (named as precision-based hereafter): Owen (1964), Faulkenberry and Weeks (1968), and Young et al. (2016).

SSNormTI requires the batch mean and within-batch variance for the batch-specific tolerance intervals, and the overall mean and between-batch variance for the process-specific tolerance intervals. The methods for those inputs will be provided using examples in the section "Illustration and Examples".

2.2 For Binary CQAs

The Shiny application RiskBinom is designed to create and evaluate the PPQ sampling plans for binary CQAs. Two levels of quality, or acceptable quality level (AQL) and rejectable quality level (RQL), are considered in the sampling plans. Thus, the sampling plans are attribute (acceptance) sampling plans. Acceptable quality level (AQL) is the largest proportion of defectives, or the largest value that is considered acceptable and desired by the producer. Rejectable quality level (RQL) is the smallest value for which the lot must be rejected. The RQL is also called the Lot Tolerance Percent Defective (LTPD). In a manner analogous to specification of the confidence levels and coverage probabilities of the tolerance intervals for continuous CQAs, the AQL and RQL are chosen with regard to prior knowledge of the associated risks for the evaluated binary CQA. Given the quality levels, the number of samples within batches is determined by controlling the producer's risk rate and/or consumer's risk rate to be lower than prespecified acceptance limit(s). The producer's risk rate is the probability of rejecting a lot having AQL quality. The consumer's risk rate is the probability of accepting a lot having RQL quality. Given prespecified AQL, RQL and acceptance and rejection limits, the number of samples within batches are determined by controlling the producer's risk rate, the consumer's risk rate, or both not exceeding prespecified acceptance limit(s).

Specifically, the Shiny application RiskBinom calculates the two risk rates according to their conventional definitions under binomial distributions or adjusts them based on historical data in both frequentist and Bayesian frameworks, as shown by Table 1 for the single sampling plans. A is defined as the acceptance number. Based on the historical data, L_p and U_p are the lower limit and upper limit of the 95% one-sided confidence interval for the lot failure rate. The 95% one-sided confidence limits are calculated via the R function binconf() using three methods: Exact, Wilson and Asymptotic. Then we set U_p as the smallest one among the three upper limits and set L_p as the largest one among the three lower limits.

The Bayesian approach assesses the two risk rates using the posterior means. The posterior means are obtained under an assumption that the number of failed samples in a historical batch follows a binomial distribution with a common failure rate across all batches. Namely, $\mathrm{X_i} \sim \mathrm{Binomial}(\mathrm{n_i}, \mathrm{p_F})$ with $\mathrm{i} = 1, \ldots, \mathrm{B}$; X_i denotes

Table 1 Risk probabilities by Shiny application RiskBinom

Name	Type	Definition
Producer's risk probability	Without historical data	$Pr(X > A \mid p_F \leq AQL)$
	With historical data via frequentist approach	$Pr\big(X > A \mid p_F \leq min\big(U_p, AQL\big)\big)$
	With historical data via Bayesian approach	Posterior mean of $Pr(X > A \mid p_F \leq AQL)$
Consumer's risk probability	Without historical data	$Pr(X \leq A \mid p_F \geq RQL)$
	With historical data via frequentist approach	$Pr\big(X \leq A \mid p \geq max\big(L_p, RQL\big)\big)$
	With historical data via Bayesian approach	Posterior mean of $Pr(X \leq A \mid p_F \geq RQL)$

the number of failed samples in the i-th batch with n_i binary measurements. p_F is the failure rate, and B refers to the number of the historical batches. With a conjugate Beta prior distribution Beta(a, b) [1, 12] for the failure rate p_F, conditional on the binomially distributed historical data, the posterior distribution of the failure rate also follows a beta distribution $\text{Beta}\left(a + \sum_{i=1}^{B} x_i, b + \sum_{i=1}^{B}(n_i - x_i)\right)$ characterized by the probability density function:

$$\frac{1}{\int_0^1 u^{a+\sum_{i=1}^{B} x_i - 1}(1-u)^{b+\sum_{i=1}^{B}(n_i - x_i)-1} du} p^{a+\sum_{i=1}^{B} x_i - 1}(1-p)^{b+\sum_{i=1}^{B}(n_i - x_i)-1}$$

In the Beta prior distribution, the two hyper-parameters a and b are set in three ways:

- to be non-informative: $a = b = 0.5$;
- under assumption of 20% failure probability: $a = 0.2 \, and \, b = 0.8$;
- empirically based on the historical data.

Besides the above capability of creation of single sampling plans using both frequentist and Bayesian approaches, the application is also designed for evaluation of acceptance sampling plans under the binomial models. Given the specified AQL, RQL or both, and batch-specific inputs including the number of measurements, and acceptance and rejection numbers for each batch, the producer's and/or consumer's risk rates are quantified separately by batch and across batches. With $X_i \sim Binomial(n_i, p_F)$, with $X_i \, and \, n_i$ as the number of failed samples and the number of samples in the i-th batch:

- For the i-th batch, the producer's risk rate, denoted as $P(X_i \geq R_i \mid p_F = AQL)$, is the probability that the number of failed samples X_i is equal to or greater than the rejection number R_i with respect to the i-th batch given $p_F = AQL$. The consumer's risk rate, denoted as $P(X_i \leq A_i \mid p_F = RQL)$, is the probability that

the number of failed samples X_i is equal to or smaller than the acceptance number A_i with respect to the i-th batch given $p_F = RQL$;

- Across all batches, the producer's and consumer's risk rates are obtained under the assumption that one batch was sampled at each sampling stage and the batches are ordered as given by the user's inputs or the imported file. Specifically, the producer's risk occurs when one or more batches of AQL quality are rejected. and consumer's risk occurs when one or more batches of RQL are accepted. Under the assumptions of batch independence and all batch failure rates equal to p_F, the two risk rates can be expressed as follows:

$$\begin{aligned}\text{Producer's Risk Rate} =& P(X_1 \geq R_1 | p_F = AQL) \\ &+ \sum_{i=2}^{I} P(X_i \geq R_i | p_F = AQL) \prod_{u=1}^{i-1} P(A_u < X_u < R_u | p_F = AQL)\end{aligned}$$

$$\begin{aligned}\text{Consumer's Risk Rate} =& P(X_1 \leq A_1 | p_F = RQL) \\ &+ \sum_{i=2}^{I} P(X_i \leq A_i | p_F = RQL) \prod_{u=1}^{i-1} P(A_u < X_u < R_u | p_F = RQL)\end{aligned}$$

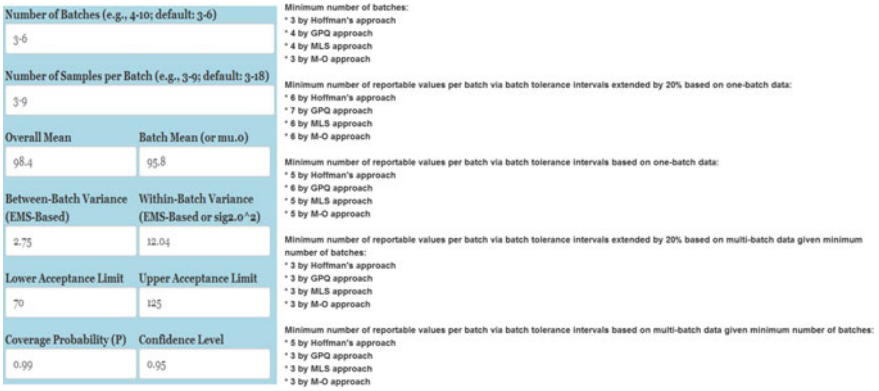
2.3 Illustration and Examples

The three Shiny applications offer the ability to assist the PPQ sampling plans as web-based tools. Such interfaces are particularly beneficial to users who are not familiar with R. Key examples are presented below to illustrate the functional and statistical ability separately for each application.

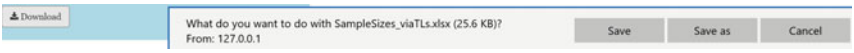
2.4 SSNormTI

SSNormTI implements the sample size calculation via tolerance intervals at three scenarios: (1) the between-batch variance and the overall mean across batches are specified; (2) the between-batch variance and the overall mean across batches are not specified without historic data imported; and (3) the between-batch variance and the overall mean across batches are not specified with historical data imported.

In the first scenario, given specified between- and within- batch variances and an overall mean, the four types of the process-specific tolerance intervals, as described before, can be calculated to examine the proposed number of batches. For example, with the user's inputs in the left window below, the minimum numbers of batches to control the process-specific tolerance intervals between 70 and 125 are summarized in the right window, with the minimum numbers of samples per batch required to control the calculated or extended-by-20% batch-specific tolerance intervals between 70 and 125.



When clicked, the Download button at the lower-left corner of the left panel allows all of the calculated batch-specific and process-specific tolerance intervals to be saved in a default file SampleSizes_viaTL.xlsx or in another file with user defined format. The file has four spreadsheets, each of which contains the results by one of the four approaches.

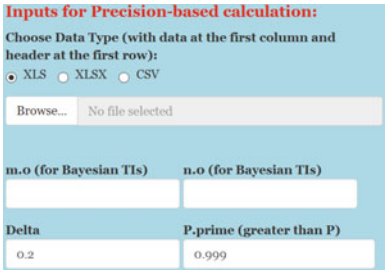


In the second scenario, with no imported historical data, only Owen's approach (1964) is applied to obtain the minimum number of samples required to control the batch-specific tolerance interval within prespecified acceptance range. For example, with batch mean of 74.6 and within-batch variance of 1.2, at least six samples per batch (as shown in the last cell of the table below) will be required to control the batch-specific tolerance interval with 95% coverage and 95% confidence by Owen (1964) above 70:

Table: Results by direct approach (Owen (1964))

Lower Spec Limit	Upper Spec Limit	Confidence	Coverage Probability	Minimum Sample Size
70.00	NA	0.9500	0.9500	6

If Delta and P.prime are provided, for example, equal to 0.2 and 0.999 as follows,

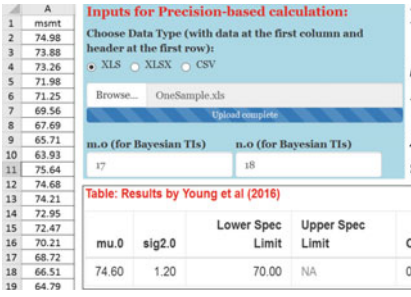


then the Faulkenberry and Weeks approach (1968) can be applied to obtain the minimum number of samples per batch, for example, as shown by the table below. With batch mean of 74.6 and within-batch variance of 1.2, at least 13 samples per batch are required to maintain the below-20% probability that the coverage probability of the 95%-confidence tolerance interval exceeds 99.9%.

Table: Results by Faulkenberry & Weeks (1968)

Confidence	Coverage Probability	Delta	P.prime	Minimum Sample Size
0.9500	0.9500	0.200	0.9990	13

The third scenario requires upload of historical data at the first column of a CSV, XLS or XLSX file, with the first row as the column or data label (as shown in the table below on the left). Then using Young's approach (2016) without specification of Delta and P.prime, the minimum number of samples per batch, together with Delta and P.prime will be determined by controlling either the exact tolerance interval without inputs of m.0 and n.0 (prior hyperparameters m_0 *and* n_0 for the R function bayesnormtol.int()), or the Bayesian tolerance interval incorporating inputs of m.0 and n.0. With batch mean of 74.6 and within-batch variance of 1.2, the table on the bottom shows 9 as the minimum number of samples per batch obtained by the Bayesian tolerance interval using the normal and scaled inverse chi-square as the priors for mean and variance, respectively with $m_0 = 17$ & $n_0 = 18$. The calculated Delta and P.prime are 0.05263158 and 0.99998661.



Priors of mean and variance:

$$\mu|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{n_0}\right) \quad \sigma^2 \sim Scaled - inv - \chi^2(m_0, \sigma_0^2)$$

When a scaled inverse chi-squared random variable $\chi^2(\upsilon, \tau^2)$ is divided by $\upsilon\tau^2$, it results in an inverse chi-squared random variable with υ degrees of freedom.

Table: Results by Young et al (2016)

mu.0	sig2.0	Lower Spec Limit	Upper Spec Limit	Confidence	Coverage Probability	Delta	P.prime	Minimum Sample Size
74.60	1.20	70.00	NA	0.9500	0.9500	0.05263158	0.99998661	9

2.5 *VarCompLM*

VarCompLM is an under-development application for variance component analysis under linear mixed models with at least one random effect. Given the analysis model, based on the uploaded data, VarCompLM can produce the estimates and the confidence limits for the model parameters. One example follows to show how the estimates and the confidence limits are derived and used for the SSNormTI inputs corresponding to the means (overall and batch-specific) and variances (between-batch and within-batches).

In the example, a stability data set was simulated. The data set (with the first 25 rows shown below on the left) includes 90 data points (in Column E) from different times (up to 36 months in Column B), batches (A1–A6 & B1–B6 in Column A), groups (A & B in Column D), and labs (L1–L3 in Column C).

A suitable analysis model contains time as a covariate and lab as a fixed class variable, and batch as a random effect. As specified below in the right panel, the model fits to the data separately by group using the R function lme() to get the estimates and confidence limits (at bottom) for all unique combinations of fixed effects (time and lab) [16]. Those estimates or confidence limits can be chosen as the overall mean and batch mean representing the worst-case scenario (e.g., with an estimate or confidence limit closest to a predefined acceptance limit).

	A	B	C	D	E
1	batch	time	lab	group	msmt
2	A1	0	L1	A	74.98
3	A1	3	L1	A	73.88
4	A1	6	L1	A	73.26
5	A1	9	L1	A	71.98
6	A1	12	L1	A	71.25
7	A1	18	L1	A	69.56
8	A1	24	L1	A	67.69
9	A1	30	L1	A	65.71
10	A1	36	L1	A	63.93
11	A2	0	L1	A	75.64
12	A2	3	L1	A	74.68
13	A2	6	L1	A	74.21
14	A2	9	L1	A	72.95
15	A2	12	L1	A	72.47
16	A2	18	L1	A	70.21
17	A2	24	L1	A	68.72
18	A2	30	L1	A	66.51
19	A2	36	L1	A	64.79
20	A3	0	L2	A	73.26
21	A3	3	L2	A	72.29
22	A3	6	L2	A	71.34
23	A3	9	L2	A	70.75
24	A3	12	L2	A	69.82
25	A3	18	L2	A	68.07

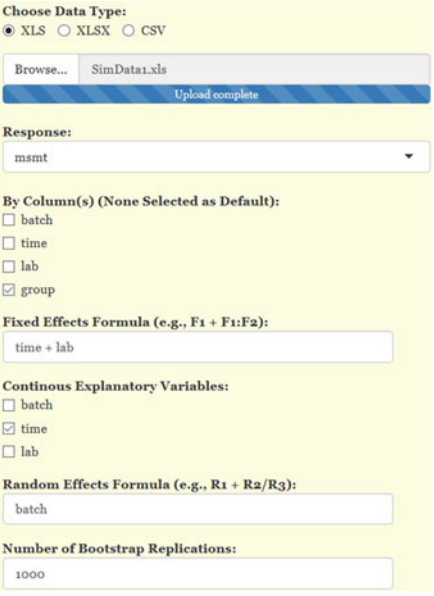


Table 2 Specification of Beta Prior Hyper-parameters a and b

Scenario	Equations for a and b
With one historical batch	$\frac{x}{n} = \hat{E}(p_F) \approx \frac{a}{a+b} \frac{\hat{E}(p_F)\left(1-\hat{E}(p_F)\right)}{(a+b+1)} = \widehat{Var}(p_F) \approx \frac{ab}{(a+b)^2(a+b+1)}$
With two or more historical batches	$\frac{1}{B}\sum_{i=1}^{B} \frac{x_i}{n_i} = \hat{E}(p) \approx \frac{a}{a+b} s^2 \approx \frac{1}{B}\sum_{i=1}^{B} \widehat{Var}\left(\frac{X_i}{n_i}\right)$ Where $s^2 = \mathrm{Var}\left(\frac{x_1}{n_1}, \cdots, \frac{x_B}{n_B}\right)$ & $\widehat{Var}\left(\frac{X_i}{n_i}\right)$ is a function of $\hat{E}(p)$ & n_i If the above equations result in negative a or b, then use $\widehat{Var}\left(\frac{X_i}{\sqrt{n_i}}\right)$

Group A

time	lab	Estimate(SE)	95% CI	90% CI	80% CI
0	L1	75.337(0.3154)	(74.699,75.975)	(74.806,75.869)	(74.926,75.749)
0	L2	73.320(0.3147)	(72.684,73.957)	(72.790,73.850)	(72.910,73.730)
0	L3	72.635(0.3147)	(71.999,73.272)	(72.105,73.165)	(72.225,73.045)
3	L1	74.428(0.3143)	(73.792,75.064)	(73.899,74.958)	(74.018,74.838)
3	L2	72.411(0.3139)	(71.776,73.046)	(71.882,72.940)	(72.002,72.820)
3	L3	71.726(0.3141)	(71.091,72.361)	(71.197,72.255)	(71.317,72.135)
6	L1	73.519(0.3134)	(72.885,74.153)	(72.991,74.047)	(73.110,73.927)
6	L2	71.502(0.3134)	(70.868,72.135)	(70.974,72.030)	(71.093,71.910)
6	L3	70.817(0.3137)	(70.182,71.451)	(70.288,71.345)	(70.408,71.226)
9	L1	72.610(0.3128)	(71.977,73.242)	(72.083,73.137)	(72.202,73.017)
9	L2	70.592(0.3130)	(69.959,71.226)	(70.065,71.120)	(70.184,71.000)
9	L3	69.907(0.3136)	(69.273,70.542)	(69.379,70.436)	(69.499,70.316)
12	L1	71.700(0.3124)	(71.068,72.332)	(71.174,72.227)	(71.293,72.108)
12	L2	69.683(0.3129)	(69.050,70.316)	(69.156,70.210)	(69.275,70.091)
12	L3	68.998(0.3137)	(68.364,69.633)	(68.470,69.527)	(68.589,69.407)
18	L1	69.882(0.3124)	(69.250,70.514)	(69.355,70.408)	(69.475,70.289)
18	L2	67.865(0.3135)	(67.230,68.499)	(67.336,68.393)	(67.456,68.273)
18	L3	67.180(0.3146)	(66.543,67.816)	(66.650,67.710)	(66.769,67.590)
24	L1	68.063(0.3133)	(67.430,68.697)	(67.535,68.591)	(67.655,68.472)
24	L2	66.046(0.3150)	(65.409,66.683)	(65.515,66.577)	(65.635,66.457)
24	L3	65.361(0.3165)	(64.721,66.001)	(64.828,65.894)	(64.948,65.774)
30	L1	66.245(0.3152)	(65.607,66.882)	(65.714,66.776)	(65.834,66.656)
30	L2	64.227(0.3175)	(63.585,64.870)	(63.692,64.762)	(63.814,64.641)
36	L1	64.426(0.3180)	(63.783,65.069)	(63.890,64.962)	(64.012,64.841)

Group B

time	lab	Estimate(SE)	95% CI	90% CI	80% CI
0	L1	79.456(0.5616)	(78.318,80.594)	(78.508,80.403)	(78.723,80.189)
0	L2	76.990(0.5613)	(75.853,78.127)	(76.043,77.937)	(76.258,77.722)
0	L3	76.635(0.5613)	(75.498,77.772)	(75.688,77.582)	(75.903,77.368)
3	L1	78.543(0.5608)	(77.407,79.680)	(77.597,79.489)	(77.811,79.275)
3	L2	76.077(0.5606)	(74.941,77.213)	(75.131,77.023)	(75.346,76.809)
3	L3	75.723(0.5609)	(74.586,76.859)	(74.776,76.669)	(74.991,76.454)
6	L1	77.631(0.5603)	(76.495,78.766)	(76.685,78.576)	(76.900,78.362)
6	L2	75.165(0.5602)	(74.030,76.300)	(74.220,76.110)	(74.434,75.896)
6	L3	74.810(0.5607)	(73.674,75.946)	(73.864,75.756)	(74.078,75.542)
9	L1	76.718(0.5599)	(75.584,77.852)	(75.773,77.663)	(75.987,77.449)
9	L2	74.252(0.5600)	(73.118,75.387)	(73.307,75.197)	(73.522,74.983)
9	L3	73.897(0.5606)	(72.761,75.033)	(72.952,74.843)	(73.166,74.629)
12	L1	75.805(0.5597)	(74.671,76.939)	(74.861,76.750)	(75.075,76.536)
12	L2	73.340(0.5599)	(72.205,74.474)	(72.395,74.284)	(72.609,74.070)
12	L3	72.985(0.5608)	(71.849,74.121)	(72.039,73.931)	(72.253,73.717)
18	L1	73.980(0.5598)	(72.846,75.114)	(73.036,74.925)	(73.250,74.711)
18	L2	71.514(0.5603)	(70.379,72.650)	(70.569,72.460)	(70.783,72.246)
18	L3	71.160(0.5616)	(70.022,72.298)	(70.212,72.107)	(70.427,71.892)
24	L1	72.155(0.5606)	(71.019,73.291)	(71.209,73.101)	(71.424,72.887)
24	L2	69.689(0.5615)	(68.552,70.827)	(68.742,70.637)	(68.957,70.422)
24	L3	69.334(0.5632)	(68.193,70.476)	(68.384,70.285)	(68.600,70.069)
30	L1	70.330(0.5622)	(69.191,71.469)	(69.381,71.278)	(69.596,71.064)
30	L2	67.864(0.5634)	(66.723,69.006)	(66.914,68.815)	(67.129,68.599)
36	L1	68.505(0.5645)	(67.361,69.649)	(67.552,69.457)	(67.768,69.241)

Also, VarCompLM generates two tables including the estimates and confidence limits of the random components, obtained via the R package VCA [15] based on the residuals from the model including only the fixed effects (time and lab in this example). The first table shows the variance estimates based on expected mean squares (EMS), and two type of confidence intervals for variances: Satterthwaite's CIs [14] under assumption of chi-square distributions for all variance components; and SAS CIs under a chi-square distribution for total and error and normal approximation for the other variance components. The second table is grounded in likelihood and bootstrap methods and implemented using the R functions intervals() and confint(), respectively. As the process-specific tolerance intervals use the EMS-based variances, it is recommended to take the estimates or 95% one-sided upper confidence limits in Table 1, as the inputs for SSNormTI. The results in Table 2 may be chosen if the EMS-based estimates or confidence limits are too large to be used.

Table 1: EMS-Based Estimates and Confidence Intervals of Variance Components

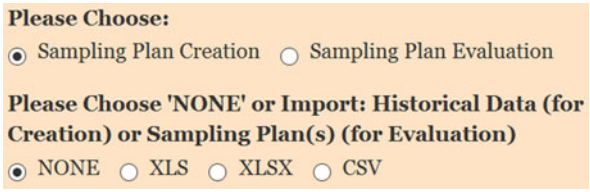
group	Name	DF	Variance	90% Satterthwaite CI	95% Satterthwaite CI	95% 2-Sided Satterthwaite CI	90% SAS CI	95% SAS CI	95% 2-Sided SAS CI
A	batch	2.86	0.204	(NA,1.110)	(NA,1.881)	(0.064,3.130)	(NA,0.422)	(NA,0.484)	(0.000,0.538)
A	Residual	39.00	0.037	(NA,0.051)	(NA,0.055)	(0.025,0.060)	(NA,0.051)	(NA,0.055)	(0.025,0.060)
B	batch	2.95	0.687	(NA,3.600)	(NA,6.024)	(0.219,9.894)	(NA,1.411)	(NA,1.617)	(0.000,1.795)
B	Residual	37.00	0.043	(NA,0.060)	(NA,0.066)	(0.028,0.071)	(NA,0.060)	(NA,0.066)	(0.028,0.071)

Table 2: Likelihood-Based Estimates and Confidence Intervals of Variance Components

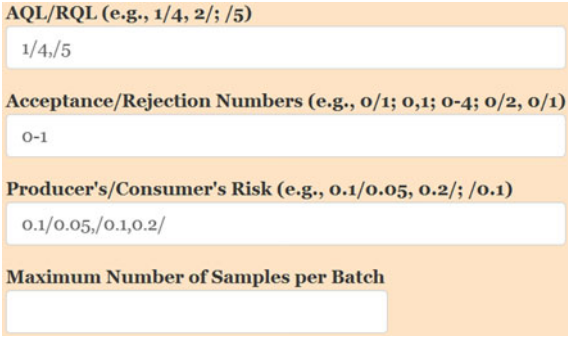
group	Name	Variance	90% Wald CI	95% Wald CI	95% 2-Sided Wald CI	90% Bootstrap CI	95% Bootstrap CI	95% 2-Sided Bootstrap CI
A	batch	0.191	(NA,0.557)	(NA,0.755)	(0.037,0.983)	(NA,0.448)	(NA,0.564)	(0.011,0.665)
A	Residual	0.037	(NA,0.049)	(NA,0.053)	(0.023,0.057)	(NA,0.048)	(NA,0.052)	(0.021,0.054)
B	batch	0.621	(NA,1.788)	(NA,2.412)	(0.123,3.128)	(NA,1.509)	(NA,1.744)	(0.062,2.116)
B	Residual	0.043	(NA,0.057)	(NA,0.062)	(0.027,0.067)	(NA,0.055)	(NA,0.058)	(0.025,0.065)

2.6 *RiskBinom*

The walk-through examples below demonstrate how to use this application for each of the two main tasks in attribute sampling plans: creation and evaluation.



If the objective is to create a sampling plan, then click the first radio-button before "Sampling Plan Creation". Secondly, if without historical data, then a user should choose "NONE" and then enter all inputs in the boxes on the left panel.



For example, as shown above, a user is required to specify:

- AQL/RQL (in the first box): it is accepted to have one or a mixture of the three formats in the parenthesis; AQL is the value before a backslash, RQL is the value after a backslash; and two or more AQL/RQL settings are separate by comma;

- Acceptance/rejection numbers (in the second box): using a backslash to separate the acceptance number and the rejection number, if without any backslash, then all input values are referred to as the acceptance numbers; for example, multiple acceptance numbers are separated using commas; a range of acceptance numbers can be specified by the minimum and the maximum and with a dash in between;
- Producer's and Consumer's risk rates (in the third box): they must be between zero and 1, separately by a backslash if both are provided; a value before a backslash defines a producer's risk rate, while a value after a backlash gives a consumer's risk rate;
- Maximum Number of Samples per Batch (in the fourth box): if provided, then the result table will only show the numbers of samples per batch below this maximum and meeting the criteria with respect to the producer's and/or consumer's risk rates; if not provided, then 2,000 will be used as the maximum.

After providing the above inputs, from the header of the right panel, choose "Sampling Plan Creation" (the middle one below) to get the table including all sampling plans associated with all combinations of the inputs.

Table: Frequentist Attribute Sampling Plans given Prespecified Quality and Risk Levels

AQL	RQL	Acceptance Number	Rejection Number	Required Producer's Risk Probability	Required Consumer's Risk Probability	Number of Samples	Actual Producer's Risk Probability	Actual Consumer's Risk Probability
1.00	4.00	0	1	0.10	0.05	NA	NA	NA
NA	4.00	0	NA	NA	0.10	57	NA	0.0976
1.00	NA	NA	1	0.20	NA	2	0.0199	NA
1.00	4.00	1	2	0.10	0.05	NA	NA	NA
NA	4.00	1	NA	NA	0.10	96	NA	0.0993
1.00	NA	NA	2	0.20	NA	2	0.0001	NA
NA	5.00	0	NA	NA	0.05	59	NA	0.0485
NA	5.00	0	NA	NA	0.10	45	NA	0.0994
NA	5.00	1	NA	NA	0.05	93	NA	0.0500
NA	5.00	1	NA	NA	0.10	77	NA	0.0973

The actual producer's and consumer's risk rates are given in the last two columns, respectively. In the third last column, the numbers of samples within a batch are listed. An "NA" at the third last column suggests non-existence of any sampling plan satisfying the corresponding specifications. For example, given the specifications in the first row of the above table, there doesn't exist any plan with the number of samples within a batch below 2,000.

If a user expects to create a sampling plan on the basis of historical data, then the historical data should be summarized into a data file in.xlsx,.xls, or.csv filetypes, with the format as follows:

Batch	Number of samples	Number of failed samples	Acceptance number
B1	30	1	0
B2	24	1	0
B3	12	0	0

Then based on the uploaded data, the largest lower confidence limit L_p and the smallest upper confidence limit U_p of the 95% one-sided confidence intervals for the failure rate are calculated and used to justify the AQL and RQL using the equations below:

$$AQL = min\left(U_p, \text{Prespecified}AQL\right) \text{ and } RQL = max\left(U_p, \text{Prespecified}RQL\right)$$

For example, given two AQL/RQL settings: AQL = 0.65%, RQL = 1%; or RQL = 5%, acceptance number equal to 1 or 2, and two risk sets: producer's risk rate = 20% & consumer's risk rate = 10%; or consumer's risk rate = 20%, then RiskBinom produces two tables, the first table below shows the frequentist results, including the calculated sample sizes at the third last column and the actual risk rates at the last two columns, with RQL adjusted from 1 to 1.01%.

Table: Frequentist Attribute Sampling Plans given Prespecified Quality and Risk Levels

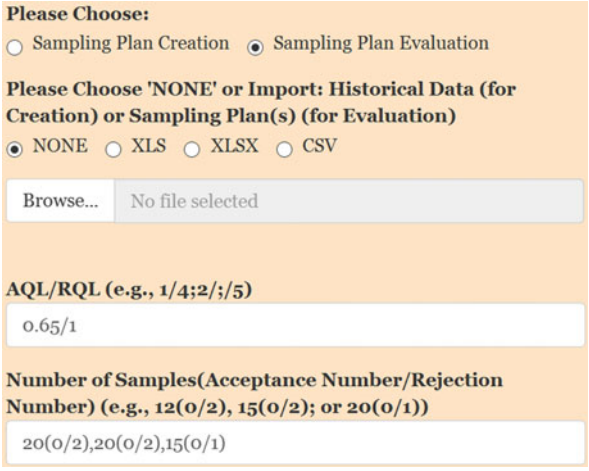
AQL	RQL	Acceptance Number	Rejection Number	Required Producer's Risk Probability	Required Consumer's Risk Probability	Number of Samples	Actual Producer's Risk Probability	Actual Consumer's Risk Probability
0.65	1.01	1	2	0.20	0.10	NA	NA	NA
NA	1.01	1	NA	NA	0.20	297	NA	0.1986
0.65	1.01	2	3	0.20	0.10	NA	NA	NA
NA	1.01	2	NA	NA	0.20	424	NA	0.1992
NA	5.00	1	NA	NA	0.10	77	NA	0.0973
NA	5.00	1	NA	NA	0.20	59	NA	0.1991
NA	5.00	2	NA	NA	0.10	105	NA	0.0992
NA	5.00	2	NA	NA	0.20	85	NA	0.1963

Following the above table, another table contains all the Bayesian results. Below the screenshot only displays the results given ALQ = 0.65% and RQL = 1%. One more column has been added to include the posterior distributions, given the three beta priors: Beta(0.5,0.5), Beta(0.2, 0.5), and Beta(a, b) with a & b empirically determined by the historical data.

Table: Bayesian Attribute Sampling Plans given Prespecified Quality and Risk Levels

AQL	RQL	Acceptance Number	Rejection Number	Required Producer's Risk Probability	Required Consumer's Risk Probability	Posterior/Prior	Number of Samples	Actual Producer's Risk Probability	Actual Consumer's Risk Probability
0.65	1.00	1	2	0.20	0.10	Beta(2.500,64.500)/Jeffrey	149	0.1510	0.0993
0.65	1.00	1	2	0.20	0.10	Beta(2.800,64.200)/80%-Prior-Success-Rate	135	0.1361	0.0990
0.65	1.00	1	2	0.20	0.10	Beta(2.023,64.915)/Empirical	173	0.1712	0.0989
NA	1.00	1	NA	NA	0.20	Beta(2.500,64.500)/Jeffrey	102	NA	0.1990
NA	1.00	1	NA	NA	0.20	Beta(2.800,64.200)/80%-Prior-Success-Rate	92	NA	0.1991
NA	1.00	1	NA	NA	0.20	Beta(2.023,64.915)/Empirical	119	NA	0.1999
0.65	1.00	2	3	0.20	0.10	Beta(2.500,64.500)/Jeffrey	214	0.0830	0.0990
0.65	1.00	2	3	0.20	0.10	Beta(2.800,64.200)/80%-Prior-Success-Rate	193	0.0698	0.0999
0.65	1.00	2	3	0.20	0.10	Beta(2.023,64.915)/Empirical	247	0.1004	0.0992

RiskBinom evaluates an attribute sampling plan by taking inputs from the left panel or an uploaded file. For example, as shown in the proceeding screenshot below, AQL and RQL take values of 0.65% and 1%, respectively. The sampling plan allows one or more batches and assumes that one batch is sampled at one stage and all batches are sampled in the order according to the inputs. For each batch, firstly, the number of samples is specified and then followed by a parenthesis, including the acceptance and rejection numbers.



Then in the “sampling Plan Evaluation” panel, two tables are displayed. The top one shows the risk rates separately calculated by batch, and the bottom one contains the overall risk rates across all batches.

Table: Risk-Based Results at Batch Level for Attribute Sampling Plans

Batch	Number of Samples(Acceptance Number/Rejection Number)	AQL	RQL	Producer's Risk Probability	Consumer's Risk Probability
Batch 1	20(0/2)	0.65	1	0.0074	0.8179
Batch 2	20(0/2)	0.65	1	0.0074	0.8179
Batch 3	15(0/1)	0.65	1	0.0932	0.8601

Table: Risk-Based Results across All Batches for Attribute Sampling Plans

Number of Samples(Acceptance Number/Rejection Number)	AQL	RQL	Overall Producer's Risk Probability	Overall Consumer's Risk Probability
20(0/2),20(0/2),15(0/1)	0.65	1	0.0095	0.9765

If a user uploads a file to evaluate the sampling plan, then two file formats should be adopted. In the sampling plan, if all batches have a common number of samples and common acceptance and rejection numbers across batches, then the template as follows should be used:

AQL	RQL	Number of batches	Number of samples per batch	Acceptance/rejection numbers per batch
0.025	1	3	15	0/2
0.065	1	1	20	0/1
NA	1	4	30	0/2
4	NA	2	18	0/2

In a multiple sampling plan, the acceptance number is generally equal to the rejection number subtracted by 1 for the last sampled batch, and it's not rare that different batches have different numbers of samples. Thus, the template below can flexibly support such a heterogeneity property of the sampling plan.

AQL	RQL	Number of batches	Batch 1	Batch 2	Batch 3
1	5	3	15(0/2)	18(0/2)	20(0/1)
4	10	2	20(0/2)	30(1/2)	

The two multiple sampling plans in the above table are evaluated through RiskBinom, and the results are presented below:

Table: Risk-Based Results at Batch Level for Attribute Sampling Plans

Batch	Number of Samples(Acceptance Number/Rejection Number)	AQL	RQL	Producer's Risk Probability	Consumer's Risk Probability
Batch 1	15(0/2)	1.00	5.00	0.0096	0.4633
Batch 2	18(0/2)	1.00	5.00	0.0138	0.3972
Batch 3	20(0/1)	1.00	5.00	0.1821	0.3585
Batch 1	20(0/2)	4.00	10.00	0.1897	0.1216
Batch 2	30(1/2)	4.00	10.00	0.3388	0.1837

Table: Risk-Based Results across All Batches for Attribute Sampling Plans

Number of Samples(Acceptance Number/Rejection Number)	AQL	RQL	Overall Producer's Risk Probability	Overall Consumer's Risk Probability
15(0/2),18(0/2),20(0/1)	1.00	5.00	0.0150	0.6579
20(0/2),30(1/2)	4.00	10.00	0.3145	0.1712

3 Discussion

Development of these shiny tools was motivated by the goal of assisting users to generate and explore PPQ sampling plans for continuous and binary CQAs. Their flexible inputs and interactive outputs may spare users a considerable amount of time, for example, consumed by searching through different options. The applications have been tested on common web browsers (Google Chrome, IE 10+).

Acknowledgements We thank Paulo Bargo for helping with configuration and installation of the Shiny applications to the Janssen Shiny server. Our gratitude extends to the team led by Kenneth Hinds for developing an internal position paper [2] regarding statistical based sampling plans for PPQ. The authors also appreciate the comments on the statistical methods and the Shiny applications from manufacturing scientists and statistician colleagues.

References

1. Bolstad, W.M., Curran, J.M.: Introduction to Bayesian Statistics. 2nd edn. John Wiley & Sons Inc (2017)
2. DS-VAL-68021: Position Paper to Address Statistically Based Sampling Plan for Process Validation Stage 2 (Process Performance Qualification). Janssen Internal Document (2014)
3. Graybill, F.A., Wang, C.: Confidence Intervals on Non-negative Combinations of Variances. JASA **75**(372), 869–873 (1980)
4. Hoffman, R.D.: One-sided tolerance limits for balanced and unbalanced random effects models. Technometrics **52**(3), 303–312 (2010)
5. Hoffman, D., Kringle, R.: Two-sided tolerance intervals for balanced and unbalanced random effects models. J. Biopharm. Stat. **15**, 283–293 (2005)
6. Howe, W.G.: Two-sided tolerance limits for normal populations, some improvements. J. Am. Stat. Assoc. **64**, 610–620 (1969)
7. Krishnamoorthy, K., Lian, X.: Closed-form approximation tolerance intervals for some general linear models and comparison studies. J. Stat. Comput. Simul. **82**(4), 1–17 (2012)
8. Krishnamoorthy, K., Mathew, T.: One-sided tolerance limits in balanced and unbalanced one-way random models based on generalized confidence intervals. Technometrics **46**(1), 44–52 (2004)
9. Krishnamoorthy, K., Mathew, T.: Statistical Tolerance Regions: Theory, Applications, and Computation. Wiley (2009)
10. Mee, R.W.: Beta-expectation and Beta-content tolerance limits for balanced one-way ANOVA random model. Technometrics **26**, 251–254 (1984)
11. Mee, R.W., Owen, D.B.: improved factors for one-sided tolerance limits for balanced one-way ANOVA random model. J. Am. Stat. Assoc. **78**(384), 901–905 (1983)
12. Puza, B.: Bayesian Methods for Statistical Analysis. ANU eView (2015)
13. Resnizky, H.G.: Learning Shiny. Packt Publishing Ltd (2015)
14. Satterthwaite, F.E.: An approximation distribution of estimates of variance components. Biom. Bull. **2**(6), 110–114 (1946)
15. Schuetzenmeister, A.: VCA: Variance Component Analysis. R package version 1.3.3 (2017)
16. West, B.T., Welch, K.B., Galeckl, A.T.: Linear Mixed Models: A Practical Guide Using Statistical Software. Chapman & Hall/CRC (2007)

Part III
Continuous Process

Risk Evaluation of Registered Specifications and Internal Release Limits Using a Bayesian Approach

Yijie Dong and Tianhua Wang

Abstract This article proposes to pursue advanced statistical approaches to quantify risks systematically through a product lifecycle for sound decision making. The work focuses on registered specifications and internal release limits as these are important elements in pharmaceutical development, manufacturing, and supply to ensure product safety, efficacy, and quality. Bayesian inference is explored as a potential valuable approach to enhance risk assessment and related decision making. A Bayesian approach is utilized to predict risks of batch failure and poor process capability associated with registered specifications and internal release limits, leading to a more effective specification setting process. The benefits are demonstrated using a real-life case.

Keywords Risk assessment · Bayesian · Specification · Release limit · Product lifecycle · Process verification · Robustness · Process capability

1 Introduction

In the lifecycle of a pharmaceutical product ("product" herein refers to both drug substance and drug product), decisions need to balance properly the producer's risk and customer's risk (*i.e.* risk to patients) without failing a producer's commitment to patients. Although the process often starts with qualitative evaluations, decision making would be more effective if built on a systematic and quantitative risk assessment of accumulated product knowledge or data through a product lifecycle.

The work was completed at Bristol-Myers Squibb Co. before Dr. Tianhua Wang joined in FDA.

Y. Dong (✉)
Bristol-Myers Squibb Co., 1 Squibb Drive, New Brunswick, NJ 08901, USA
e-mail: yijie.dong@gmail.com

T. Wang
U.S. Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_7

Registered specification limits and internal release limits are important elements in pharmaceutical development, manufacturing, and supply. Registered specification limits are criteria to which a drug substance or drug product should conform to be considered acceptable for its intended use, which are proposed and justified by the manufacturer and approved by regulatory authorities focusing on safety and efficacy. Registered specification limits for a market can be at release and/or at shelf-life per the Health Authority (HA) requirements, between which this work will focus on shelf-life specification limits (SSLs). Internal Release Limits (IRLs) are company-specified limits for range of acceptability at time of manufacture, usually more restrictive than registered specifications in order to ensure that a product remains compliant with the registered specifications with a high confidence through the assigned expiration period.

Setting SSLs and IRLs needs to consider clinical relevance, process performance, analytical variability, and stability behavior, and more importantly, risks associated with the limits to the producer and patients. Although data is presented to justify SSLs and IRLs, the conventional approach of quantifying risks mostly focuses on the worst scenario and reflects only a fragment of data accumulated from product history, e.g. developmental phase 3a for new filings.

This article proposes to pursue advanced statistical approaches to evaluate risks systematically through a product lifecycle. The idea is inspired to better support the current regulatory expectations and industrial trends, including continued process verification, quality by design, and end-to-end (E2E) proactive quality management, for which SSLs and IRLs are decision rules and control mechanisms. Focusing on the specification setting process, a Bayesian approach is explored as a potential useful tool to enhance risk assessments and hence the related decision makings.

2 Decision Making Pertaining to Shelf-Life Specification Limits and Internal Release Limits

The ICH guidelines Q6A [1] and Q6B [2] provide the principles for setting SSLs. Both topics indicate the importance to include four elements: toxicology and clinical studies, long term stability, analytical performance, and manufacturing consistency. The limits should support the E2E robustness from the drug substance release to drug product shelf. In addition, the process needs to consider key business factors, such as needs for future shelf life extensions and commercial time out of refrigeration requirements. The calculated SSLs will be evaluated further based on toxicology and clinical experience, technical/scientific inputs, business perspectives, and filing considerations, which may lead to tightening of the calculated specifications. The proposed SSLs could be further revised during the review by regulators to ensure product efficacy and safety. Note that the above is the current practice balancing product robustness and clinical relevance. There has been an elevating interest in

building clinically relevant specifications, which is fundamentally different from the current common practice.

Starting at the product performance qualification stage, IRLs needs to be evaluated for each round of SSLs considered, which can be the revisions discussed above or any further updates during commercial production due to product evolvements. By definition, IRLs should be back calculated from SSLs to allow for stability change [3], assay variability, and other change and uncertainties over time. Since release test is usually the last check of a batch by the producer, IRLs serve as a decision rule for batch disposition and a critical element in ensuring product quality. In addition, IRLs are appropriate acceptance criteria for process performance as such performance is largely reflected in the CQA results at release. Since IRLs are internal practice, they may be changed to reflect the dynamics of continuously accumulated data and business needs. A commonly used method for calculating IRLs is the approach proposed by Allen et al. [4]. Other approaches were proposed by Shao and Chow [5], Wei [6], and Murphy and Hofer [7].

SSLs and IRLs need to be set at appropriate levels to control both producer's risk and customer's risk. If an Out of Specification (OOS) result is observed, it will trigger various investigations, retests, and potentially lead to a product recall. Furthermore, OOS results present risk to patients and are monitored in Quality Metrics, which collectively may hurt the reputation of a producer and prompt HA actions. Out of internal Release Limit (ORL) results will increase investigation costs and elicit questions about product robustness and quality. A confirmed ORL may cause rejection of a batch, stress inventory and supply, and increase operational cost. Note that both OOS and ORL cases may motivate technical and operational improvements.

Sound risk assessment is evidently vital to make decisions pertaining to SSL and IRL. The assessment needs to address at minimum process capability (against IRLs), probability of OOS and ORL, sources and control of variabilities, and impacts to filing and distribution. In reality, there are challenges to quantify such risks. The most familiar one is the limited amount or history of data, which is a typical situation for new product filings and frequently for method or process updates of a mature product. For instance, it is common to use three long-term stability study (LTSS) batches for registration with no reference to other developmental stability or kinetic studies; then the LTSS data remains as the main source of stability changes for setting SSLs and IRLs until a substantial amount of data from commercial stability batches is accumulated.

Ideally, a producer can leverage knowledge and data accumulated during years of developmental work and production to define the risks quantitatively. In the case of SSL and IRL, key questions around risks include:

- What would be the probabilities of observing OOS and/or ORL?
- What would be the predicted process capability in commercial manufacturing?
- What areas should be improved if the OOS and/or ORL risks are high and which factors should be targeted to ensure the risk mitigation plan is sustainable: e.g., analytical or process variability?

- Can the sources of variability be decomposed and controlled in development and manufacturing?

If these questions can be answered with accuracy and confidence, one can expect more robust business decisions and more targeted improvement in investments.

Two classic frequentist approaches have been tempted. First, confidence level and coverage level are commonly included in the formulas for SSL and IRL calculation, which gives a rough risk alert to the boundary situations: for instance, if a batch were released at the IRL value, what would be the chance that the batch mean will remain within the specifications. It fails to answer questions that are more relevant to decision making in practice: what the overall risks, e.g. OOS/ORL probabilities, based on the product performance are. Second, frequentist simulation propagating multiple parameters may help to quantify the overall risks, but with little reference to historical information or uncertainties in the simulation parameters, such as degradation rate uncertainty, analytical precision, and process variability in the SSL/IRL case.

To overcome the limitations of the frequentist approaches, Bayesian inference is considered as a framework for making consistent and sound decisions in the face of uncertainties and evolvements [8]. Bayesian paradigm offers philosophical consistency by structuring data and knowledge as in the natural learning process: use new evidence to update beliefs through the application of Bayes' rule. Moreover, Bayesian updating can apply the Bayes' rule interactively: a new posterior probability can be computed from new evidence using the previous posterior probability as a prior for the current cycle. Considering that the product lifecycle dynamically generates sequences of data, the iterative learning capability of Bayesian technique is particularly important. Furthermore, Bayesian inference improves evaluation of uncertainties and therefore risk assessments by evaluating the distribution of each parameter rather than often an optimum point estimate per parameter in frequentist approaches. Modern computational methods such as Markov Chain Monte Carlo allow drawing samples from a joint posterior distribution. The sampled results can then be used to estimate distributions of the model parameters and to predict future observations or to assign probabilistic statements to possible decisions (*i.e.* how each alternative decision is affected by the uncertainties in model parameters), which may provide direct and reliable answers to risk and prediction questions.

In this work, such potential benefits are demonstrated with a real-life case for a small molecule potency test. Statistical approaches are applied with two primary aims: (1) to assess the probabilities of OOS/ORL occurrence; and (2) to predict process capability during commercial manufacturing.

3 Case: The Potency Test for a Recent Approved Small Molecule Product

The purpose is to evaluate how Bayesian inference performs in estimating OOS/QRL risks and in predicting process capability.

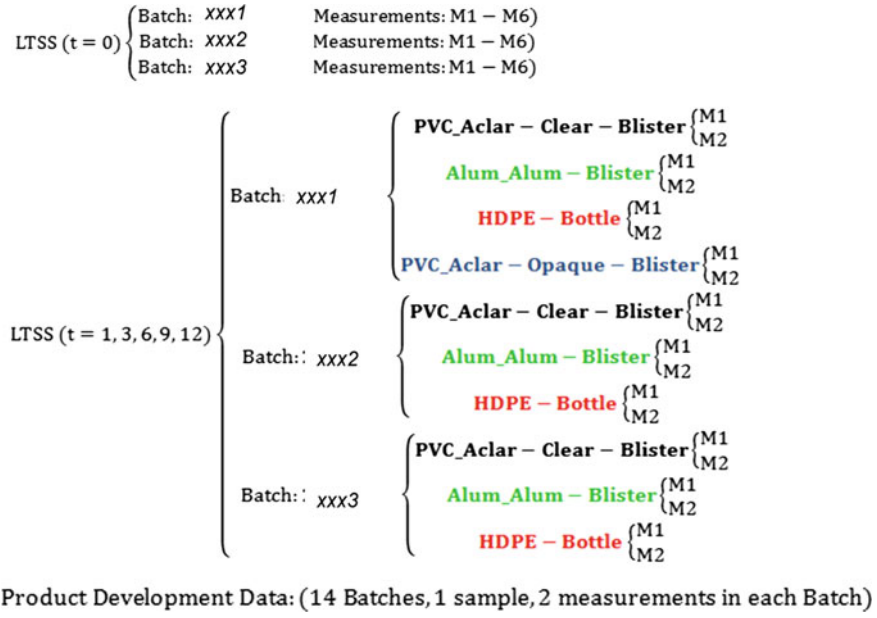


Fig. 1 Data structure of the data at filing in the case

Since the product was recently approved, data from the initial filing (17 batches as in Fig. 1) and data from the subsequent commercial production after the filing approval (27 commercial batches) are available. The structure of the data in the filing is shown in Fig. 1. There are three LTSS batches; and three or four packages were presented in each batch. Six measurements M_1 to M_6 were taken at the release time point ($t = 0$) and two measurements (M_1 and M_2) were taken at other testing time points for each package per LTSS batch. The other 14 batches at filing were manufactured during product development, with two measurements taken at the release time points ($t = 0$) for each batch.

As shown in Fig. 2, a total of four hierarchical Models are developed based on the data structure in Fig. 1. Note that the zero-order kinetic model is assumed in the analysis of the LTSS stability data. Collectively, the likelihood functions and posterior distribution are defined as in Fig. 3.

In the model as described in the case above, the uniform non-informative priors are used for the parameters. Due to the complexity of the model case, it is rather difficult to derive the Jeffreys' priors. The choice of uniform non-informative priors herein has little impact on the convergence of the MCMC chains and consequently the results in this case because most of the scale parameters are not of interest in the final analysis (nuisance parameters). The models are incorporated and simulated using WinBUGS with subsequent analysis performed in R. Based on the posterior distribution, MCMC simulations are performed to get statistical references for the parameters of interest. Three chains are used with n.iter = 20,000 iterations. The

At t = 0; let i = 1,2,3 (3 Batches) and j = 1(1 sample in each Batch) and k = 1,2, … , 6 (6 measurements in one sample within each batch)	$Y_{0ijk} \sim N(\mu_{0ij}, \tau_{LS}) \begin{cases} \mu_{0ij} \sim N(\mu_{0i}, \tau_L) \begin{cases} \mu_{0i} \sim N(\mu_g, \tau_g) \\ \tau_L = \frac{1}{\sigma_L^2} \end{cases} \\ \tau_{LS} = \frac{1}{\sigma_{LS}^2}; \end{cases}$
At t = 1, 3, 6, 9, 12; let i = 1, 2, 3 (3 Batches) and j = 1, 2, 3(3 samples/packages in each Batch) and k = 1,2(2 measurements in one sample within each batch)	$Y_{tijk} \sim N(\mu_{tij}, \tau_{LS}) \begin{cases} \mu_{tij} \sim N(\mu_{ti}, \tau_L) \begin{cases} \mu_{ti} = \mu_{0i} - \beta_i t \\ \tau_L = \frac{1}{\sigma_L^2} \end{cases} \\ \tau_{LS} = \frac{1}{\sigma_{LS}^2} \end{cases}$
At t = 6, 12; Batch: xxx1 (i = 1) has 1 extra sample/package (j = 4); k = 1,2 measurements.	$Y_{t14k} \sim N(\mu_{t14}, \tau_{LS}) \begin{cases} \mu_{t14} \sim N(\mu_{t1}, \tau_L) \\ \tau_{LS} = \frac{1}{\sigma_{LS}^2} \end{cases}$
Product Development Data: 14 Batches (i = 1, 2, … , 14), 1 sample (j = 1), 2 measurements (k = 1, 2) in one sample within each batch, (t=0).	$Y_{0ijk}^{PD} \sim N(\mu_{0ij}^{PD}, \tau_{LS}) \begin{cases} \mu_{0ij}^{PD} \sim N(\mu_{0i}^{PD}, \tau_L) \begin{cases} \mu_{0i}^{PD} \sim N(\mu_g, \tau_g) \\ \tau_L = \frac{1}{\sigma_L^2} \end{cases} \\ \tau_{LS} = \frac{1}{\sigma_{LS}^2} \end{cases}$
Non-informative priors: $\mu_g \sim N(100, 10^{-6})$; $\tau_g = \frac{1}{\sigma_g^2}$ and $\sigma_g^2 \sim U[0, 1000]$ $\sigma_L^2 \sim U[0, 1000]$ and $\sigma_{LS}^2 \sim U[0, 1000]$ $\beta_i \sim N(0, 10^{-6})$ Where t is the stability testing time in month, μ denotes parameter means, τ and σ denote parameter uncertainties, β_i denotes the fixed effect stability change rate for the i^{th} batch, g indicates between-batch process component, L indicates within-batch process component, and LS indicates measurement component. Y_{tijk} is the k^{th} measured value for the j^{th} sample within the i^{th} batch at the time point t; μ_{tij} is the mean value for the j^{th} sample within the i^{th} batch at the time point t; τ_{LS} is the measurement precision (σ_{LS}^2 as variability) for each sample measurement within a batch; μ_{ti} is the mean value for the i^{th} batch at the time point t; τ_L is the sample precision (σ_L^2 as variability) within a batch; μ_g is the mean value for all batches at the release time point; τ_g is the between batch precision (σ_L^2 as variability) at the release time point. In the case of Product Development Data, the "PD" is added to the notations and other meanings of the notations keep the same.	

Fig. 2 Bayesian hierarchical models built based on the case data structure in Fig. 1

length of burn-in is specified to be 5000. Most of the initial values of the parameters in the simulation chain are randomly chosen but cover the possible practical values. Based on the posterior distribution, the results at release and various stability testing stations (*i.e.*, 0, 3, 6, 12… months according to the stability test plan) are sampled using the same model structure. Using the sampled results, product performance predictions are made for various scenarios based on the simulation values after burn-in. For example, if the release batch values (Y_0) and the batch values at the shelf life time (Y_S) are of interest, the joint distributions of Y_0 and Y_S can be simulated using the sampled results based on the posterior distribution of the model.

$$L1 = \left[\prod_{i=1}^{3}\prod_{j=1}^{1}\prod_{k=1}^{6} P(Y_{0ijk}|\mu_{0ij}, \tau_{LS})\right] \times \left[\prod_{i=1}^{3}\prod_{j=1}^{1} P(\mu_{0ij}|\mu_{0i}, \tau_L)\right] \times \left[\prod_{i=1}^{3} P(\mu_{0i}|\mu_g, \tau_g)\right]$$

$$L2 = \left[\prod_{t\in\{1,3,6,9,12\}}\prod_{i=1}^{3}\prod_{j=1}^{3}\prod_{k=1}^{2} P(Y_{tijk}|\mu_{tij}, \tau_{LS})\right] \times \left[\prod_{t\in\{1,3,6,9,12\}}\prod_{i=1}^{3}\prod_{j=1}^{3} P(\mu_{tij}|\mu_{0i} - \beta_i t, \tau_L)\right]$$

$$L3 = \left[\prod_{t\in\{6,12\}}\prod_{k=1}^{2} P(Y_{t14k}|\mu_{t14}, \tau_{LS})\right] \times [P(\mu_{t14}|\mu_{01} - \beta_1 t, \tau_L)]$$

$$L4 = \left[\prod_{i=1}^{14}\prod_{j=1}^{1}\prod_{k=1}^{14} P(Y^{PD}_{0ijk}|\mu^{PD}_{0ij}, \tau_{LS})\right] \times \left[\prod_{i=1}^{14}\prod_{j=1}^{1} P(\mu^{PD}_{0ij}|\mu^{PD}_{0i}, \tau_L)\right] \times \left[\prod_{i=1}^{14} P(\mu^{PD}_{0i}|\mu_g, \tau_g)\right]$$

Joint Posterior Distribution:

$$\pi\left(\{\mu_{0ij}\}_{\substack{i=1,2,3;\\ j=1}}, \{\mu_{0i}\}_{i=1,2,3}, \{\mu_{tij}\}_{\substack{t\in\{1,3,6,9,12\}\\ i=1,2,3;j=1,2,3.}}, \{\mu_{t14}\}_{t\in\{6,12\}},\right.$$

$$\left.\{\mu^{PD}_{0ij}\}_{\substack{i=1,\ldots,14;\\ j=1}}, \{\mu^{PD}_{0i}\}_{i=1,\ldots,14}, \tau_{LS}, \tau_L, \mu_g, \tau_g, \{\beta_i\}_{i=1,2,3} \,|\text{All Data}\right)$$

$$\propto L1 \times L2 \times L3 \times L4 \times (\text{Non} - \text{informative Priors})$$

Fig. 3 The likelihood functions and the posterior distribution in the case

The OOS and ORL risks can be calculated jointly using these Bayesian sampled results. Figure 4 presents the probability of OOS at 24 month as an example. Based on this analysis, the approved SSL of 93.0–105.0% (*i.e.* the SSL endorsed by regulators) and the IRL of 95.0–104.5% at the time of filing is acceptable. As highlighted by the dashed box, the combination controls the OOS risks to a minimal level. However, if business inputs indicated that a higher risk level could be tolerated, more stringent limits could have been considered if the proposal were pushed back by agencies or internal stakeholders.

Process capability index (Cpk or long term Cpk) and its credibility interval (CI) are calculated for various IRLs as summarized in Fig. 5. Specifically, two scenarios in this case are examined for different number of batches (denoted by N):

- N = 27, the available number of commercial batches reflecting the current manufacturing and analytical performance.
- N = 200, an estimated total number of batches in the product life assuming the manufacturing and analytical performance will remain the same.

The graph illustrates two points aligning with statistical intuitions. First, the tighter the IRL, the worse the process capability, because the acceptance criteria (*i.e.* the IRLs if set appropriately) for the process becomes more stringent. Second, the larger the number of batches, the narrower the 95% CI of capability index, which reflects increasingly the long-term performance. The commercial data of 27 batches is used to validate the Bayesian results. Process capability index is generated when the number

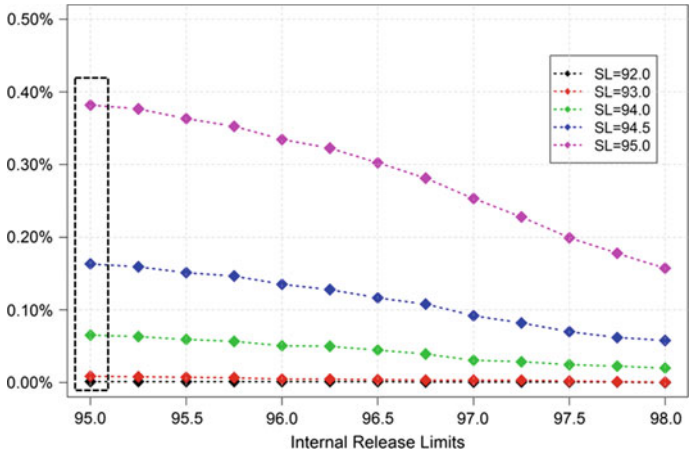


Fig. 4 Probability of OOS occurrence at 24 months for combinations of shelf-life specifications and internal release limits in the case. (SL = Shelf-life Specification Limits; unit for the limits is % of label.)

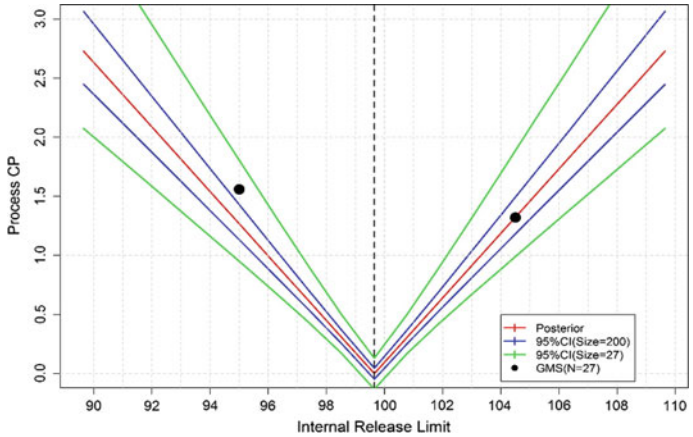


Fig. 5 Process capability analysis in the case (Red lines are capability indexes for different internal release limits. Size denotes the number of batches simulated. Green and blue lines are 95% credibility intervals (CI) of the capability index for Size = 27 and Size = 200, respectively. Black dots are the capability indexes calculated using the data from N = 27 based on the implemented internal release limits of 95.0 to 104.5% of label)

of batches is greater than 25 for process robustness monitoring. In Fig. 5, the black dot is the process capability index reported on the available 27 commercial batches using the classic method, which is within the 95% CI (green lines) from the Bayesian prediction for N = 27. The results indicate that the Bayesian process capability prediction is acceptable for the early commercial production. In the scenario of N = 200, the black dot falls out of the 95% CI for N = 200 at the lower side, which is more likely because data of the 27 commercial batches is not fully representative of long term commercial production performance. The Bayesian approach offers the potential to reduce such inaccuracy in prediction, which needs to be confirmed by further Bayesian updating and ideally incorporation of prior development knowledge.

4 Discussion

The real-life case indicates that Bayesian inference has the potential to improve confidence in risk predictions to inform the specification/limit setting process, even only based on the limited data used in the filing. The use of non-informative prior distributions yields results in line with those from conventional statistical analysis, which is expected as the information from the evidence, *i.e.* the simulated data, dominates the not very informative prior.

The posterior distributions can be used as a prior distribution in the next round of Bayesian modeling, in which the pre-existing evidence is taken into account in the new prior. Through iterative updating, the posterior is determined increasingly by the evidence rather than any original prior distribution as data accumulates. If employed through a product lifecycle, Bayesian updating offers the potential to capture the knowledge accumulated through development and commercial production, which will address the limitations of frequentist approaches. Starting with non-informative priors may help to mitigate the frequent doubt about the Bayesian modelling: analysis manipulation by designing the prior.

Even in a single round application as in the real-life case herein, the Bayesian approach generates more robust risk assessment. First, the joint distribution of release data and 24-month stability values can be carefully examined as shown in Fig. 6. Since the batch number for each simulated observation can be tracked, the OOS/ORL risks can be reviewed in depth and categorized. One may try various combinations of SSL and IRL to reduce the probability of false rejection and false acceptance. Moreover, distribution of parameters can be derived and used for investigation and for continuous improvements. For instance, in the real-life case discussed, the analytical variability, between batch process variability, and within batch process variability can be better defined, monitored, and improved as needed.

While the approach generates the probabilities of events, the resulting cost or benefit can be added relatively easily to compare different risk options. For example, the costs of a typical OOS investigation, a retest, and a batch rejection can be estimated and even incorporated as parameters in the modelling. Nevertheless, risks can be interpreted from multiple perspectives. For instance, a higher ORL probability

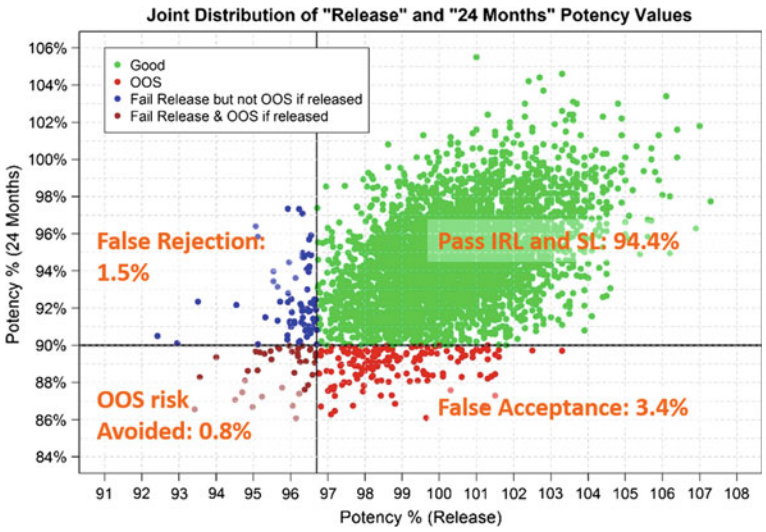


Fig. 6 Illustration of joint distribution of the simulated release results and 24-month stability results for the potency test from a Bayesian approach

translates into (1) a lower capability which is not a favorable scenario to production; (2) potentially lower OOS occurrence and less other potential consequences such as recalls, which is positive from a quality perspective. Therefore, the decision rules need to be aligned among business units, preferably to be standardized for consistency and, most importantly, to balance producer's risk and customer's risk but still fulfill commitments to patients.

If implemented, a Bayesian approach needs to demonstrate consistency and robustness. To convince stakeholders, structured processes need to be developed to drive best practices among practitioners. Furthermore, collaboration across stages and functional areas are critical to integrate knowledge and data accumulated into defendable prior distributions in Bayesian updating and to align on the decision framework.

5 Conclusion

This work shares vision and effort to improve pharmaceutical risk assessments, specifically focusing on risks associated with shelf-life specification and internal release limits as important factors in development, production, and supply. A Bayesian approach is considered for its inherent advantages in incorporating evolving knowledge, estimating uncertainties, and predicting risks. The values of the concept are discussed and preliminarily demonstrated using a real-life case. The Bayesian

paradigm can enhance effectiveness of specification setting process and optimize improvement opportunities to ensure process, analytical, and product robustness.

Based on the proof of concept, Bayesian inference is proposed as a potentially useful tool to quantitatively build systemic knowledge through product lifecycle. The approach can be vital to delivering on regulatory expectations and industrial trends, including continued process verification, quality by design, and E2E proactive quality management.

Therefore, the plan is to further evaluate performance of Bayesian approaches, particularly Bayesian updating, in real-life situations, to expand from potency to other critical quality attributes, and to broaden the application from specification setting to other decision processes. To fully realize the potential value of the approach, it requires engagement in quantitative modeling from development to production, a structured framework with aligned processes and rules, as well as seamless collaborations across various functional areas.

Acknowledgements The work was first presented at the 2015 Nonclinical Biostatistics Conference in October 2015 and then 2016 Midwest Biopharmaceutical Statistics Workshop in May 2016. The authors thank Mary Ann Gorko for guidance and insightful review as the 2016 Midwest Biopharmaceutical Statistics Workshop CMC Section Chair. They acknowledge Dr. Jose Tabora and Dr. Renfei Yan for support to developing the case for demonstration. They also appreciate Dr. Siheng He, Joel Young, Dr. Ronald Behling, Jennifer Walsh, Dr. Lindsay Hobson, and Don Buglino for helpful discussions pertaining to this work.

References

1. ICH Guideline Q6A: Specifications: test procedures and acceptance criteria for new drug substances and new drug procedures: chemical substances (1999)
2. ICH Guideline Q6B: Specifications: Test procedures and acceptance criteria for biotechnological/biological products (1999)
3. ICH Guideline Q1E: Evaluation of stability data (2003)
4. Allen, P.V., Dukes, G.R., Gerger, M.E.: Determination of release limits: a general methodology. Pharm. Res. **8**(9), 1210–1213 (1991)
5. Shao, J., Chow, S.C.: Constructing release targets for drug products: A Bayesian decision theory approach. Appl. Stat. 381–390 (1991)
6. Wei, G.C.: Simple methods for determination of the release limits for drug products. J. Biopharm. Stat. **8**(1), 103–114 (1998)
7. Murphy, J.R., Hofer, J.D.: Establishing shelf life, expiry limits, and release limits. Drug Inf. J. **36**(4), 769–781 (2002)
8. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian data analysis, vol. 2. CRC Press, Boca Raton, FL (2014)

Development of Statistical Computational Tools Through Pharmaceutical Drug Development and Manufacturing Life Cycle

Fasheng Li and Ke Wang

Abstract Statisticians at Pfizer who support Chemistry, Manufacturing, and Controls (CMC), and Regulatory Affairs (Reg CMC) have developed many statistical R-based computational tools to enable high efficiency, consistency, and fast turnaround in their routine statistical support to drug product and manufacturing process development. Most tools have evolved into web-based applications for convenient access by statisticians and colleagues across the company. These tools cover a wide range of areas, such as product stability and shelf life or clinical use period estimation, process parameter criticality assessment, and design space exploration through experimental design and parametric bootstrapping. In this article, the general components of these R-programmed web-based computational tools are introduced, and their successful applications are demonstrated through an application of estimating a drug product shelf life based on stability data.

Keywords Statistical computing · Web based tool · R programming · Regression analysis · Design of experiment

1 Introduction

Through the regulatory, chemistry, manufacturing and controls (Reg CMC) development lifecycle of a drug product, a series of compendial requirements, quality standards, and performance criteria must be well established and met. It usually takes years to perform data collection, analysis, and reporting on chemical process, formulation and manufacturing process, and analytical method development. Common practice in current pharmaceutical industry to "optimize" product compositions,

F. Li (✉) · K. Wang
Pharmaceutical Science and Manufacturing Statistics, Pfizer Inc., MS 8220-2356, Eastern Point Road, Groton, CT 06340, USA
e-mail: fasheng.li@pfizer.com

K. Wang
e-mail: ke.wang2@pfizer.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*, Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_8

manufacturing processes, and analytical methods is to apply designed experiments (DOEs), statistical models and statistical sampling techniques. Data generated in these procedures, which could be in a large amount, are usually analyzed or evaluated by statisticians or statistically trained professionals with commercial statistical software systems such as Design Expert, SAS, or SAS-JMP. With the increasing demand of statistical application and the challenge of limited number of trained statisticians, it is desirable to develop computational tools to conduct routine statistical analyses in more efficient and consistent ways. The computational tools promote consistency, efficiency, and reproducibility for routine statistical analysis. Version control, monitoring and regular maintenance are an integral part of developing the computational tools. The features of the computational tools align well with the requirements of Title 21 Code of Federal Regulations (CFR) Part 11 that the software systems should be readily available for and subject to FDA inspection (3) [1]. Working as statisticians at Pfizer supporting pharmaceutical development and Reg CMC, we have identified many opportunities and areas that benefit from statistical computation tools. Most tools are developed using a language such as R and have evolved into web-based applications for easy access by statisticians and colleagues at Pfizer. This article introduces the general requirements and structure of web-based statistical tools. The computational application is demonstrated through one tool which evaluates product stability and predicts shelf life or clinical use period.

2 Overview of Available Web-Based Statistical Tools

2.1 Introduction of Components of Web-Based Statistical Applications

Figure 1 illustrates three standard components of typical web-based applications: computer server, GUI platform server and application user. In practice, applet authors utilize the application servers to construct the computation script and the graphical user interface (GUI) of the application, and ensure successful communication between the application servers (usually a web browser) and the computer servers. General users only need to compile data into a required format by the applications. For a statistical computational application, additional software systems, such as R and SAS need to be installed onto the computer server for statistical analysis. The following section provides an overview of the web-based statistical applications developed by Pfizer pharmaceutical development and Reg CMC statisticians.

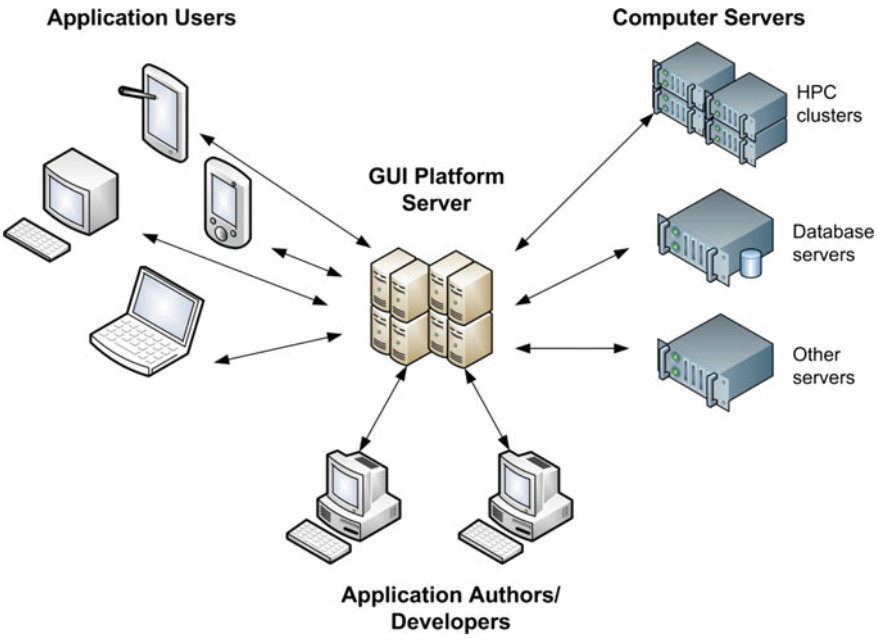


Fig. 1 Components of a typical web-based application

2.2 Overview of Web-Based CMC Development/Regulatory Statistics Applications

Most computational tools developed at Pfizer to support analytical method, product, and process development are written in script codes using R, SAS, MATLAB, JMP, Minitab, or MS Excel spread sheet templates. One example is drug product shelf life prediction. Long-term stability data are collected under various storage conditions, per ICH Q1A (2) and are evaluated per ICH Q1E (1) [2, 3]. The statistical analysis is coded in SAS and R to generate summary results and plots.

The commercial software packages, nevertheless are important tools for statisticians to carry out data analysis. However, individual usage of the software presents issues in portability, limited version control, and reproducibility. With support from Pfizer Information Technology group, statisticians have been able to turn the individual pieces of code into web-based applications. Figure 2 illustrates various web-based statistical applications developed by the CMC statisticians at Pfizer and the targeted areas throughout the life cycle of drug development and manufacturing. These applications are searchable and accessible to Global Pfizer colleagues.

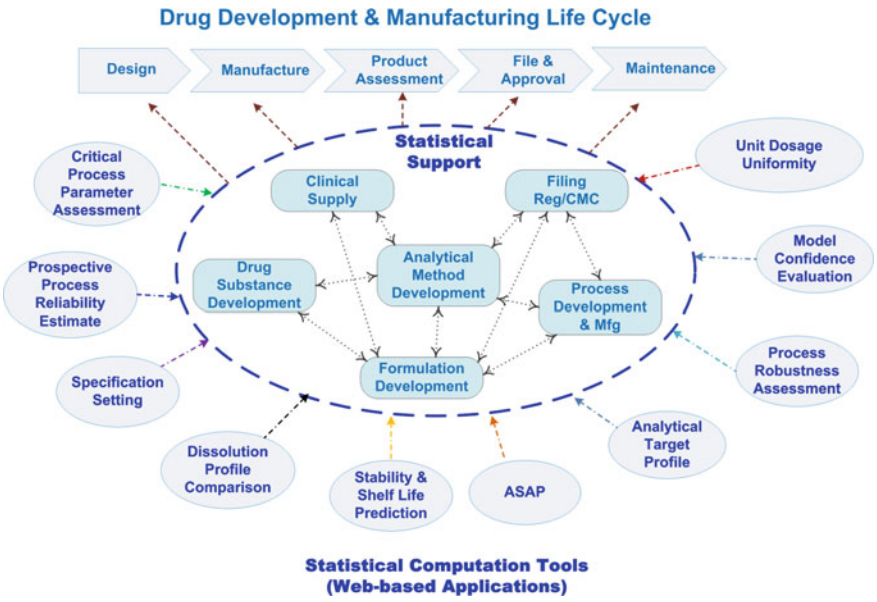


Fig. 2 Examples of statistical computational web-based applications throughout drug development and manufacturing life cycle

3 An Example Web-Based Statistical Computation Tool

Below, details are provided on the development and usage of one of the web-based applications listed in Fig. 2, Stability & Shelf Life Prediction.

For this application, assume that the stability data are collected from a registration stability program that follows ICH Q1A guidelines or a clinical stability program. Most stability programs have three registration batches per combination of strength, packaging configuration, and storage condition, whereas clinical stability program usually has only one batch. The online application of analyzing stability data is programmed in R, following ICH Q1E guidance for a specific combination of product, strength, package, and storage condition. The shelf life is determined by the decision criteria in the guidance. The clinical stability data is analyzed using a simple linear regression model, and the use period is determined, according to an internal criterion. For example, the use period of a clinical material is the shorter of the intersection of the 95% confidence interval and the specification limit or real stability time plus 12 months or longer if statistically supported. Therefore, the shelf life or clinical use period can be determined by a two-step procedure: model selection and projection of shelf life/use period.

3.1 Statistical Model Selection

For the statistical analysis of typical registration stability data, the following model selection procedure is performed based on the poolability of the data from the three batches. Assume $Y_b = y_{b1}, y_{b2}, \ldots, y_{bT}$ are the stability data for an attribute at time period $t = 1, 2, \ldots, T$ months for batch $b = 1, 2, \ldots, B$ for a certain combination of strength, package type, and storage condition.

(a) Fit a full model (the SSSI model—separate slopes and separate intercepts model):

$$y = \beta_0 + \beta_1 * time + \beta_{21} * batch + \beta_{12} * time * batch + \varepsilon \quad (1)$$

where the error Ɛ is normally distributed with mean 0, and standard deviation σ. This model is referred to as the separate slopes and separate intercept model (SSSI), as it allows for different slopes and different intercepts for each batch.

Decision: If the p-value of the interaction of time and batch (time*batch) is <0.25, STOP and use Eq. (1) for the shelf life projection; if the p-value of the interaction of time and batch (time*batch) is ≥0.25, GOTO step (b).

(b) Fit a reduced model (the CSSI model—common slope and separate intercepts model):

$$y = \beta_0 + \beta_1 * time + \beta_{21} * batch + \varepsilon \quad (2)$$

This model is referred to as a common slope and separate intercepts model (CSSI), as it permits the same slope estimate but different intercepts for all batches.

Decision: If the p-value of batch is <0.25, STOP and use Eq. (2) for the shelf life projection; if the p-value of batch is ≥0.25, GOTO step (c).

(c) Fit a reduced model (the CSCI model—common slope and common intercept model):

$$y = \beta_0 + \beta_1 * time + \varepsilon \quad (3)$$

This model is referred to as a common slope and common intercept model (CSCI), since the same slope and intercept are used for all batches.

Decision: Eq. (3) is used for the shelf life projection.

The above described procedure for the statistical analysis of long-term registration stability data is summarized into a flow chart in Fig. 3. For typical one-batch clinical stability data, a simple linear regression model is used.

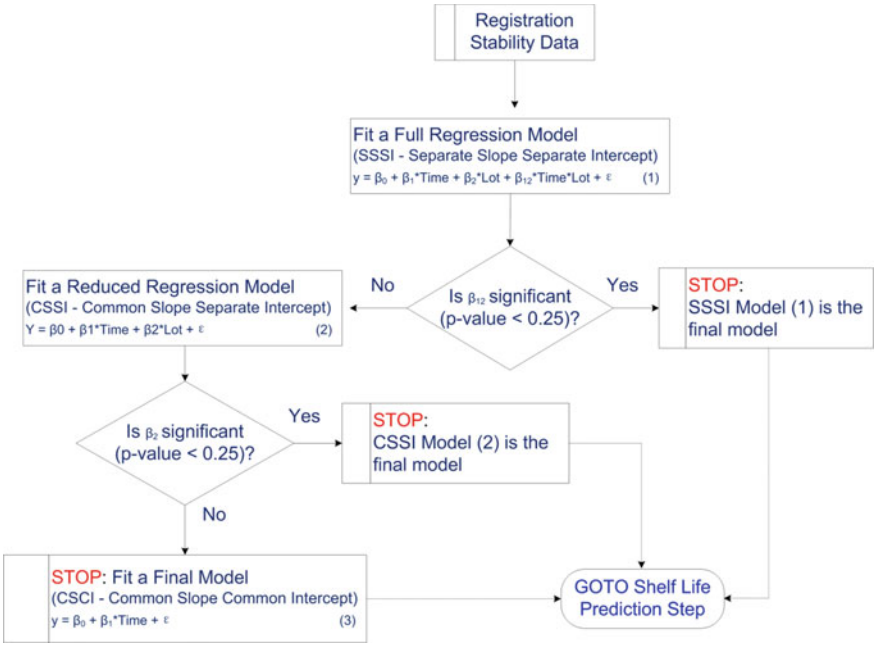


Fig. 3 Typical regression model selection per ICH Q1E stability data analysis

3.2 Shelf Life or Use-Period Projection

Once the regression model is determined, the 95% confidence interval (CI) can be calculated for any stability time point. The predicted shelf life/use period is determined as the shortest time point when the confidence limit intersects with the specification limit of the product. Notice that it is necessary to extrapolate the predictions and 95% CIs in order to determine the shelf life/use period beyond the maximum storage time of the stability data. Per ICH Q1E, the maximum extrapolation is two times of the maximum storage time (T_{max}) when T_{max} is <12 months or an extrapolation of 12 months when T_{max} is > = 12 months. Figure 4 illustrates how to establish the shelf life for an example data set. For this set of stability data, a separate slope and separate intercept model is selected and the shelf life is determined by the limiting lot (i.e. Lot 3). This shelf life limiting lot is determined, due to its fastest impurity A growth (largest slope) and thus its 95% CI intercepts with the specification limit of 1%, the earliest at 32.1 months. Therefore, 32.1 months (or 32 months) is the longest shelf life can be proposed. Practically, a shelf life of either 24 months or 30 months can be proposed for this product based on this set of data.

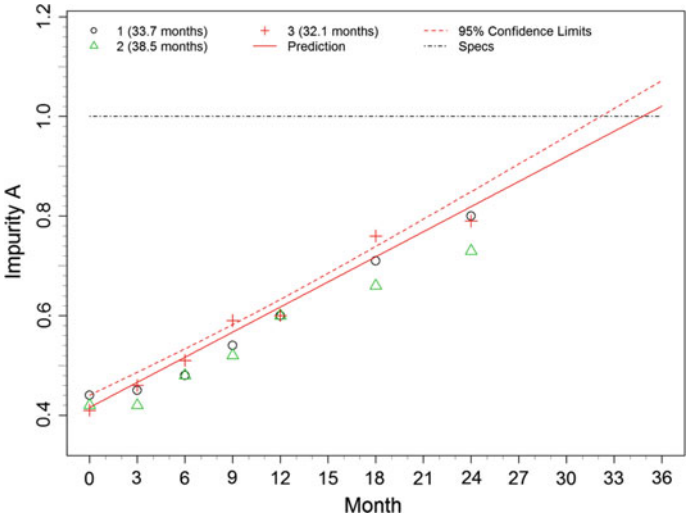


Fig. 4 Prediction of product shelf life based on regression model per ICH Q1E stability analysis: the predicted shelf life is the interception point (i.e., 32.1 months) of the upper 95% confidence limit with specification limit (i.e. the upper limit 1.0%)

3.3 The Internal Web-Based Online Application

Both long-term registration stability data and clinical stability data are collected routinely for all filed products. The repeated stability data analysis, including stability data plotting and drug product shelf life prediction, necessitated the development of a web-based application tool to standardize these statistical activities.

The web application for Registration and Clinical Stability Data Analysis and Shelflife/Use Period Prediction is programmed in R. A graphical user interface (GUI) is built to allow users to upload the relevant stability data to the program for analysis. The GUI of this application is displayed in Figs. 5 and 6 where the main interface contains links to various features, such as the user manual, example data sets in required formats, dialogues for uploading data, and choices of analyses.

Once stability data is uploaded and choices of statistical analyses and parameters are determined, the job is submitted and run in the background through the HPC computing cluster. As soon as the job is finished, users can view the results (including tables and graphical plots) through the web browser (e.g., Internet Explorer, Chrome). The application also provides the ability to download tables and graphs as well as consolidating the results in a .pdf formatted report. Figure 6a, b are snapshots of the output on a web browser.

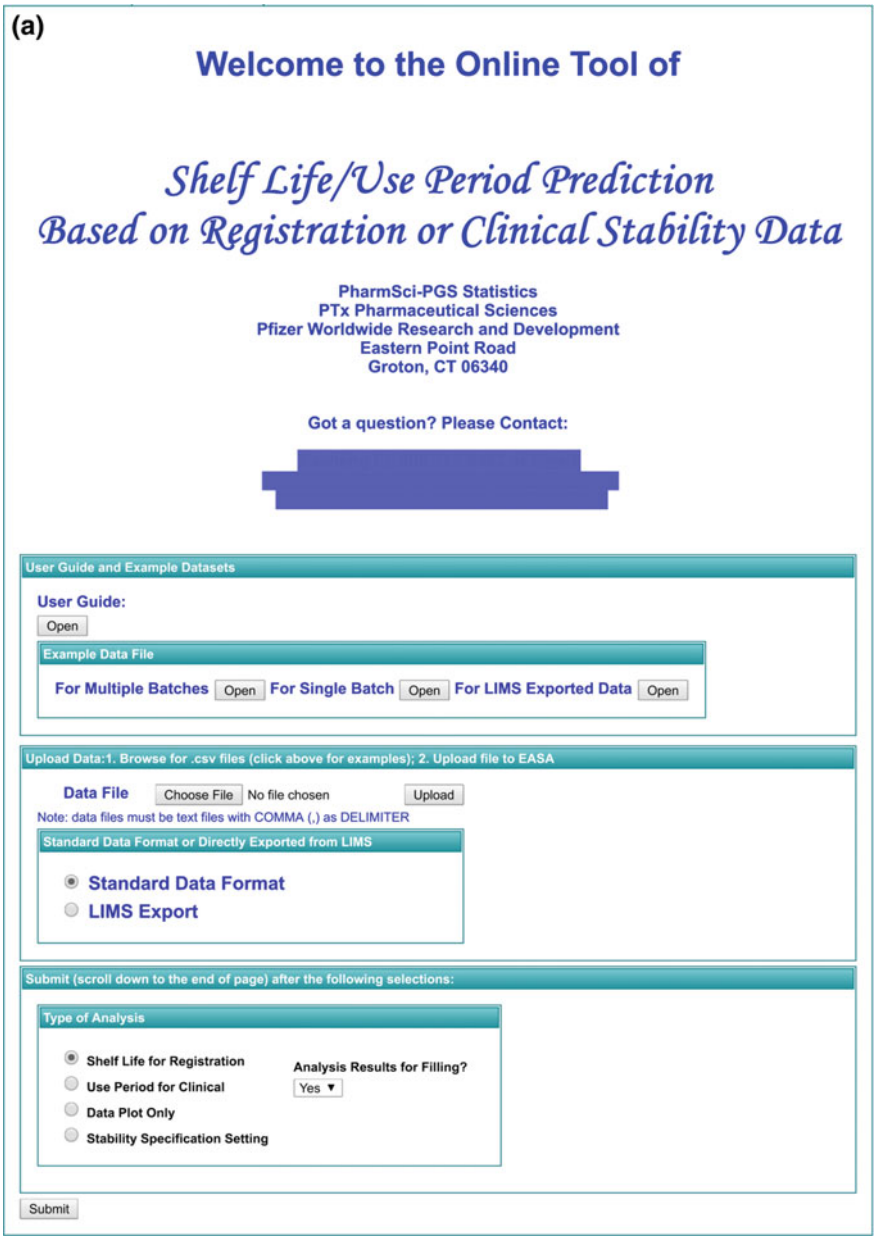


Fig. 5 **a** Web-based application—registration and clinical stability data analysis and shelflife/use period prediction: GUI—main interface **b** Web-based application—registration and clinical stability data analysis and shelflife/use period prediction: GUI—further dialogues

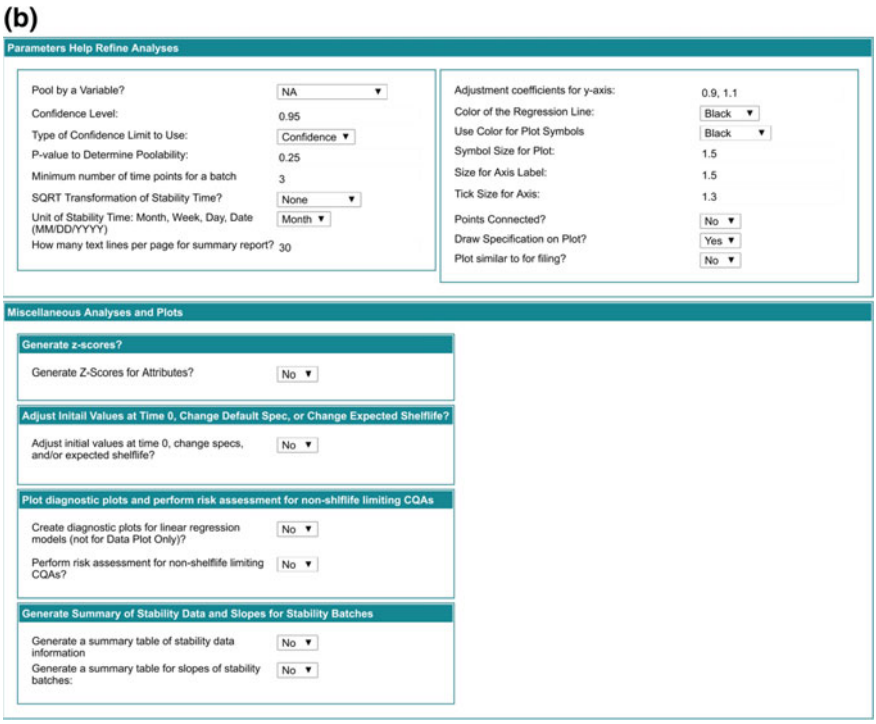


Fig. 5 (continued)

In summary, the implementation of the web-based statistical application of "registration and clinical stability data analysis and shelflife/use period prediction" is able to offer benefits and features such as,

- Align the statistical analyses of long term stability data
- Offer quick and convenient turnaround to analyze stability data, to generate shelf life plots and tables, and summary report
- Allow easy maintenance for feature updates due to the version controlled R program
- Run jobs in the background on HPC cluster or cloud computers.

4 Conclusions

The benefits and features of web-based statistical applications have been demonstrated through a selected program "registration and clinical stability data analysis and shelflife/use period prediction". Statisticians and scientists supporting drug development and Reg CMC areas can offer their routine statistical activities with

(a)

1 . Data Read-In

Data_Analyzed.csv (click to Open or Save)

	PRODUCT	STRENGTH	STORAGE CONDITION	PACKAGE	LOT	METHOD	TIME	RESULT	SPECIFICATION	EXPIRY
1	Product B	x mg	25C/60%RH	bottle	Lot 1	assay	0	100	95-105	24
2	Product B	x mg	25C/60%RH	bottle	Lot 1	assay	3	98	95-105	24
3	Product B	x mg	25C/60%RH	bottle	Lot 1	assay	6	98	95-105	24
4	Product B	x mg	25C/60%RH	bottle	Lot 1	assay	9	100	95-105	24
5	Product B	x mg	25C/60%RH	bottle	Lot 1	assay	12	99	95-105	24
6	Product B	x mg	30C/65%RH	bottle	Lot 1	assay	0	100	95-105	24

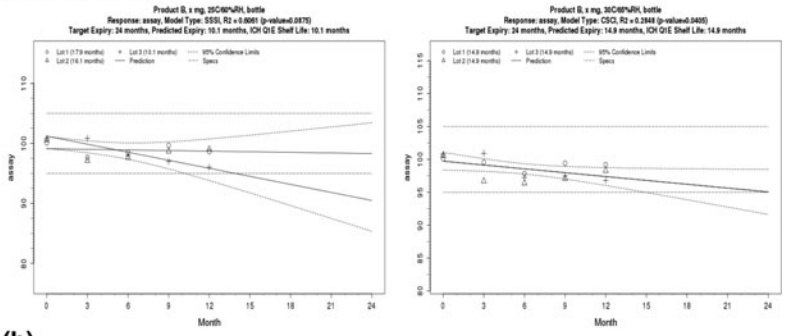
2 . Summary of Shelflife Prediction

ShelfLife_Estimate.csv (click to Open or Save)

	Product	Strength	Storage condition	Package	Response	Model	Lot	Pred Shelf Life	Q1E Shelf Life
1	Product B	x mg	25C/60%RH	bottle	assay	Maximum Allowable	Expiry	10	10
2	Product B	x mg	25C/60%RH	bottle	assay	SSSI	Lot 1	18	18
3	Product B	x mg	25C/60%RH	bottle	assay	SSSI	Lot 2	16	16
4	Product B	x mg	25C/60%RH	bottle	assay	SSSI	Lot 3	10	10

3. Plots of Shelf Life Prediction

ShelfLifePlots.pdf (click to Open or Save)



(b)

4 . Summary of Stability Data Info

Stability_Data_Information_Summary.csv (click to Open or Save)

	PRODUCT	STRENGTH	STORAGE CONDITION	PACKAGE	Lot	Time Point	Attribute	Specification
1	Product B	x mg	25C/60%RH	bottle	Lot 1, Lot 2, Lot 3	0, 3, 6, 9, 12	assay	assay: 95-105
2	Product B	y mg	25C/60%RH	bottle	Lot 4, Lot 5, Lot 6	0, 3, 6, 9, 12	~	assay: 95-105
3	Product B	x mg	30C/65%RH	bottle	Lot 1, Lot 2, Lot 3	0, 3, 6, 9, 12	~	assay: 95-105
4	Product B	y mg	30C/65%RH	bottle	Lot 4, Lot 5, Lot 6	0, 3, 6, 9, 12	~	assay: 95-105

5 . Slopes of CQAs

Stability_Data_Slope_Summary.csv (click to Open or Save)

	PRODUCT	STRENGTH	STORAGE CONDITION	PACKAGE	Response	Model	Lot	Pred Shelf Life	Intercept	Slope	P-Value
1	Product B	x mg	25C/60%RH	bottle	assay	SSSI	Lot 1	18	99	-0.033	0.852
2	Product B	x mg	25C/60%RH	bottle	assay	SSSI	Lot 2	16	99	-0.055	0.758
3	Product B	x mg	25C/60%RH	bottle	assay	SSSI	Lot 3	10	101	-0.447	0.030
4	Product B	x mg	30C/65%RH	bottle	assay	CSCI	Lot 1	15	100	-0.195	0.040

6 . Summary Statistics of CQA Slopes

Stability_Data_Slope_Summary_Statistics.csv (click to Open or Save)

	PRODUCT	STRENGTH	STORAGE CONDITION	PACKAGE	Response	N Batches	n	Mean Slope	Min Slope	Max Slope	sqrt(Total Variance)	sqrt(Between Batch Variance)	sqrt(Within Batch Variance)	Mean+3*StDev
1	Product B	x mg	25C/60%RH	bottle	assay	3	3	-0.18	-0.45	-0.033	0.23	0.22	0.082	0.52
2	Product B	y mg	25C/60%RH	bottle	assay	3	3	-0.15	-0.15	-0.150	0.00	0.00	0.000	-0.15
3	Product B	x mg	30C/65%RH	bottle	assay	3	3	-0.19	-0.19	-0.195	0.00	0.00	0.000	-0.19
4	Product B	y mg	30C/65%RH	bottle	assay	3	3	-0.10	-0.10	-0.105	0.00	0.00	0.000	-0.10

7 . Report of Stability Data Analysis

Stability Data Analysis Report.pdf (click to Open or Save)

Run by: **Fasheng Li**
Date run: **28-Aug-2017 15:16**

Fig. 6 **a** Abbreviated Result—displayed in a browser of the web-based application—registration and clinical stability data analysis and shelflife/use period prediction: Data read-in, shelf life results and plots **b** Abbreviated Result—displayed in a browser of the web-based application—registration and clinical stability data analysis and shelflife/use period prediction: Summary of data, slopes, reports, etc.

increased consistency, improved efficiency, better alignment of statistical analyses, and easily retrievable results by deploying web-based statistical applications. These web-based statistical applications can standardize statistical approaches, centralize software pieces, validate and verify software pieces, and utilize high performance and cloud computer resources.

Acknowledgements The authors acknowledge Kimberly Vukovinsky, Robert J. Timpano, and colleagues in Pharmaceutical Science and Manufacturing Statistics group at Pfizer for their generous support of evaluating and commenting the web-based applications.

References

1. US Food and Drug Administration Department of Health and Human Services. 21 CFR 11: Electronic records; electronic signatures (2017)
2. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Q1E: Evaluation for Stability Data (2004)
3. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Q1A (R2): Stability Testing of New Drug Substances and Products (2003)

Application of Advanced Statistical Tools to Achieve Continuous Analytical Verification: A Risk Assessment Case of the Impact of Analytical Method Performance on Process Performance Using a Bayesian Approach

Iris Yan and Yijie Dong

Abstract The criticalness of robust analytical performance is becoming more and more recognized in the pharmaceutical industry. An effective analytical control strategy needs to be defined, along with a process control strategy, to ensure that the measurement uncertainties are controlled to achieve the intended purposes of analytical methods. The principles of Continuous Process Verification (CPV) have been applied to the lifecycle management of analytical robustness, which leads to our vision of Continuous Analytical Verification (CAV) through a product lifecycle. This work proposes to apply advanced statistical tools to deliver on the vision of CAV. A Bayesian hierarchical modeling approach is a potential solution to integrate a risk-based control strategy into the framework of CAV from design, qualification, to continued verification. A case study is included to illustrate the benefits of a Bayesian-based systematic tool in assessing the impact of analytical performance on process performance and in informing decisions related to analytical control strategy, in order to ensure analytical and process robustness.

Keywords Continuous process verification · Analytical control strategy · Life cycle management · Risk assessment

1 Introduction

To achieve Continuous Process Verification (CPV) following the EU and FDA process validation expectations [1], it is essential to understand, monitor and control significant sources of variabilities that could affect product quality, among which analytical variability is a critical component. From an even broader perspective, robust analytical performance is a prerequisite for robust process performance. Analytical variability is always part of the observed variability in the process outputs and

I. Yan (✉) · Y. Dong
Global Statistics, Bristol-Myers Squibb Co., 1 Squibb Drive, New Brunswick, NJ 08901, USA
e-mail: ennayan@gmail.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_9

is often confounded with other sources of variations, such as materials, equipment, operation, etc. Unsatisfactory analytical performance, and most commonly high analytical variability, can cause negative effects, such as lower process capability and higher risk of out of specification.

The CPV principles can be applied to the lifecycle management of analytical robustness, which leads to our vision of Continuous Analytical Verification (CAV): effective design, qualification, and continued verification of analytical performance through a product lifecycle.

In line with the above efforts (CPV and CAV) to modernize the pharmaceutical development and operation, advanced statistical tools need to be explored and applied for more effective risk assessments. For CAV, such tools would enhance the understanding of major sources of analytical variations and the associated impacts on process performance and quality, and consequently, the effectiveness of a risk-based analytical control strategy.

Bayesian hierarchical modeling is potentially a suitable choice. In comparison to the frequentist approaches, Bayesian provides a structured framework for combining prior information from historical process and analytical data, integrating analytical and manufacturing factors across multiple units and/or stages, and making predictive inferences based on varied hypotheses. Moreover, its continuous learning capability enables verification of risk assessment results and informs decisions related to control strategies through a product lifecycle.

A case study is included to illustrate the development of an effective control strategy for analytical robustness by using a Bayesian approach for risk assessments.

2 Advanced Statistical Modelling to Build Effective Analytical Control Strategy Through Lifecycle

In this session, business needs for a systematic risk assessment approach to achieve analytical robustness are described. A flow of building a risk-based analytical control strategy is presented. Advanced statistical modelling tools, particularly Bayesian, are discussed to leverage their advantages in building an effective control strategy to achieve CAV.

a. **Robustness through Continuous Analytical Verification**

The FDA Process Validation Guidance issued in [1] formalized the framework of CPV to ensure a consistent and reliable delivery of quality product through a product lifecycle. The benefits of the scientific and risk-based approach are more and more recognized across the industry. As all CPV-related data is generated by analytical methods, it is critical to ensure analytical performance through a product lifecycle. Analogous to the CPV model, a CAV model can include the following stages:

Stage 1: Design

- Apply the Quality by Design (QbD) concepts during the design stage [2–4].
  - Identify method and performance expectations: Establish Analytical Target Profile (ATP) to drive the design, development, and validation of appropriate analytical methods.
  - Design and develop optimal method: Further define the analytical method performance criteria in line with the ATP. Enhanced risk assessment and modeling tools are applied to optimize method performance across the entire design space.

Stage 2: Qualification

- Perform a formal method validation or method transfer to demonstrate that the ATP and the analytical method performance criteria is fulfilled.

Stage 3: Continued Verification

- Utilize method performance monitoring tools to assess method performance and identify opportunities for improvement. Revisit the validity of the established ATP and performance criteria on a regular basis as augmented knowledge are gained through development and operation.

Through the three stages, an effective control strategy is required to ensure that the analytical performance, especially for critical quality attributes and critical process parameters, is properly controlled to fit for the intended purposes.

b. **Proposed Risk-based Control Strategy**

A risk-based control strategy can be developed following the procedures in Fig. 1 to achieve CAV.

Define: understand the major sources of variations including analytical variability, and decide on where in the process the analytical methods need to be employed and controlled.

Measure: collect available data from development and manufacturing to enable the evaluation of variations and risk assessments.

Assess: evaluate the variations and quantify the associated risks. When applicable, determine the analytical method control limits that support specifications and the desired process performance across the product design space.

Control: inform the control strategy that is commensurate with the level of risks.

Improve: review the analytical and process control strategy regularly or as needed as more data are being generated from routine monitoring. Increasing knowledge about the method and process may offer improvement opportunities. Events such as major process change, specification revisions, or new sources of variations can

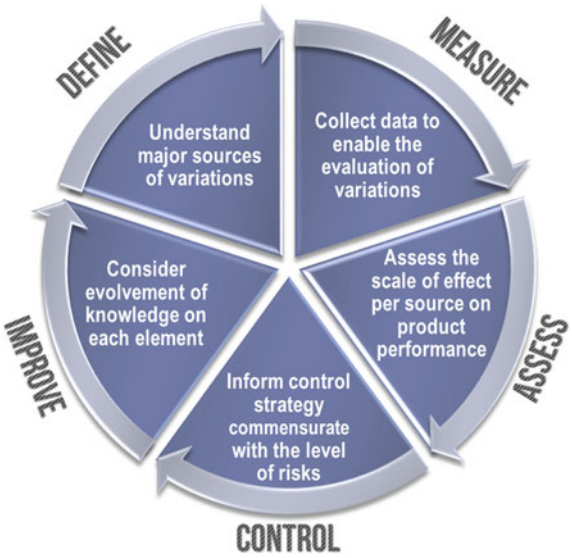


Fig. 1 Build risk-based control strategy for continuous analytical verification

impact the analytical performance. Such events, when they happen, may reduce the effectiveness of the current analytical control and the impacts need to be mitigated.

c. **Bayesian's Advantages in Continuous Analytical Verification**

To build an effective control strategy through CAV, advanced statistical risk assessment tools need to be utilized to assess and predict the impact of the analytical performance over the product manufacturing space. The desired features for the risk assessment tools are depicted in Fig. 2, including:

(1) Integrate the effects that impact process performance and analytical performance;
(2) Evaluate properly the impacts of major sources of variations, which have their own inherent uncertainties and collectively explain the risks;
(3) Predict reliably analytical and process performances and quantify the producer's risk and customer's risk;
(4) Learn from the prior knowledge accumulated through development to production, and provide a systematic way to update the above analyses and risk assessment.

Bayesian hierarchical modeling is a potential solution with the above desired features. In this work, we propose a Bayesian-based systematic approach to assess the impact of analytical performance on process performance, and to inform decisions related to analytical control strategy to ensure analytical robustness and process robustness.

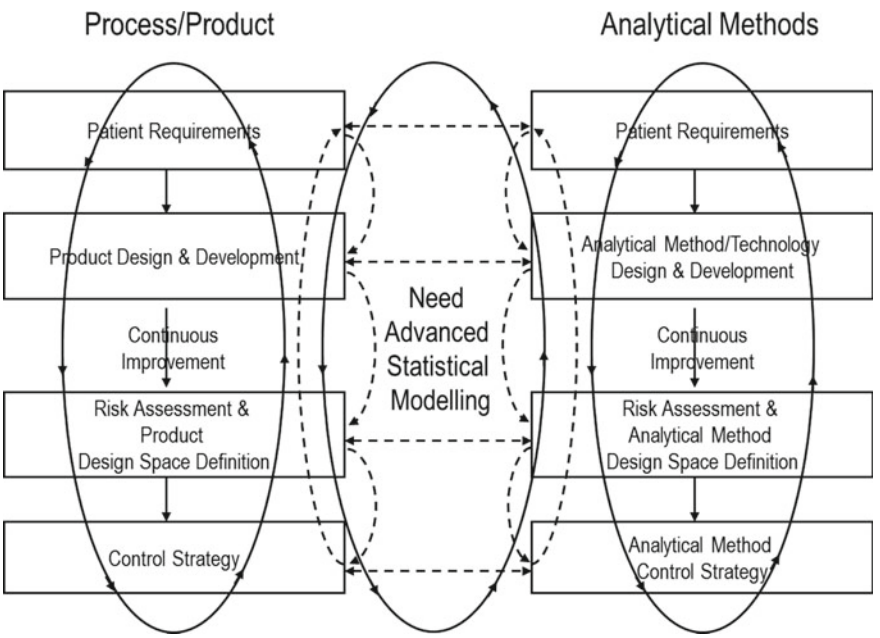


Fig. 2 Values of advanced statistical modeling in continuous analytical verification (adapted from PhRMA ATG presentation to FDA [2])

3 A Case Study

A case study is shown in this section to demonstrate the benefits of Bayesian modeling as a risk assessment tool for CAV. The case models the impact of a potential change in the system suitability criterion of the protein concentration method, on the process performance of both the drug substance and drug product for a biologics product. Analytical variability and variabilities from critical manufacturing steps are taken into consideration, and predictive inferences are made to decide whether the system suitability criterion change can be justified for the protein concentration method.

a. **Background**

The existing system suitability criterion for working reference standard (WRS) is to conduct three tests on the Reference Material (RM) and the average protein concentration must be within ±2.5% from the RM release value. The limits were originally determined based on limited method historical data. The system suitability criterion is applied to both drug substance (DS) (reported as protein concentration in unit of mg/mL) and the drug product (DP) (reported as drug content in unit of mg). The goal of the study is to evaluate if it is appropriate to widen the system suitability criterion from ±2.5% to ±3.0%. A critical consideration for decision making is how

the quality performance of DS and DP would be impacted by widening the system suitability criterion. More specifically, the analysis is to evaluate the impact on DS protein concentration process capability (Cpk at release) and DP drug content out-of-specification (OOS) risk.

b. **Building Risk-based Control Strategy**

(1) **Define**

To identify the key sources of variation in DS and DP outcomes, a process flowchart is presented in Fig. 3 with the major sources identified, the corresponding specifications or control limits, and the theoretical calculation per stage. The manufacturing and analytical testing procedures could impact the process and the final product quality through the following major stages,

- DS is formulated at a target protein concentration of 50 mg/mL (specification: 45–55 mg/mL). The sample is tested using an A280 method, with results containing two sources of variations: the DS manufacturing variability (Δ_{DS}) and the analytical variability from the DS testing lab (Δ_1).
- During DP manufacturing, multiple formulated DS batches can be mixed if needed, and then filled into vials. The control limit on fill weight is 8.923–9.381 mg/vial. The true protein concentration in the vial is the weighted average across the mixed DS batches. The vial filling procedure introduces a weight variability (Δ_{FW}).
- The filled DP vials are then lyophilized and stored at 2 °C–8 °C. To test the drug content in DP, the lyophilized DP is diluted with water to a final volume of 8.8 mL and tested using an A280 method. The DP sample contained variabilities from DS manufacturing (Δ_{DS}), DP fill weight (Δ_{FW}), as well as analytical variability from the DP testing lab (Δ_2).

(2) **Measure**

To evaluate the identified sources of variations and quantify the associated risks, the historical RM results from three labs (A, B and C) were provided, where lab A and B are the DP testing labs and lab C is the DS testing lab. A run chart is presented in Fig. 4, which reveals the performance differences across the three labs.

- Lab C reveals certain bias to the negative direction, comparing to the WRS release value;
- Lab A reveals certain bias to the positive direction;
- Lab A and C show larger variability compared to lab B.

As suggested by analytical scientists, RM testing results from Lab B and C are used for the risk assessment, as Lab B and C are the primary testing labs for DP and DS, respectively.

Paired data for measured protein concentration results with the corresponding RM testing results were provided for 34 DS batches. In addition, data for filled weight of DP vials was provided with average and standard deviation of vial fill weight from 10 DP batches.

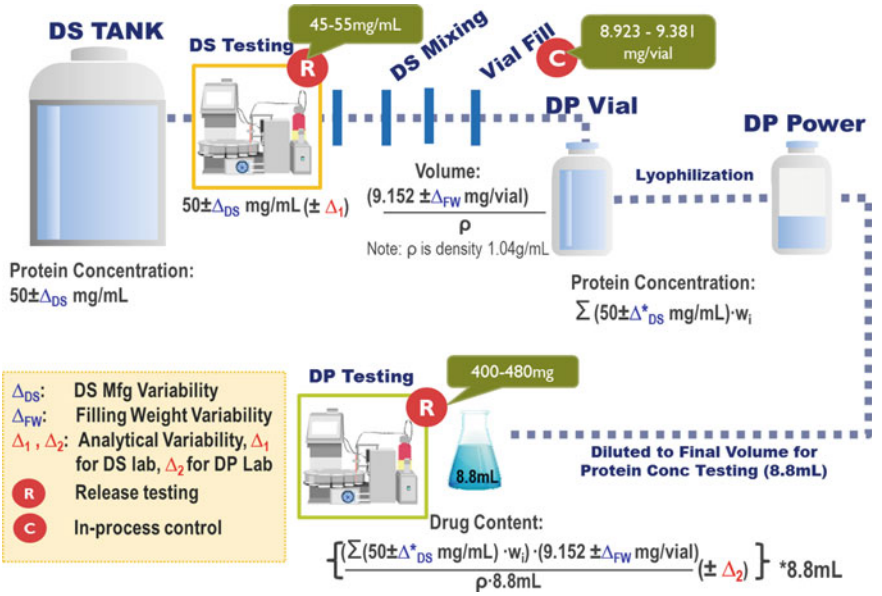


Fig. 3 Flow chart of product manufacturing and theoretical values of the key parameters

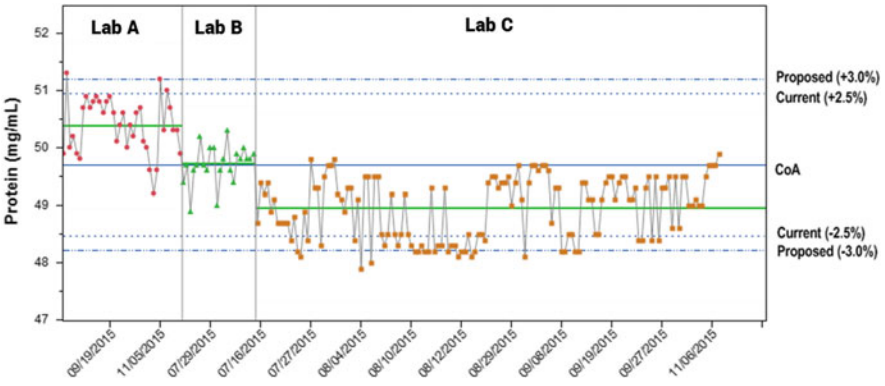


Fig. 4 Run chart of system suitability results by lab

(3) **Assess**

Widening of the system suitability criteria can potentially introduce more analytical errors into the measured results and thus, impact the process performance. Bayesian modeling techniques are applied to estimate such effects and the statistical models are presented as follows.

Model 1: Analytical Variability of DS Lab
Objective:

- Obtain predictive inference for true DS protein concentration (without analytical error).
- Obtain predictive inference for DS protein concentration (factoring in true DS protein concentration and analytical error).
- Predict the results of system suitability test and the corresponding analytical errors in the DS testing results.

Statistical Model:

$$
\begin{aligned}
PC_i^{DS.Obs} &\sim PC_i^{DS} + \Delta_1 \\
PC_i^{DS} &\sim N(\mu_{DS}, \sigma_{DS}) \\
\Delta_1, SS_1, SS_2, SS_3 &\sim N(\mu_i^{DS}, \sigma_i^{DS}) \\
SS_{report}^{DS} &\sim \sum_{k=1,2,3} SS_k/(3 \cdot CoA) \cdot I(L, U) \\
\mu_i^{DS} &\sim N(\mu_{M\mu}^{DS}, \tau_{M\mu}^{DS})
\end{aligned}
$$

where

PC_i^{DS} = true protein concentration for the ith DS lot (i= 1, 2, …, 34)
$PC_i^{DS.Obs}$ = measured protein concentration which contains analytical error
Δ_1 = analytical error in the measured protein concentration (DS lab)
SS_1, SS_2, SS_3 = analytical errors in a series of three individual system suitability tests
SS_{report}^{DS} = average of the three system suitability tests which were only reported if passing the existing system suitability criterion of ±2.5%
μ_{DS}, σ_{DS} = DS process mean and process standard deviation
μ_i^{DS}, σ_i^{DS} = population mean and standard deviation for analytical error under the same testing circumstance (repeatability)
$\mu_{M\mu}^{DS}$, $\tau_{M\mu}^{DS}$ = population mean and standard deviation for analytical error under varied testing circumstances (intermediate precision).
L, U = lower and upper acceptance criteria for the system suitability test.

Priors:

$$\mu_{DS} \sim N(50, 100) \qquad \mu_{DS} \sim U(0, 100)$$

$$\mu_{M\mu}^{DS} \sim N(0, 100) \qquad \tau_{M\mu}^{DS} \sim U(0, 10) \qquad \sigma_i^{DS} \sim U(0, 1)$$

Model 2: DP Fill Weight
Objective:

- Obtain predictive inference for filling weight of individual vials

Model:

$$\begin{aligned} FW_{ij} &\sim N(FW_i, \sigma_i) \\ FW_i &\sim N(\mu_g, \sigma_g) \\ \tau_i = \frac{1}{\sigma_i^2} &\sim \Gamma(\alpha, \beta) \end{aligned}$$

where

FW_{ij} = fill weight for the jth vial ($j = 1, 2, \ldots, 200$) from the ith lot ($i = 1, 2, \ldots, 10$)
FW_i = average fill weight for the ith lot ($i = 1, 2, \ldots, 10$)
μ_g, σ_g = process mean and between batch standard deviation for DP fill weight
σ_i = within batch standard deviation of filling weight
τ_i = precision of within batch variability of filling weight

Priors:

$$\mu_g \sim N(9.152, 100) \quad \sigma_g \sim U(0, 10)$$

$$\alpha \sim \Gamma(0.001, 0.001) \quad \beta \sim \Gamma(0.001, 0.001)$$

Model 3: Analytical Variability of DP Lab
Objective:

- Obtain predictive inference for analytical variability from the DP Lab. Specifically, predict the results of system suitability test at the DP lab and the corresponding analytical errors in the tested DP results.

Model:

$$\begin{aligned} \Delta_2, SS_1, SS_2, SS_3 &\sim N(\mu_k^{DP}, \sigma_k^{DP}) \\ SS_{report}^{DP} &\sim \sum_{k=1,2,3} SS_k/(3 \cdot CoA) \cdot I(L, U) \\ \mu_k^{DP} &\sim N(\mu_{M\mu}^{DP}, \sigma_{M\mu}^{DP}) \end{aligned}$$

where

SS_1, SS_2, SS_3 = analytical errors in a series of three individual system suitability tests in the DP lab
Δ_2 = analytical error in the measured protein concentration in the DP lab

SS_{report}^{DP} = average of the three system suitability tests which were only reported if passing the existing system suitability criterion of ±2.5%
$\mu_k^{DP}, \sigma_k^{DP}$ = population mean and standard deviation for analytical error under the same testing circumstance in the DP lab (repeatability)
$\mu_{M\mu}^{DP}, \sigma_{M\mu}^{DP}$ = population mean and standard deviation for analytical error under varied testing circumstance in the DP lab (intermediate precision)
L, U = lower and upper acceptance criteria for the system suitability test.

Priors:

$$\mu_{M\mu}^{DP} \sim N(0, 100) \quad \sigma_{M\mu}^{DP} \sim U(0, 5) \quad \sigma_k^{DP} \sim U(0, 1)$$

Model 4: DP Drug Content
Objective:

- Obtain predictive inference on true DP drug content, based on the predicted true DS protein concentration from Model 1, the predicted DP fill weight per vial from Model 2, and the analytical errors per lab (Δ_1 for the DS lab from Model 1 and Δ_2 for the DP lab from Model 3).

To simplify the model, it is assumed that there was no DS mixing during DP manufacturing and that DS and DP followed a one-to-one mapping relationship.
Model:

$$PC_{ij}^{DP*} = \frac{FW_{ij}^{*} * PC_i^{DS*}}{\rho}$$
$$PC_{ij}^{DP.Obs*} = PC_{ij}^{DP*} + \Delta_2^{*}$$
$$|SS_{report}^{DP*}| \leq \text{WRS Criterion}$$

where

PC_{ij}^{DP*} = estimated true protein concentration for the *j*th vial (j = 1, 2,…, 200) produced from the *i*th DS lot (i= 1, 2, …, N)
$PC_{ij}^{DP.Obs*}$ = estimated tested protein concentration for the *j*th vial produced from the *i*th DS lot
PC_i^{DS*} = predicted true protein concentration for the *i*th DS lot from Model 1
FW_{ij}^{*} = predicted fill weight for the *j*th vial from the *i*th lot from Model 2
SS_{report}^{DP*} = predicted results of the system suitability test (average of three individual tests)
Δ_2^{*} = predicted analytical error in the measured protein concentration in the DP lab corresponding to the reported system suitability test
r = DS density of 1.04 g/mL.

Prediction with Widened System Suitability Criteria:

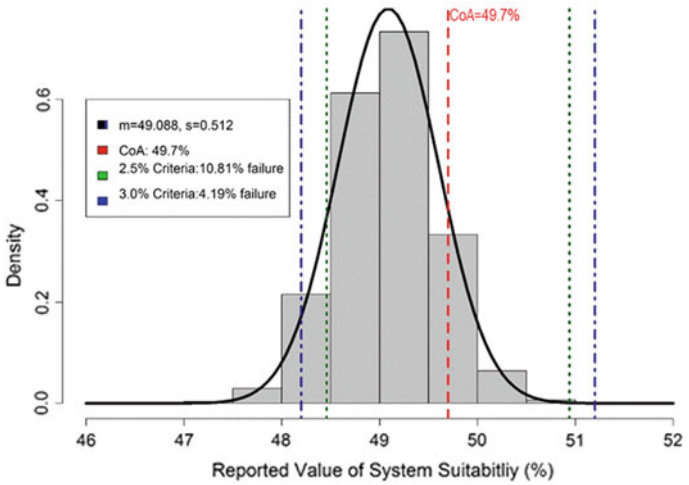


Fig. 5 Predicted distribution of reported system suitability results (drug substance lab)

Upon widening the system suitability criteria from ±2.5% to ±3.0%, larger values from system suitability tests would be accepted (SS^{DS*}_{report} for DS and SS^{DP*}_{report} for DP). The corresponding analytical error (Δ^*_1, Δ^*_2) being introduced to the DS and DP testing results are also expected to be larger.

The distributions of DS and DP results, as well as the DS Cpk and DP OOS, are calculated based on the obtained predictive inferences from Models 1 to 4 and by applying the widened system suitability criteria of ±3.0% to filter the simulated results. The results are compared against the results filtered with the current system suitability criteria of ±2.5%.

Results:

From Model 1, 10,000 reported system suitability results from the DS testing lab are generated and as shown in the histogram in Fig. 5. The system suitability results has a negative bias of −0.61 mg/mL, comparing to the RM release value of 49.7%. By widening the system suitability criterion from ±2.5% to ±3.0%, the failure rate would be reduced by 6.6% for the DS testing lab, from 10.8% to 4.2%.

From Model 3, 10,000 system suitability results (each result is an average of three individual tests) from the DP testing lab are generated and the histogram is presented in Fig. 6. The system suitability results center around the RM release value of 49.7%. The failure rates of system suitability test are estimated to be 0.33% if applying the existing criterion of ±2.5%, versus 0.08% if applying the proposed criterion of ±3.0%.

The true analytical error presented in the measured results is plotted against the analytical error observed in the system suitability testing results in Fig. 7. The analytical error observed in the system suitability testing results is defined as the relative difference between system suitability results and RM release value, while the RM release value is considered as the true value of RM. The plot shows at what level

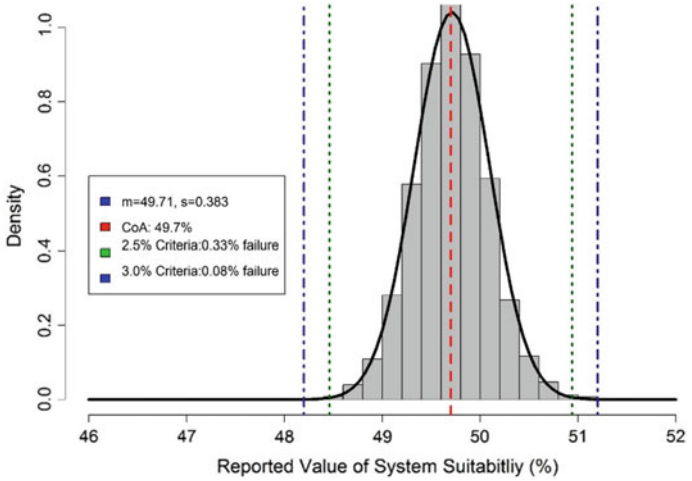


Fig. 6 Predicted distribution of reported system suitability results (drug product lab)

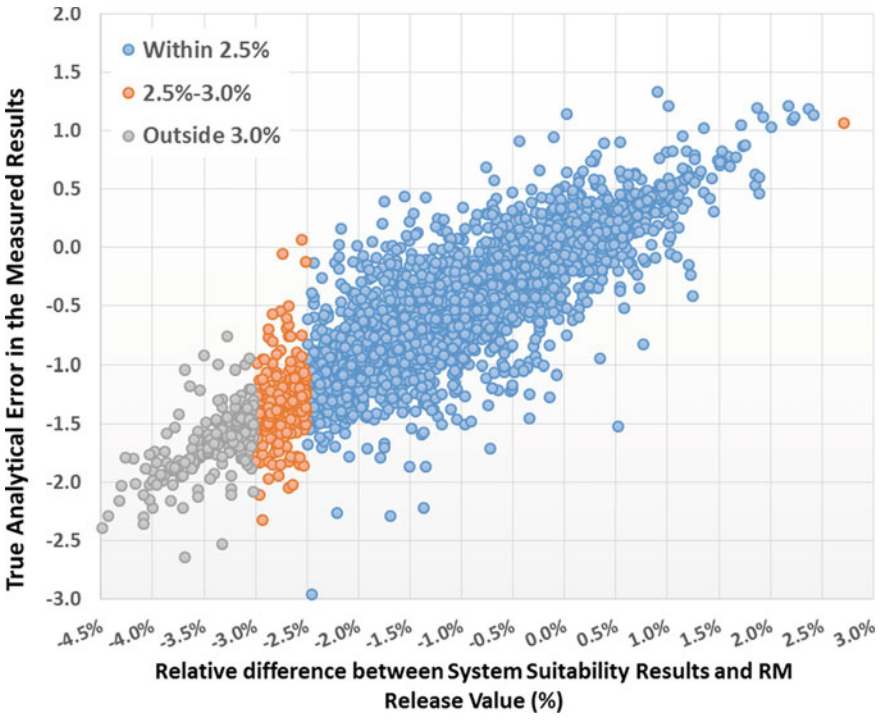


Fig. 7 Simulation results of analytical error in measured drug substance results versus in system suitability results

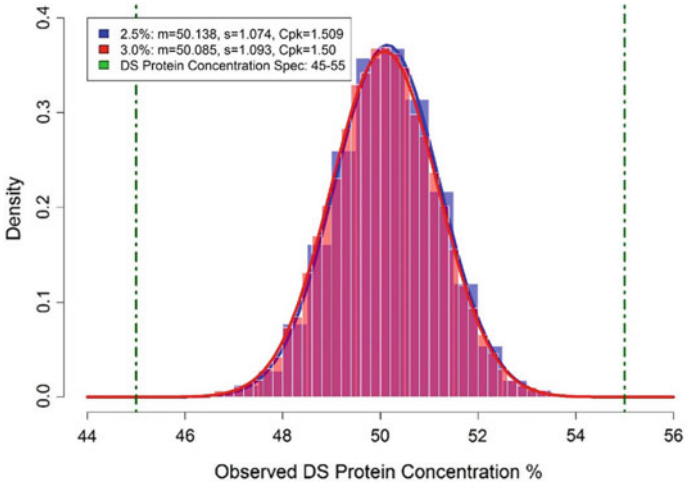


Fig. 8 Predicted distribution of measured drug substance protein concentration (with analytical error introduced by drug substance lab)

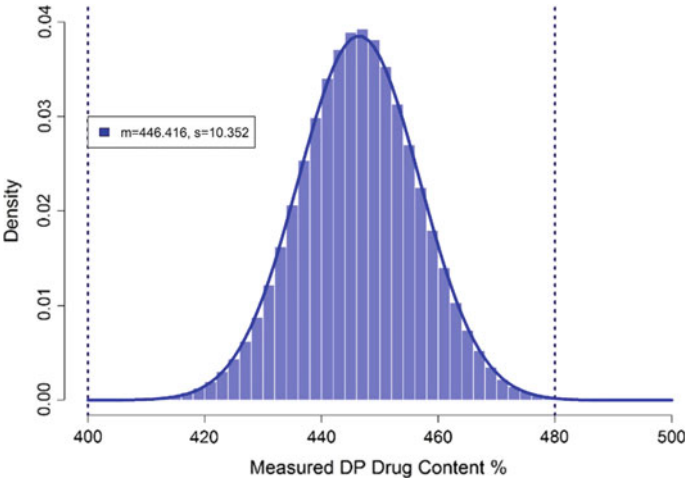


Fig. 9 Predicted distribution of measured drug product content (with analytical error introduced by drug product lab)

the system suitability testing can efficiently capture unacceptable analytical variation and reject the run before proceeding to the real samples testing. Widening the system suitability criteria from ±2.5% to ±3.0% will potentially reduce the system suitability failure rate, but introduce more negative analytical errors into the measured DS results. The combined effect is further evaluated, as shown in Fig. 8.

The histogram of predicted results for measured DS protein concentration is presented in Fig. 8, with different system suitability criteria applied. Upon widening the system suitability criteria, the center of the measured results is slightly shifted and the spread of the measured results become slightly larger. The DS Cpk against the specification of 45.0–55.0 mg/mL are almost the same between applying either system suitability criterion (Cpk = 1.51 vs. 1.50).

The predicted DP drug content tested by DP lab is shown in Fig. 9. The probabilities of OOS occurrence are estimated to be 0.11%, when applying either system suitability criterion.

(4) **Control decision**

The study examines the opportunity of widening the system suitability criterion from ±2.5% to ±3.0%, which would significantly reduce the failure rate during the system suitability testing.

Bayesian modeling techniques are applied to obtain predictive inference on both the true and measured DS/DP protein concentration, the vial fill weight, and the analytical variation per lab. Further predictions are performed to estimate the impacts on DS process capability and DP OOS risk, due to the widening of system suitability criterion. The risk assessment suggests that both effects are relatively small. Therefore, widening the system suitability criterion from ±2.5% to ±3.0% is of low risks.

Limitations of this analysis are noted in the decision making process, including: the available RM results were censored by the current system suitability criterion; not all RM results were distinctive measurements (i.e., multiple DS tests could have shared the same system suitability testing results if they were tested on the same day); the reported RM results were an average of three tests, while individual numbers were not available, etc.

(5) **Improve**

Opportunities for improvement are identified through the risk assessment and decision making discussions. The testing labs plan to pursue operational improvement in the system suitability test, which may change the distribution of analytical variability. Data collection practices can be improved to address the limitations discussed in Step 4.

Steps 1 to 5 can be repeated regularly or as needed through the product lifecycle to ensure the effectiveness of the analytical control strategy. Bayesian's benefit, in terms of continuous learning, can be explored through cycles of updating analyses.

4 Conclusion

In this work, we propose to apply advanced statistical tools in developing a risk-based control strategy to ensure CAV through a product lifecycle. A Bayesian based systematic approach is proposed to assess the impact of analytical performance on product robustness, integrate impacts of analytical and process components, inform decisions related to analytical control strategy, and ultimately, ensure analytical and process robustness.

A case study is presented to demonstrate the values of the proposed Bayesian-based risk assessment tool. The analysis models the impacts of a potential change of the system suitability criterion of the protein concentration method for a biologics product. Analytical variability and variabilities from critical manufacturing steps are considered in making predictive inferences related to analytical and process performances. The predicted results clearly show that relaxing the system suitability criterion would improve analytical performance without posing major risks to future process performance.

Once more data becomes available from future testing and improvements, the analytical control strategy can be revisited and refined, to which Bayesian updating can provide further benefits. The approach can be expanded to other tests and products, and enable more effective risk management to ensure CAV through a product lifecycle.

References

1. US Department of Health and Human Services, Food and Drug Administration: Guidance for industry. Process validation: General principles and practices (2011)
2. Borman, P., Nethercote, P., Chatfield, M., Thompson, D., Truman, K.: The application of quality by design to analytical methods. Pharm. Technol. **31**, 142–152 (2007)
3. Schweitzer, M., Pohl, M., Hanna-Brown, M., Nethercote, P., Borman, P., Hansen, G., Smith, K., Larew, J.: Implications and opportunities of applying QbD principles to analytical measurements. Pharm. Technol. **34**(2), 52–59 (2010)
4. Borman, P., Popkin, M., Oxby, N., Chatfield, M., Elder, D.: Analytical methods and control strategies: the forgotten interface. Pharm. Outsourcing **16**, 34–39 (2015)

Part IV
Clinical Trial Design and Analysis

Exact Inference for Adaptive Group Sequential Designs

Cyrus Mehta, Lingyun Liu, Pranab Ghosh and Ping Gao

Abstract In this paper we present a method for estimating the treatment effect in a two-arm adaptive group sequential clinical trial that permits sample size re-estimation, alterations to the number and spacing of the interim looks, and changes to the error spending function based on an unblinded look at the accruing data. The method produces a median unbiased point estimate and a confidence interval having exact coverage of the parameter of interest. The procedure is based on mapping the final test statistic obtained in the modified trial into a corresponding backward image in the original trial. Methods that were developed for classical (non-adaptive) group sequential inference can then be applied to the backward image.

Keywords Estimation in adaptive design · Exact adaptive confidence intervals · Adaptive median unbiased estimates · Group sequential estimation

1 Introduction

An adaptive group sequential trial permits data dependent alterations of the key design parameters such as sample size, number and spacing of interim looks, and the error spending function. The primary motivation for these adaptive modifications is the uncertainty regarding the efficacy of the new treatment relative to the control. Mehta and Pocock [7] and Mehta [8] present several case studies of actual trials in which provision was made for such adaptive modifications. The two major statistical problems for an adaptive group sequential trial are hypothesis testing and parameter estimation. Specifically, how can we prevent inflation of the type-1 error, and how

C. Mehta (✉) · L. Liu · P. Ghosh
Cytel Corporation, Cambridge, MA, USA
e-mail: mehta@cytel.com

C. Mehta
Harvard School of Public Health, Boston, MA, USA

P. Gao
Brightech-International, 285 Davidson Ave, Somerset, NJ 08873, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_10

can we obtain valid p-values, confidence intervals and point estimates in an adaptive group sequential trial?

Cui et al. [2], and Lehmacher and Wassmer [6] showed that the type-1 error of an adaptive group sequential trial can be preserved by combining the independent data from the different stages of the trial with pre-specified weights. A more general approach that permits, among other options, changes in the sample size, the number of interim looks, the spacing of interim looks, the error spending function and subgroup selection, was proposed by [10]. Their method is based on the conditional error rate principle. Specifically, one must ensure that the conditional type-1 error after an adaptive change does not exceed the conditional type-1 error of the original design. For the related problem of parameter estimation, [6] proposed extending [4] repeated confidence intervals method by applying it to the inverse-normal weighted statistic. Mehta et al. [9] also proposed an approach based on extending [4] repeated confidence intervals. Their solution, based on a generalization of the hypothesis testing procedure of [10], is applicable to a broader class of adaptive changes than the method of [6]. Repeated confidence intervals do not, however, exhaust the entire type-1 error and hence produce conservative coverage of the efficacy parameter. More recently, [1] proposed a one-sided lower confidence bound for the efficacy parameter, based on extending the stage wise adjusted confidence intervals of [11]. A different approach, applicable to two-sided confidence intervals, was proposed by [3]. Their method generalizes the stage wise adjusted confidence intervals developed by [11] for classical group sequential designs, and the hypothesis tests developed by [10] for adaptive group sequential designs, and combines these two ideas in a novel manner to map the observed value of the test statistic from the sample space of the adaptive design to the sample space of the original non-adaptive design. The usual stage wise adjusted confidence intervals are then derived for the mapped image of the test statistic. These confidence intervals are referred to as *backward image* confidence intervals (BWCI). This paper presents the highlights from the paper by [3] but refers the reader to the original paper for technical details. The main results of [3] are summarized in Sect. 2. Section 3 presents extensive simulation results that demonstrate median unbiasedness and exact coverage. We end with some concluding remarks in Sect. 4.

2 Backward Image Method

Consider a two-arm randomized clinical trial comparing a new treatment to an active control. The treatment effect is captured by a single parameter θ that might denote the difference of means for two normal distributions, the difference of proportions for two binomial distributions, the log hazard ratio for two survival distributions, or more generally, the coefficient of the treatment effect in a regression model. The accumulating data are captured by the efficient score statistic

$$W(t) = \hat{\theta} t$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ and

$$t = [\mathrm{se}(\hat{\theta})]^{-2}$$

is the Fisher information for θ obtained from the available data. Since t depends on unknown parameters it is replaced, in practice, by its large sample estimate. Furthermore, as is well known (e.g., [4]), $W(t)$ converges in distribution to a Brownian motion with drift θ. That is,

$$W(t) \xrightarrow{D} B(t) + \theta t \tag{1}$$

where $B(t) \sim N(0, t)$, and for any $t_2 > t_1$, $\mathrm{cov}\{B(t_1), B(t_2)\} = t_1$.

We shall be interested in testing the null hypothesis H_δ: $\theta = \delta$ for arbitrary values of δ and inverting this hypothesis test to produce point and interval estimates for θ. We will assume throughout that a positive value of θ indicates a better prognosis for the treatment arm relative to the control arm. In the absence of adaptive changes, the following group sequential trial will be employed to test H_0. Analyses are planned at information times $t_1^{(1)}, t_2^{(1)}, \ldots t_{K_1}^{(1)}$ with corresponding critical values $c_1^{(1)}, c_2^{(1)}, \ldots c_{K_1}^{(1)}$. The trial is terminated and null hypothesis H_0 is rejected at the first information time, $t_j^{(1)}$ say, such that $W(t_j^{(1)}) \geq c_j^{(1)}$. If $W(t_j^{(1)}) < c_j^{(1)}$ for all $j = 1, 2, \ldots K_1$, then H_0 is retained. For a one-sided level-α test of H_0, the critical values, $c_1^{(1)}, c_2^{(1)}, \ldots c_{K_1}^{(1)}$, must satisfy the relationship

$$P_0(\bigcup_{i=1}^{K_1} [W(t_i^{(1)}) \geq c_i^{(1)}]) = \alpha\,, \tag{2}$$

where $P_\delta(.)$ represents probability under the assumption that $\theta = \delta$.

At any look $L < K_1$, with $W(t_L^{(1)}) = x_L^{(1)}$, it is possible to alter the number and spacing of the future looks based on an examination of the data already obtained. Suppose it is decided to take K_2 future looks, at information times $t_1^{(2)}, t_2^{(2)}, \ldots t_{K_2}^{(2)}$. Let $c_1^{(2)}, c_2^{(2)}, \ldots c_{K_2}^{(2)}$ be corresponding critical values, so selected that

$$P_0\{\bigcup_{j=L+1}^{K_1} W(t_j^{(1)}) \geq c_j^{(1)} | W(t_L^{(1)}) = x_L^{(1)}\} = P_0\{\bigcup_{j=1}^{K_2} W(t_j^{(2)}) \geq c_j^{(2)} | W(t_L^{(1)}) = x_L^{(1)}\}\,. \tag{3}$$

We will continue to monitor the accumulating data and will reject H_0 at the first information time $t_I^{(2)} > t_L^{(1)}$ such that $W(t_I^{(2)}) \geq c_I^{(2)}$. If $W(t_i^{(2)}) < c_i^{(2)}$ for all $i = 1, 2, \ldots K_2$, then we will retain H_0 and set $t_I^{(2)} = t_{K_2}^{(2)}$. Müller and Schäfer [10] have shown that, despite this data driven modification of the trial, the unconditional probability that such a procedure will reject H_0 remains α. Equation (3) is referred to by [10] as the principle of preserving the conditional error rate.

Suppose the trial terminates at information time $t_I^{(2)}$ with observed statistic $x_I^{(2)}$. We now compute $(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$, the *backward image* of the observed outcome $(t_I^{(2)}, x_I^{(2)})$, such that

$$P_\delta\{\bigcup_{i=1}^{I-1}[W(t_i^{(2)}) \geq c_i^{(2)}] \cup [W(t_I^{(2)}) \geq x_I^{(2)}]|x_L^{(1)}\} \\ = P_\delta\{\bigcup_{i=L+1}^{J_\delta-1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J_\delta}^{(1)}) \geq x_{J_\delta}^{(1)}]|x_L^{(1)}\} . \tag{4}$$

We have shown in [3] that the backward image of any observed outcome in the adaptive trial is unique and can easily be computed.

Given a final outcome $(t_I^{(2)}, x_I^{(2)})$ in the adaptive trial, we compute $(\delta_{\alpha/2}, \delta_{1-\alpha/2})$, the $100 \times (1-\alpha)\%$ two sided confidence interval for θ, and $\delta_{0.5}$, the median unbiased point estimate for θ by noting, as proven in [3], that the one-sided p-value of the observed event for the test of H_δ can be computed from its backward image as

$$f_\delta(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)}) = P_\delta\{\bigcup_{i=1}^{J_\delta-1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J_\delta}^{(1)}) \geq x_{J_\delta}^{(1)}]\} . \tag{5}$$

The lower confidence bound, $\delta_{\alpha/2}$ and corresponding backward image $(t_{J_{\delta_{\alpha/2}}}^{(1)}, x_{J_{\delta_{\alpha/2}}}^{(1)})$ are computed such that

$$f_{J_{\delta_{\alpha/2}}}(t_{J_{\delta_{\alpha/2}}}^{(1)}, x_{J_{\delta_{\alpha/2}}}^{(1)}) = \alpha/2 . \tag{6}$$

The upper confidence bound $\delta_{1-\alpha/2}$ and corresponding backward image $(t_{J_{\delta_{1-\alpha/2}}}^{(1)}, x_{J_{\delta_{1-\alpha/2}}}^{(1)})$ are computed such that

$$f_{J_{\delta_{1-\alpha/2}}}(t_{J_{\delta_{1-\alpha/2}}}^{(1)}, x_{J_{\delta_{1-\alpha/2}}}^{(1)}) = 1 - \alpha/2 . \tag{7}$$

Finally, find the median unbiased point estimate $\delta_{0.5}$ and corresponding backward image $(t_{J_{\delta_{0.5}}}^{(1)}, x_{J_{\delta_{0.5}}}^{(1)})$ are computed such that

$$f_{J_{\delta_{0.5}}}(t_{J_{\delta_{0.5}}}^{(1)}, x_{J_{\delta_{0.5}}}^{(1)}) = 0.5 . \tag{8}$$

3 Simulation Experiments

We evaluated the operating characteristics of the backward image method for estimating θ by repeatedly simulating a number of adaptive group sequential designs. In this section we report the results of three such simulation experiments. Each ex-

periment involved simulating an adaptive group sequential design with five different values of θ. We simulated the adaptive group sequential trial 100,000 times with each value of θ, thereby producing 100,000 confidence intervals whose coverage of θ we then assessed. All the simulations utilized normally distributed data with mean θ and $\sigma = 1$ (assumed known).

First Simulation Experiment. In this simulation experiment the original trial is designed for up to four equally spaced looks with the Lan and DeMets [5] O'Brien-Fleming type error spending function (LD(OF) error spending function). The total sample size of 480 subjects provides slightly over 90% power to detect $\delta = 0.3$ with a one-sided level-0.025 group sequential test. At look 1, with 120 subjects enrolled, the conditional power under the estimated value of θ is evaluated and if it falls between 30 and 90%, the so called "promising zone" (see Mehta and Pocock, [2]), the sample size is increased by the amount necessary to boost the conditional power up to 90%, subject to a cap of 1000 subjects. The trial then proceeds with the new sample size, up to three additional equally spaced looks, and new stopping boundaries derived from the LD(OF) error spending function. The α error of the new stopping boundaries for the adaptive extension is derived from Eq. (3) so as to preserve the unconditional type-1 error of the trial despite the data dependent adaptation. This trial is simulated 100,000 times with a fixed value of θ. At the end of each simulation the point estimate of θ, $\delta_{0.5}$, and the corresponding 95% two-sided confidence interval, $(\delta_{0.025}, \delta_{0.975})$, are computed. If the trial crosses the stopping boundary at look 1, there is no adaptation and the classical stage wise adjusted point and interval estimates are obtained. If, however, there is a sample size adaptation at look 1, the point and interval estimates for θ are computed by the backward image method using Eqs. (6), (7) and (8). Simulation results for $\theta = -0.15, 0, 0.15, 0.3$ and 0.45 are presented in Table 1. Column 1 contains the true value of θ that was used in the simulations. Column 2 contains the median of the 100,000 $\delta_{0.5}$ estimates and demonstrates that $\delta_{0.5}$ is indeed a median unbiased point estimate for θ. Column 3 contains the proportion of the 100,000 confidence intervals that contain the true value of θ. These intervals demonstrate 95% coverage up to Monte Carlo accuracy. Columns 4 and 5 display the proportion of intervals that exclude the true value of θ from below and above respectively.

Second Simulation Experiment. In this simulation experiment the original trial is designed for up to three equally spaced looks with the LD(OF) error spending function. The total sample size of 390 subjects provides about 90% power to detect $\delta = 0.3$ with a one-sided level-0.05 group sequential test. If the trial does not cross an early stopping boundary at look 1 or look 2, then at look 2, with 240 subjects enrolled, the conditional power under the estimated value of θ is evaluated and if it falls in the promising zone, here specified to between 20 and 90%, the sample size is increased by the amount necessary to boost the conditional power up to 90%, subject to a cap of 780 subjects. The trial then proceeds with the new sample size for up to three additional equally spaced looks with new stopping boundaries derived from the Lan and DeMets [5] Pocock type error spending function (the LD(PK) error spending function). This trial was simulated 100,000 times with different values of θ. The median of the 100,000 point estimates for θ and the coverage proportion of

Table 1 Results from 100,000 simulations of a 4-look LD(OF) GSD with adaptation at look 1 to a 3-look LD(OF) GSD, demonstrating that the point estimate is median unbiased and the two-sided 95% confidence intervals provide exact coverage of the true value of θ up to Monte Carlo accuracy

True value of θ	Median of 100,000 point estimates	Proportion intervals containing θ	Proportion of intervals that exclude θ	
			From below	From above
−0.15	−0.14971	0.94893	0.02568	0.02539
0.0	0.000363	0.94976	0.02486	0.02538
0.15	0.149574	0.94939	0.02484	0.02577
0.3	0.30028	0.95111	0.02442	0.02447
0.45	0.44996	0.95017	0.02489	0.02494

Table 2 Results from 100,000 simulations of a 3-look LD(OF) GSD with adaptation at look 2 to a 3-look LD(PK) GSD demonstrating that the point estimate is median unbiased and the two-sided 90% confidence intervals provide exact coverage of the true value of θ up to Monte Carlo accuracy

True value of θ	Median of 100,000 point estimates	Proportion intervals containing θ	Proportion of intervals that exclude θ	
			From below	From above
−0.15	−0.14972	0.90007	0.05022	0.04971
0.0	0.00027	0.90073	0.04920	0.05007
0.15	0.14986	0.89866	0.04955	0.05179
0.3	0.2999	0.90087	0.04940	0.04973
0.45	0.44963	0.89929	0.05083	0.04988

the corresponding 90% confidence intervals for θ are reported in Table 2. It is seen that the point estimates are median unbiased and the confidence intervals have exact 90% coverage up to Monte Carlo accuracy.

Third Simulation Experiment—Comparison with [6]. An alternative two-sided confidence interval was proposed by [6] based on extending the repeated confidence intervals of [4]. It is well known that these repeated confidence intervals provide conservative coverage for classical group sequential designs because of the possibility that the trial might stop early and not exhaust all the available α. It would therefore be instructive to assess the extent to which these repeated confidence intervals are conservative in the adaptive setting. Accordingly we created a design with three equally spaced looks derived from the LD(OF) spending function and a planned adaptation at the end of look 1. The total sample size of 480 subjects has 90.44% power to detect $\theta = 0.3$ with a one sided test operating at significance level $\alpha = 0.025$. If the trial does not cross the early stopping boundary at look 1 then, with 160 subjects enrolled, the conditional power under the estimated value of θ is evaluated and if it falls in the promising zone, here specified to between 30 and 90.44%, the sample size is increased by the amount necessary to boost the conditional power up to 90%, subject to a cap of 960 subjects. The trial then proceeds with the new sample size for up to two additional equally spaced looks with new

Table 3 Comparison of the coverage 100,000 simulated 95% confidence intervals generated by the BWCI, SWCI and RCI methods. The underlying design is a 3-look LD(OF) GSD with adaptation at look 1 to a 2-look LD(OF) GSD.

True value of θ	Median of 100,000 Point Estimates			Actual Coverage of 95% CIs		
	BWCI	SWCI	RCI	BWCI	SWCI	RCI
−0.15	−0.15027	−0.149794	NA	0.95062	NA	0.95771
0.0	0.000118	−0.000421	NA	0.95014	NA	0.95213
0.15	0.150858	0.149064	NA	0.95016	NA	0.95017
0.3	0.300286	0.301016	NA	0.95062	NA	0.97597
0.45	0.449971	0.451704	NA	0.94936	NA	0.9875

stopping boundaries derived from the LD(OF) error spending function. This trial was simulated 100,000 times with different underlying values of θ. Table 3 compares the actual coverage of θ by 100,000 95% confidence intervals obtained by the backward image method (BWCI) and the repeated confidence intervals method (RCI) due to [6]. The median of the 100,000 point estimates generated by the BWCI method and by the stage wise adjusted confidence interval method (SWCI) due to [1] methods is also reported. No corresponding method for obtaining a point estimate from the RCI method was developed by [6] hence none is reported.

As expected the BWCI method produces median unbiased point estimates and 95% confidence intervals with exact coverage up to Monte Carlo accuracy. The SWCI method also produces median unbiased point estimates but does not provide two-sided confidence intervals. The RCI method does not provide valid point estimates and produces confidence intervals with increasingly conservative coverage as θ increases. The reason for the increase in conservatism is that as θ increases, the probability of stopping early, and hence of not exhausting the entire α increases.

It is also informative to examine the extent of the one sided coverage by the three methods. This is shown in Table 4. The BWCI interval excludes the true value for θ with 0.025 probability symmetrically from below and above, whereas the RCI method is both extremely asymmetric as well as extremely conservative. The SWCI method excludes the true value for θ with probability 0.025 from below but is not applicable for exclusion from above.

4 Concluding Remarks

We have presented a new method for computing confidence intervals and point estimates for an adaptive group sequential trial. The confidence intervals are shown to produce exact coverage and the point estimates are median unbiased. These results close an important gap that previously existed for inference on adaptive group sequential designs. Hypothesis tests that control the type-1 error have been available for over a decade [2, 6, 10]). The development of procedures to produce valid

Table 4 Comparing the BWCI, SWCI and RCI methods in terms of the probability that the lower and upper bounds, respectively, of a 95% confidence interval will exclude θ. The underlying design, a 3-look LD(OF) GSD with adaptation at look 1 to a 2-look LD(OF) GSD, is simulated 100,000 times

True value of θ	Probability of Low CL $> \theta$			Probability of Up CL $< \theta$		
	BWCI	SWCI	RCI	BWCI	SWCI	RCI
−0.15	0.02505	0.0256	0.01905	0.02529	NA	0.02324
0.0	0.02462	0.0251	0.02448	0.02524	NA	0.02339
0.15	0.02473	0.0256	0.02585	0.02511	NA	0.02238
0.3	0.02411	0.0253	0.00654	0.02527	NA	0.01749
0.45	0.02470	0.0259	0.00075	0.02594	NA	0.01050

confidence intervals and point estimates proved to be much more challenging. The first methods to guarantee two-sided coverage [6, 9] were shown to be conservative and did not produce valid point estimates. Subsequently [1] proposed a procedure that does produce exact coverage and valid point estimates. However, it only produces one-sided intervals. In contrast the two sided interval discussed here provides a bounded region within which it is possible to verify monotonicty with standard search procedures. This has enabled us to provide an operational proof that the intervals have exact coverage and the point estimates are median unbiased. Refer to [3] for details.

Finally, the entire development in this paper was expressed in terms of score statistics and so is applicable to all types of efficacy endpoints including normal, binomial and survival endpoints and model-based endpoints derived from contrasts of regression parameters and estimated by maximum likelihood methods.

References

1. Brannath, W., Mehta, C., Posch, M.: Exact confidence intervals following adaptive sequential tests. Biometrics **64**, 1–22 (2009)
2. Cui, L., Hung, M.J., Wang, S.-J.: Modification of sample size in group sequential clinical trial. Biometrics **55**, 853–857 (1999)
3. Gao, P., Liu, L., Mehta, C.R.: Exact inference for adaptive group sequential designs. Stat. Med. **32**, 3991–4005 (2013)
4. Jennison, C., Turnbull, B.W.: Interim analyses: the repeated confidence interval approach (with discussion). J. R. Stat. Soc. B **51**(3), 305–61 (1989)
5. Lan, K.K.G., DeMets, D.L.: Discrete sequential boundaries for clinical trials. Biometrika **70**, 659–663 (1983)
6. Lehmacher, W., Wassmer, G.: Adaptive sample size calculations in group sequential trials. Biometrics **55**, 1286–1290 (1999)
7. Mehta, C.R., Pocock, S.J.: Adaptive increase in sample size when interim results are promising: a practical guide with examples. Stat. Med. **30**(28), 3267–3284 (2011)

8. Mehta, C.R.: Sample size re-estimation for confirmatory clinical trials. In: Harrington, D. (ed.) Designs for Clinical Trials: Perspectives on Current Issues, Chapter 4. Springer, New York (2012)
9. Mehta, C.R., Bauer, P., Posch, M., Brannath, W.: Repeated confidence intervals for adaptive group sequential trials. Stat. Med. **26**(30), 5422–5433 (2007)
10. Müller, H.H., Schäfer, H.: Adaptive group sequential designs for clinical trials: combining the advantage of adaptive and of classical group sequential approaches. Biometrics **57**, 886–891 (2001)
11. Tsiatis, A.A., Rosner, G.L., Mehta, C.: Exact confidence intervals following a group sequential test. Biometrics **40**, 797–803 (1984)

A Novel Framework for Bayesian Response-Adaptive Randomization

Jian Zhu, Ina Jazić and Yi Liu

Abstract The development of response-adaptive randomization (RAR) has taken many different paths over the past few decades. Some RAR schemes optimize certain criteria, but may be complicated and often rely on asymptotic arguments, which may not be suitable in trials with small sample sizes. Some Bayesian RAR schemes are very intuitive and easy to implement, but may not always be tailored toward the study goals. To bridge the gap between these methods, we proposed a framework in which easy-to-implement Bayesian RAR schemes can be derived to target the study goals. We showed that the popular Bayesian RAR scheme that assigns more patients to better performing arms fits in the new framework given a specific intention. We also illustrated the new framework in the setting where multiple treatment arms are compared to a concurrent control arm. Through simulation, we demonstrated that the RAR schemes developed under the new framework outperform a popular method in achieving the pre-specified study goals.

Keywords Response-adaptive randomization · Bayesian adaptive design · Goal function · Multi-arm comparative trials

1 Introduction

Clinical trials face the challenges of low success rates, rising costs, and prolonged timelines. In an effort to streamline clinical trials and expand the range of questions that may be explored in a single study, investigators have developed adaptive designs

J. Zhu (✉)
Global Statistics, Takeda Pharmaceutical Company Limited, Tokyo, Japan
e-mail: jian.zhu2@takeda.com

I. Jazić
Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, Cambridge, MA, USA

Y. Liu
Takeda Pharmaceutical, Cambridge, MA, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_11

that allow a trial in progress to be modified in a pre-specified fashion based on the accumulated data. Such designs may shorten trials, reduce costs, improve power, increase the number of patients treated effectively, and allow for the identification of clinically meaningful subgroups [3, 35]. Many types of adaptive designs have become part of the standard toolbox for conducting clinical trials [10].

Traditionally, fixed randomized designs have been the gold standard for clinical trials: they balance potential confounders and eliminate bias arising from treatment assignment by physicians. Equal allocation is justified under the principle of equipoise, in which there is no *a priori* belief that one treatment is superior. However, if interim results demonstrate that one treatment is superior to another, an adjustment of the allocation ratio may be warranted. This is one argument for the use of response-adaptive randomization (RAR), an adaptive strategy that modifies the allocation ratio in a trial based on interim response data to achieve favorable trial characteristics.

There are many RAR schemes in the literature, differing from each other in terms of design, implementation, and most importantly, goals. Such goals may include—but are not limited to—maximizing the power to identify the most efficacious treatment, maximizing the power to identify one treatment that is efficacious, minimizing the number of non-responders in the trial, and minimizing sample size/cost while maintaining sufficient power. However, these favorable characteristics are often in conflict, and a single RAR scheme cannot achieve them all. Therefore, investigators must identify and prioritize study objectives in order to select an RAR scheme that is aligned with these objectives.

As described further in the next section, frequently used simple RAR methods are easy to implement, but cannot be tailored to address particular study goals. On the other hand, more complex RAR methods can be tailored for particular study goals, but are harder to implement. In this manuscript, we propose a unifying framework for a class of simple Bayesian RAR methods that can be targeted to a broader range of study goals, yet still easy to implement. We illustrate the potential of this framework by evaluating two novel RAR methods for multi-arm trials with a concurrent control that were developed with specific study goals in mind.

Overview of RAR designs

RAR in practice can be dated back to the play-the-winner rule in two-arm trials [33, 36]. Since then, a variety of RAR schemes have been developed in both the frequentist and Bayesian paradigms. Frequentist RAR designs specifically developed to optimize certain criteria have been extensively studied, and their theoretical properties have been well established [12–14, 24, 25, 30, 37]. Rosenberger et al. [24] describe an RAR scheme based on Neyman allocation that maximizes the statistical power to test the difference between treatments, using an allocation ratio proportional to the empirical standard deviation calculated from the accruing data. Another design aims to minimize the expected number of failures in the trial while fixing the variance of the test statistic, thereby maintaining desirable power [14, 24]. The doubly adaptive biased coin design updates allocation ratios depending on both the observed allocation ratio and the estimated target allocation ratio to converge to a pre-specified optimal allocation function [8, 13]. However, the theoretical properties of such RAR

schemes generally rely on asymptotic approximation and have not been evaluated for small sample sizes, and the ratios themselves may not be intuitive.

Another set of RAR schemes in the literature maximize specific utility functions under a Bayesian decision-theoretic framework [2, 5, 16, 31]. For example, [2] consider the conditional expected successes lost, a loss function that incorporates the total number of patients anticipated to receive the drug, both within and beyond the trial (known as the patient horizon). Such designs extend two- and multi-armed bandit problems, in which resources must be optimally allocated among options providing random rewards according to their own probability distributions. However, this class of methods has some drawbacks—some methods require extensive recursive computations, some are deterministic and are vulnerable to associated bias, while others require that the response for each patient must be observed before the next can be randomized, which may pose challenges in practice.

The most frequently used RAR scheme in medical research and the pharmaceutical industry stems from [29] paper on calculating the probability that one underlying response rate exceeds another, given two samples. In the version that is currently used in practice, updated allocation probabilities are computed at each interim analysis based on the posterior probability of each arm being the most efficacious. In the absence of a name for this method in the literature, we refer to it as Thompson RAR. Beginning in the late 1990s, Donald Berry and his colleagues at the MD Anderson Cancer Center designed many Bayesian adaptive clinical trials in oncology, employing Thompson RAR as one of a host of adaptive strategies. For example, the I-SPY 2 breast cancer trial [1] and the BATTLE lung cancer trial [15] are two well known large-scale phase II clinical trials that adopted this method for RAR. Thompson RAR has also been used in trial designs in a variety of other therapeutic areas, such as cardiovascular disease [6, 17], gastrointestinal disorders [23], diabetes [9, 26], and neurology [4, 18].

Thompson RAR is appealing from a practical standpoint—it is intuitive, simple to implement, can accommodate block randomization, and does not rely on asymptotics. Unlike the optimal frequentist and Bayesian methods described above, though, the theoretical properties of Thompson RAR have not been extensively studied, and it has not been explicitly linked with a specific intention or goal. The fact that Thompson RAR is the preeminent RAR method used in practice carries with it the risk of a "one size fits all" approach in applying it across studies with possibly disparate study goals. In the next section, we describe our proposed RAR framework that enables Thompson RAR to be linked with a specific intention and, in effect, generalizes it to a larger class of simple Bayesian RAR methods that can be customized to a particular goal.

2 Proposed Framework

Consider a clinical trial with a total of K arms. For $k = 1, \ldots, K$, let $\theta_k \in \Omega_k$ denote the response parameter vector for the k_{th} arm, where Ω_k is the parameter space for θ_k

and $\Omega = \Omega_1 \times \cdots \times \Omega_K$. For example, θ_k may represent the response rate if the endpoint is binary, or the mean and variance if the endpoint is normal. Traditionally, fixed equal randomization designs use the allocation ratio $(1/K, \ldots, 1/K)$. Some fixed randomization designs determine the optimal allocation ratio $w = (w_1, \ldots, w_K)$, where $\sum_{k=1}^{K} w_k = 1$, based on assumptions of some constant values for the response parameters $\theta = \{\theta_1, \ldots, \theta_K\}$ prior to the study. This is desirable since w is specifically derived based on the study goal. For instance, in a comparative trial with a treatment and a control arm, one can derive an optimal fixed allocation to maximize the statistical power to detect the treatment difference using a Wald test, which is the Neyman allocation assigning patients to each arm with weight proportional to its standard deviation. However, such design is usually sensitive to the pre-specification of θ. In other words, for each $\theta \in \Omega$, the desirable randomization weight w is different and can be viewed as a function of θ. If there is no additional information regarding θ, it is impractical if not impossible to implement the weight $w(\theta)$.

In the context of RAR, suppose there are I interim analyses, and let $\mathbf{Y}_i = \{\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{iK}\}$ denote the observed response data for all arms at the i_{th} interim analysis, $i = 1, \ldots, I$. The accumulated data can be used for two purposes. Firstly, since $w(\theta)$ is derived before any data is collected, we can update w based on the study goal for any θ given $\mathbf{Y}_i$, which is denoted as $w(\theta \mid \mathbf{Y}_i)$. In other words, $w(\theta \mid \mathbf{Y}_i)$ can be regarded as an optimal weight function in the presence of historical data $\mathbf{Y}_i$. Secondly, in the Bayesian framework, instead of pre-specifying the exact values for θ, one can assume a prior distribution $p(\theta)$, often non-informative, and summarize the information regarding θ into a joint posterior distribution $p(\theta \mid \mathbf{Y}_i)$.

The above two components give a complete picture about the parameter space Ω: given observed data, $p(\theta \mid \mathbf{Y}_i)$ tells us where the parameters are likely to be, and $w(\theta \mid \mathbf{Y}_i)$ tells us for such θ what the allocation ratio should be. Combining the two components, we propose a novel framework to calculate the adaptive randomization ratio as below:

At the i_{th} interim analysis with observed data $\mathbf{Y}_i$, let $G(\theta, w \mid \mathbf{Y}_i)$ denote the metric for the study goal. Then for any θ and data, we define the weight function:

$$w^*(\theta \mid \mathbf{Y}_i) = argmax_w G(\theta, w \mid \mathbf{Y}_i),$$

subject to the constraint that $\forall\, \theta$, $\sum_k w_k^*(\theta \mid \mathbf{Y}_i) = 1$. The allocation ratio for the k_{th} arm is calculated as:

$$\mathbf{w}_{ik} = E[w_k^*(\theta \mid \mathbf{Y}_i) \mid \mathbf{Y}_i] = \int w_k^*(\theta \mid \mathbf{Y}_i)\, p(\theta \mid \mathbf{Y}_i) \mathrm{d}\theta.$$

If the optimal weight function is difficult to obtain, one may choose a weight function that aligns with the goal.

Although the analytic form of the allocation ratio can be derived for some cases, in practice, it is easier to obtain such weight through Monte Carlo integration following three steps:

1. Generate D draws $\{\theta^{(1)}, \ldots, \theta^{(d)}, \ldots, \theta^{(D)}\}$ from the joint posterior distribution $p(\theta \mid \mathbf{Y}_i)$;
2. For each drawn parameter vector $\theta^{(d)}$, derive the weight $w^{(d)} = w^*(\theta^{(d)} \mid \mathbf{Y}_i)$;
3. Define $\mathbf{w}_i = \sum_{d=1}^{D} w^{(d)}/D$.

2.1 Revisiting Thompson RAR

In this subsection we will demonstrate that Thompson RAR fits neatly in the proposed framework.

We consider applying RAR in a common type of multi-arm comparative trials without a concurrent control arm, where the intention of the RAR scheme is to maximize the average number of patients assigned to the most efficacious treatment arm. Following the notation described earlier, we assume that the trial includes a total of K candidate treatments. For the rest of the paper, we focus on a measure of treatment efficacy. To differentiate efficacy from the general response parameter θ, let μ_k denote the efficacy parameter for the k_{th} arm, where $\mu = \{\mu_1, \ldots, \mu_K\}$. μ can be identical with θ (response rate for binary endpoint), can be a part of θ (mean for continuous endpoint), or can be a transformation of θ (effect size for continuous endpoint). Without loss of generality, we assume that a larger value of μ_k corresponds to a better efficacy.

Given observed data $\mathbf{Y}_i$ at the interim analysis, let n_{ik} be the sample size for $\mathbf{Y}_{ik}$ and n_{next} be the total sample size for the next cohort of patients to be randomized. For any allocation weight w, the average total sample size for the k_{th} arm at the end of the next cohort is $N_k = n_{ik} + n_{\text{next}} w_k$. The average number of patients assigned to the most efficacious treatment can be defined as

$$G(\mu, w \mid \mathbf{Y}_i) = \sum_{k=1}^{K} 1_{\{\mu_k = \mu_{(K)}\}} \times N_k = n_{i(K)} + \sum_{k=1}^{K} 1_{\{\mu_k = \mu_{(K)}\}} \times n_{\text{next}} w_k,$$

where $\mu_{(K)} = \max_{1 \le j \le K} \mu_j$, and $n_{i(K)}$ is the corresponding sample size for that arm.

The allocation weight vector w can be determined so that the average sample size defined above is maximized. Note that $n_{i(K)}$ is fixed given the observed data, and $G(\mu, w \mid \mathbf{Y}_i) \le n_{i(K)} + n_{\text{next}}$. Without loss of generality, we can assume μ to be a continuous efficacy measure, therefore it is almost surely that only one arm has $\mu_{(K)}$. It is then easy to show that $G(\mu, w \mid \mathbf{Y}_i) = n_{i(K)} + n_{\text{next}}$ if and only if $w_k^* = 1_{\{\mu_k = \mu_{(K)}\}}$ $a.s.$ An intuitive interpretation suggests that we should assign all patients in the next cohort to the most efficacious arm and none to the other arms if we know the true efficacy profile.

Combining both, we can calculate

$$\mathbf{w}_{ik} = E[w_k^*(\mu) \mid \mathbf{Y}_i] = \Pr(\mu_k = \mu_{(K)} \mid \mathbf{Y}_i).$$

This is exactly the Thompson RAR allocation ratio. This allocation ratio can also be obtained through Monte Carlo integration. One can generate D draws $\{\mu^{(1)}, \ldots, \mu^{(d)}, \ldots, \mu^{(D)}\}$ from the joint posterior distribution $p(\mu \mid \mathbf{Y}_i)$. For the d_{th} drawn $\mu^{(d)} = (\mu_1^{(d)}, \mu_2^{(d)}, \ldots, \mu_K^{(d)})$, let B_d denote the index of the most efficacious treatment arm. Then the weight function $w^{(d)}$ has weight 1 for the B_d-th arm and 0 for all other arms. The average of $w^{(d)}$ is then a numerical approximation of the Thompson RAR allocation ratio.

By applying the proposed framework on this particular example, we provide an intention for Thompson RAR as a by-product. In the next section we will demonstrate that this framework can be applied to a much broader range of studies.

3 Application in Multi-arm Trials with a Concurrent Control

We consider another common type of multi-arm trials in which multiple candidate treatment arms are compared to a concurrent control. For specific study goals, we propose RAR schemes under the new framework.

3.1 Background

Thompson RAR is widely used in multi-arm comparative trials without a concurrent control. However, when the trial has a control arm that is generally worse than the treatment arms, naive implementation of Thompson RAR often assigns too few patients to the control and thus reduces power to detect differences between the treatment arms and the control arm.

A popular approach to protect the allocation to the control, while still adopting Thompson RAR, is to assign a pre-specified fixed allocation ratio to the control arm, and distribute the rest of the patients to the treatment arms according to the posterior probability of each treatment arm being the most efficacious. We refer to this approach as Fixed Control (FC) RAR. Our literature review [7, 11, 18, 19, 22, 23, 26] suggests that, among all published multi-arm trials with a control that use Thompson RAR, almost all adopted FC RAR.

However, there are two major issues with FC RAR. Firstly, while this constant control allocation ratio throughout the trial prevents under-allocation, there is no clear guidance on what value this ratio should take, as different published trials have used different control allocation ratios ranging from 20% to 35%. In general, the appropriate fixed control ratio depends on the number of treatment arms in the study as well as the true underlying efficacy profiles of the treatment and control arms, which is typically unknown. For a particular assumed efficacy profile, one may pick a desirable fixed ratio through simulation, but it may no longer be desirable

if the efficacy profile changes. Secondly, in most trials using Thompson RAR, the study goal is rarely considered before FC RAR is implemented or evaluated. FC is often applied in studies with different goals. Some investigators may be interested in identifying the most efficacious treatment arm, while others are interested in finding any efficacious treatment arm. For example, suppose a trial has a true efficacy profile $(\pi_0, \pi_1, \pi_2, \pi_3) = (0.2, 0.2, 0.3, 0.35)$, where π_0 is the control response rate and π_1, π_2, π_3 are the treatment response rates, and any treatment arm with a response rate higher than 0.2 is regarded as efficacious. Suppose treatment arm 2 is selected at the end of the study ($\pi_2 = 0.3$). If the study goal is to select the most efficacious treatment, then selecting arm 2 does not reach the goal; if the study goal is to select any efficacious treatment, then selecting arm 2 is a correct decision that meets the goal.

To address these issues, we developed new RAR schemes through our proposed framework, which will be described in the next subsection.

3.2 New RAR Through Proposed Framework

Notation adjustment: To accommodate the control arm, we adjust our notation slightly by shifting the arm index from $1, \ldots, K$ to $0, 1, \ldots, K-1$, where 0 refers to the control arm, and $k = 1, \ldots, K-1$ refers to the k_{th} treatment arm.

We assume that the study tests whether μ_k is more efficacious than μ_0 by comparing the frequentist test statistic $\frac{\hat{\mu}_k - \hat{\mu}_0}{SE(\hat{\mu}_k - \hat{\mu}_0)}$ with a critical value. The critical value will be determined later to ensure that the type I error is controlled. Note that the selection of this test is for demonstration purpose only. In actual trials one may choose to use a different type of tests.

This subsection will focus on introducing the new schemes to derive the adaptive allocation ratio at interim analyses.

3.2.1 RAR Aiming to Pick the Most Efficacious Treatment

Let's consider a study goal, which is to maximize the power to detect the most efficacious treatment if at least one treatment arm is better than the control arm. We propose a RAR scheme targeting this goal, which borrows the Neyman allocation ratio [14] to determine the weight function. For any value of $\mu = \{\mu_0, \mu_1, \ldots, \mu_{K-1}\}$, in order to maximize the power to detect the difference between the most efficacious treatment arm ($K-1$) and the control arm, all patients from the next cohort should be assigned to these two arms, with allocation weight proportional to the standard deviation in each arm. Since this heavily depends on the **s**tandard **d**eviation in each arm, we name this scheme as the SD RAR.

Given observed data $\mathbf{Y}_i$ at the interim analysis, the SD RAR scheme is executed in three steps:

1. Generate D draws $\{\theta^{(1)}, \ldots, \theta^{(d)}, \ldots, \theta^{(D)}\}$ from the joint posterior distribution $p(\theta \mid \mathbf{Y}_i)$; Within each draw, derive the corresponding efficacy parameter draw $\mu^{(d)}$ for all arms if it is not identical with the $\theta^{(d)}$.

2. For each drawn efficacy parameter vector $\mu^{(d)} = \{\mu_0^{(d)}, \mu_1^{(d)}, \ldots, \mu_{K-1}^{(d)}\}$, derive the standard deviation vector $\sigma^{(d)}$ if it is not included in $\theta^{(d)}$ (e.g. $\sigma_k^{(d)} = \sqrt{\mu^{(d)}(1-\mu^{(d)})}$ if $\mu^{(d)}$ denotes the response rate for a binary endpoint). Let B_d denote the index of the most efficacious treatment arm in the d_{th} draw, i.e., $\mu_{B_d}^{(d)} = \max_{k>0} \mu_k^{(d)}$. Calculate

$$w^{(d)} = \left(\frac{\sigma_0^{(d)}}{\sigma_0^{(d)} + \sigma_{B_d}^{(d)}}, 0, \ldots, 0, \frac{\sigma_{B_d}^{(d)}}{\sigma_0^{(d)} + \sigma_{B_d}^{(d)}}, 0, \ldots, 0 \right),$$

where $w^{(d)}$ only has non-zero elements when $k = 0$ or $k = B_d$.

3. Calculate $\mathbf{w}_i = \sum_{d=1}^{D} w^{(d)}/D$.

3.2.2 RAR Aiming to Pick One Efficacious Treatment

As shown in the example in Sect. 3.1, the treatment with the best estimated efficacy may not necessarily be the most efficacious. However, for some studies this is acceptable as long as the selected treatment arm is truly efficacious. We then introduce a RAR developed with such study goal being considered.

We define a treatment arm k to be promising if $\Pr(\mu_k > \mu_0 \mid \mathbf{Y}_i) > P$, where P is a pre-specified promising threshold and is generally lower than the thresholds for early stopping for efficacy. For any value of $\mu = \{\mu_0, \mu_1, \ldots, \mu_{K-1}\}$, in order to increase the power to detect one treatment arm that is efficacious, we assign all patients from the next cohort equally to the arms with $\mu_k > \mu_0$ if there is no strong evidence suggesting that one arm among those efficacious arms is promising. On the other hand, if the accumulated data suggest that one arm is the most promising arm to be declared as efficacious, even though it may not be the best, we assign all patients from the next cohort to only this arm and the control arm. To further increase the power to select this arm, the allocation weights to this arm and the control arm are proportional to the standard deviations respectively. This is particularly desirable for the case where there are several treatment arms with similar efficacy. We name this method as PT RAR since the **p**romising **t**hreshold is a key component.

Given observed data $\mathbf{Y}_i$ at the interim analysis, the PT RAR scheme is executed in three steps:

1. Generate D draws $\{\theta^{(1)}, \ldots, \theta^{(d)}, \ldots, \theta^{(D)}\}$ from the joint posterior distribution $p(\theta \mid \mathbf{Y}_i)$; Within each draw, derive the corresponding efficacy parameter draw $\mu^{(d)}$ for all arms if it is not identical with the $\theta^{(d)}$.

2. For each treatment arm $k > 0$, calculate $\Pr(\mu_{\mathrm{k}} > \mu_0 \mid \mathbf{Y}_{\mathrm{i}})$, and let $(K-1)$ denote the index of the arm with the highest posterior probability of being better than the control, i.e., $\Pr(\mu_{(\mathrm{K}-1)} > \mu_0 \mid \mathbf{Y}_{\mathrm{i}}) = \max_{k>0} \Pr(\mu_{\mathrm{k}} > \mu_0 \mid \mathbf{Y}_{\mathrm{i}})$.

If $\Pr(\mu_{(\mathrm{K}-1)} > \mu_0 \mid \mathbf{Y}_{\mathrm{i}}) > P$, where P is a pre-specified threshold, for each drawn efficacy parameter vector $\mu^{(d)} = \{\mu_0^{(d)}, \mu_1^{(d)}, \ldots, \mu_{K-1}^{(d)}\}$, derive the standard deviation vector $\sigma^{(d)}$ if it is not included in $\theta^{(d)}$. Then calculate

$$w^{(d)} = \left(\frac{\sigma_0^{(d)}}{\sigma_0^{(d)} + \sigma_{(K-1)}^{(d)}}, 0, \ldots, 0, \frac{\sigma_{(K-1)}^{(d)}}{\sigma_0^{(d)} + \sigma_{(K-1)}^{(d)}}, 0, \ldots, 0 \right).$$

Note: If one treatment arm is classified as (the most) promising, then for all draws, $w^{(d)}$ only has non-zero elements for the control arm and the most promising treatment arm $(K-1)$.

This is not equivalent to early dropping of all other treatment arms. If accumulated data in later cohorts suggest this arm is not the most promising arm, other treatment arms will be considered again for randomization.

If $\Pr(\mu_{(\mathrm{K}-1)} > \mu_0 \mid \mathbf{Y}_{\mathrm{i}}) \leq P$, for each drawn efficacy parameter vector $\mu^{(d)}$, let $S_d = \{k | k > 0, \mu_k^{(d)} > \mu_0^{(d)}\}$ and L_d be the length of S_d. $w^{(d)}$ is calculated as $w_k^{(d)} = 1/(L_d + 1)$ if $k = 0$ or $k \in S_d$, and $w_k^{(d)} = 0$ otherwise.

Note: If none of the treatment arms is promising, for each draw, $w^{(d)}$ assigns equal weights to the treatment arms with better efficacy parameters than the control arm as well as the control arm. The set of treatment arms with better efficacy parameters than the control arm is different across draws.

3. Calculate $\mathbf{w}_i = \sum_{d=1}^{D} w^{(d)}/D$.

4 Simulation Study

In the setting of multi-arm trials with a concurrent control, we performed a simulation study comparing existing methods to the methods we have developed under our new framework, and evaluated their relative performance on the relevant study goals. Specifically, we compared equal randomization (ER) and fixed control (FC) to (1) the standard deviation (SD) method based on the probability of selecting the best arm and (2) the promising threshold (PT) method based on the probability of selecting one efficacious arm. We considered six sets of true response rates (in addition to the null scenario where all response rates equal 0.3):

- One efficacious arm ("One eff"): (0.3, 0.3, 0.3, 0.5)
- Two efficacious arms, similar ("Two eff, S"): (0.3, 0.3, 0.45, 0.46)
- Two efficacious arms, different ("Two eff, D"): (0.3, 0.3, 0.45, 0.5)
- Three efficacious arms, similar ("Three eff, S"): (0.3, 0.45, 0.45, 0.46)
- Three efficacious arms, different ("Three eff, D"): (0.3, 0.4, 0.45, 0.46)
- Staircase ("Staircase"): (0.3, 0.36, 0.42, 0.48).

For each set of response rates and each randomization method (ER, FC, SD, PT), 10,000 trials were simulated using the procedure described in the next section. For simulations under FC, we used a fixed 30% allocation to the control arm, and for simulations under PT, we used a promising threshold $P = 0.75$.

Simulation procedure

All simulated trials described here used a binary endpoint and involved three treatment arms and a concurrent control arm. Each trial began with an equal randomization burn-in period, in which 10 patients were assigned to each arm. This allowed some information on all arms to accumulate before adaptive randomization was initiated. Outcomes were simulated for these patients according to the specified true response rates for each arm, and an allocation vector $\mathbf{w}_i$ for the next cohort of patients was computed according to the specified randomization scheme. Note that under equal randomization, $\mathbf{w}_i = (0.25, 0.25, 0.25, 0.25)$ is fixed for the entirety of the trial. If the specified randomization scheme fell under the new framework, then Monte Carlo integration via sampling from the joint posterior distribution of response rates may have been necessary to calculate $\mathbf{w}_i$. For all simulations, a prior distribution of $Beta(0.2, 0.8)$ was used for each response rate, as in [32]. Then, a cohort of 20 patients was assigned to the four arms through block randomization based on $\mathbf{w}_i$. The process of simulating outcomes for newly assigned patients, calculating an updated $\mathbf{w}_i$, and assigning patients to treatment accordingly was repeated until a total of 160 patients (six cohorts) were allocated. Without early stopping, each trial consisted of six interim analyses ($I = 6$) and one final analysis.

After outcomes were simulated for all patients in the trial, test statistics T_k were calculated for all treatment arms $k \in \{1, 2, 3\}$ at the final analysis:

$$T_k = \frac{\hat{p}_k - \hat{p}_0}{\sqrt{\tilde{p}_k(1 - \tilde{p}_k)(\frac{1}{n_k} + \frac{1}{n_0})}}; \quad \hat{p}_k = \frac{r_k}{n_k}; \quad \tilde{p}_k = \frac{r_k + r_0}{n_k + n_0}$$

where n_k and n_0 respectively denote the total number of patients assigned to arm k and the control arm, and r_k and r_0 respectively denote the number of patients assigned to arm k and the control arm who responded. Arm k was selected as efficacious in this final analysis if its test statistic T_k exceeded a pre-specified critical value C_F. If more than one arm had a test statistic exceeding this critical value, the arm with the highest T_k was selected. The computation of C_F is described further in the next section.

We also performed simulations incorporating an early stopping rule for efficacy. In this case, after outcomes were simulated for 120 patients ($i = 5$) and 140 patients ($i = 6$), test statistics T_{ik} were computed as above using the accumulated data. If test statistic T_{ik} exceeded a pre-specified critical value C_S at either of these interim analyses, then the trial was terminated early and arm k was selected as efficacious. Once again, if more than one arm had a test statistic exceeding this critical value, the arm with the highest T_{ik} was selected. No early stopping rules for futility were used in this simulation study.

Type I error control

C_F is a critical value specific to each randomization method that was computed via simulation to guarantee control of the type I error at 0.05. When no early stopping rules were used, we simulated 10,000 trials under the null, where all true response rates were equal to 0.3, for under each randomization method. For each trial, the highest test statistic T_k at the end of the trial was recorded. Then, C_F for the corresponding randomization method was set to the 95^{th} percentile of these maximum test statistics. Under our selection rule, 5% of simulated trials under the null would select some treatment arm as efficacious.

For simulations with early stopping for efficacy, we set $C_S = 3$ as the critical value used in interim analyses after four and five cohorts. In this setting, we still perform simulations under the null in the same fashion, but including an early stopping rule. For a particular randomization method, C_F is chosen such that the proportion of trials that are terminated early and the proportion of trials where a treatment arm is deemed efficacious in the final analysis sum to 0.05. Consequently, C_F differs depending on whether an early stopping rule for efficacy is used.

Note that this procedure is sound because no more than one treatment arm can be selected at the end of each trial. For trials in which multiple treatment arms may be chosen (for example, under an adaptive randomization scheme tailored to the intention of choosing all efficacious treatment arms), this procedure would need to be modified to control the family-wise error rate (FWER).

Results

The tables below compare the performance of these randomization methods for selecting the best arm and selecting one effective arm, both with and without an early stopping rule for efficacy. Results for simulations under the null are not shown, since they were explicitly used to compute C_F and by definition selected one of the treatment arms with probability 0.05.

Selecting the best arm

In all scenarios examined, RAR methods substantially outperform equal randomization in selecting the best arm, supporting previous conclusions that RAR is particularly valuable in the multi-arm trial setting. FC does perform well here, although the performance of this method is highly dependent on the proportion allocated to control. SD always has similar or better performance than FC, with the advantage of not needing to fix a control allocation proportion in advance. Selection probabilities are lower and closer together across all methods in the "two efficacious arms, similar" and both "three efficacious arms" scenarios—these are cases where the difference between the response rates of the best arm and the next-best arm is very small (Tables 1 and 2).

Selecting one efficacious arm

As in the previous pair of tables, RAR methods have substantially greater power to select one efficacious arm than equal randomization does (Tables 3 and 4). Across

Table 1 Probability of selecting the best arm, no stopping rule

	One eff	Two eff, S	Two eff, D	Three eff, S	Three eff, D	Staircase
ER	0.38	0.19	0.31	0.15	0.18	0.26
FC	0.56	0.28	0.43	0.20	0.23	0.36
SD	0.58	0.27	0.44	0.22	0.25	0.38

Table 2 Probability of selecting the best arm, early stopping for efficacy

	One eff	Two eff, S	Two eff, D	Three eff, S	Three eff, D	Staircase
ER	0.38	0.20	0.32	0.16	0.18	0.27
FC	0.55	0.27	0.42	0.20	0.24	0.36
SD	0.53	0.27	0.43	0.21	0.25	0.36

Table 3 Probability of selecting one efficacious arm, no stopping rule

	One eff	Two eff, S	Two eff, D	Three eff, S	Three eff, D	Staircase
ER	0.38	0.37	0.46	0.45	0.40	0.39
FC	0.56	0.50	0.60	0.54	0.49	0.49
PT	0.59	0.54	0.65	0.61	0.55	0.54

Table 4 Probability of selecting one efficacious arm, early stopping for efficacy

	One eff	Two eff, S	Two eff, D	Three eff, S	Three eff, D	Staircase
ER	0.38	0.38	0.46	0.45	0.41	0.40
FC	0.55	0.48	0.59	0.54	0.49	0.50
PT	0.58	0.53	0.65	0.59	0.54	0.54

all scenarios examined, PT has the best performance, regardless of whether an early stopping rule is used. PT performs exceptionally well in the "three efficacious arms, similar" case—the fact that the posterior probability of each arm being the best is roughly the same on average means that methods based more closely on Thompson RAR may end up with essentially equal allocations among the treatment arms. PT, on the other hand, avoids this pitfall with its initial step in which one treatment arm may be identified. Moreover, sensitivity analyses show that the performance of PT is quite robust to the choice of the initial posterior probability threshold.

5 Discussion

There are two major classes of RAR schemes in the literature. RAR schemes in one class are optimized but with the price of complexity and reliance on large sample approximation, while RAR schemes in the other class are easy to implement without

relying on asymptotics, but are not tailored toward specific study goals. The two classes complement each other so much that it seems natural to unify both by borrow both advantages. This paper is an attempt to bridge the gap. The proposed framework aligns with the study goals, is intuitive for statisticians to derive and for investigators to understand, and its Bayesian approach allows implementation for both small and large sample sizes.

It should be noted that, although it is often straightforward to optimize the goal function in the response parameter space, in general the proposed framework does not optimize the average goal function. This is a compromise between simplicity and optimality. Despite this, the simulation results show that the proposed RAR schemes under the new framework robustly outperform FC RAR, the current popular method.

Thompson RAR has been shown to perform well in cases where there is one efficacious arm that truly stands out. It has also been pointed out that when the treatment efficacies are similar, its performance is not necessarily desirable [20]. However, the proposed RAR schemes, especially PT RAR, have higher power to detect one efficacious treatment arm among similar treatment arms.

Although the simulation study only considers binary endpoints, in general, it is easy to extend to continuous endpoints, where investigators may choose the mean or effect size as the efficacy parameter. The new framework opens a new avenue for researchers to develop simple Bayesian RAR schemes more closely tailored to specific goals, and can be applied for various endpoints with or without control arms.

References

1. Barker, A.D., Sigman, C.C., Kelloff, G.J., Hylton, N.M., Berry, D.A., Esserman, L.J.: I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clin. Pharmacolology Ther. **86**(1), 97–100 (2009)
2. Berry, D.A., Eick, S.G.: Adaptive assignment versus balanced randomization in clinical trials, a decision analysis. Stat. Med. **14**, 231–246 (1995)
3. Berry, D.A.: Adaptive clinical trials in oncology. Nat. Rev. Clin. Oncol. **9**, 199–207 (2012)
4. Bleck, T., Cock, H., Chamberlain, J., Cloyd, J., Connor, J., Elm, J., Fountain, N., Jones, E., Lowenstein, D., Shinnar, S., Silbergleit, R., Treiman, D., Trinka, E., Kapur, J.: The established status epilepticus trial 2013. Epilepsia **54**, 89–92 (2013)
5. Cheng, Y., Berry, D.A.: Optimal adaptive randomized designs for clinical trials. Biometrika **94**, 673–689 (2007)
6. Collins, S.P., Lindsell, C.J., Pang, P.S., Storrow, A.B., Peacock, W.F., Levy, P., Rahbar, M.H., Del Junco, D., Gheorghiade, M., Berry, D.A: Bayesian adaptive trial design in acute heart failure syndromes: moving beyond the mega trial. Am. Hear. J. **164**, 138-145 (2012)
7. Connor, J.T., Luce, B.R., Broglio, K.R., Ishak, K.J., Mullins, C.D., Vanness, D.J., Fleurence, R., Saunders, E., Davis, B.R.: Do Bayesian adaptive trials offer advantages for comparative effectiveness research? Protocol for the RE-ADAPT study. Clin. Trials **10**, 807–827 (2013)
8. Eisele, J.: The doubly adaptive biased coin design for sequential clinical trials. J. Stat. Plan. Inference **38**, 249–261 (1994)
9. Fiore, L.D., Brophy, M., Ferguson, R.E., D'Avolio, L., Hermos, J.A., Lewa, R.A., Doros, G., Conrada, C.H., O'Neil, J.A., Sabina, T.P., Kaufman, J., Swartz, S.L., Lawler, E., Lianga, M.H., Gaziano, M., Lavori, P.W.: A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. Clin. Trials **8**, 183–195 (2011)

10. Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., Pinheiro, J.: PhRMA Working Group: adaptive designs in clinical drug development–an executive summary of the PhRMA Working Group. J. Biopharm. Stat. **16**, 275–283 (2006)
11. Genovese, M.C., Lee, E., Satterwhite, J., Veenhuizen, M., Disch, D., Berclaz, P.Y., Myers, S., Sides, G., Benichou, O.: A phase 2 dose-ranging study of subcutaneous tabalumab for the treatment of patients with active rheumatoid arthritis and an inadequate response to methotrexate. Ann. Rheum. Dis. **72**(9), 1453–60 (2013)
12. Hu, F., Rosenberger, W.F.: Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. J. Am. Stat. Assoc. **98**, 671–678 (2003)
13. Hu, F., Zhang, L.X.: Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. Ann. Stat. **32**, 268–301 (2004)
14. Hu, F., Rosenberger, W.F.: The Theory of Response-adaptive Randomization in Clinical Trials. Wiley, Hoboken, NJ (2006)
15. Kim, E.S., Herbst, R.S., Wistuba, I.I., Lee, J.J., Blumenschein Jr., G.R., Tsao, A., Stewart, D.J., Hicks, M.E., Erasmus Jr., J., Gupta, S., Alden, C.M., Liu, S., Tang, X., Khuri, F.R., Tran, H.T., Johnson, B.E., Heymach, J.V., Mao, L., Fossella, F., Kies, M.S., Papadimitrakopoulou, V., Davis, S.E., Lippman, S.M., Hong, W.K.: The BATTLE trial: personalizing therapy for lung cancer. Cancer Discov. **1**, 44–53 (2011)
16. Komaki, F., Biswas, A.: Bayesian optimal response-adaptive design for binary responses using stopping rule. Stat. Methods Med. Res. (2016). https://doi.org/10.1177/0962280216647210
17. Krams, M., Lees, K.R., Hacke, W., Grieve, A.P., Orgogozo, J.M., Ford, G.A.: Acute stroke therapy by inhibition of neutrophils (ASTIN): an adaptive dose-response study of UK-279,276 in acute ischemic stroke. Stroke **34**, 2543–2548 (2003)
18. Lenz, R.A., Pritchett, Y.L., Berry, S.M., Llano, D.A., Han, S., Berry, D.A., Sadowsky, C.H., Abi-Saab, W.M., Saltarelli, M.D.: Adaptive, dose-finding phase 2 trial evaluating the safety and efficacy of ABT-089 in mild to moderate Alzheimer disease. Alzheimer Dis. Assoc. Disord. **29**, 192–199 (2015)
19. Lewis, R.J., Viele, K., Broglio, K., Berry, S.M., Jones, A.E.: An adaptive, phase II, dose-finding clinical trial design to evaluate L-carnitine in the treatment of septic shock based on efficacy and predictive probability of subsequent phase III success. Crit. Care Med. **41**(7), 1674–1678 (2013)
20. Lin, J., Bunn, V.: Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials. Contemp. Clin. Trials **54**, 48–59 (2017)
21. Luce, B.R., Connor, J.T., Broglio, K.R., Mullins, C.D., Ishak, K.J., Saunders, E., Davis, B.R.: Using Bayesian adaptive trial designs for comparative effectiveness research: a virtual trial execution. Ann. Med. **165**(6), 431–438 (2016)
22. Parmar, S., Andersson, B.S., Couriel, D., Munsell, M.F., Fernandez-Vina, M., Jones, R.B., Shpall, E.J., Popat, U., Anderlini, P., Giralt, S., Alousi, A., Cano, P., Bosque, D., Hosing, C., Silva Lde, P., Westmoreland, M., Wathen, J.K., Berry, D., Champlin, R.E., de Lima, M.J.: Prophylaxis of graft-versus-host disease in unrelated donor transplantation with pentostatin, tacrolimus, and mini-methotrexate: a phase I/II controlled, adaptively randomized study. J. Clin. Oncol. **29**(3), 294–302 (2011)
23. Popescu, I., Fleshner, P.R., Pezzullo, J.C., Charlton, P.A., Kosutic, G., Senagore, A.J.: The ghrelin agonist TZP-101 for management of postoperative ileus after partial colectomy: a randomized, dose-ranging, placebo-controlled clinical trial. Dis. Colon Rectum **53**, 126–134 (2010)
24. Rosenberger, W.F., Stallard, N., Ivanova, A., Harper, C.N., Ricks, M.L.: Optimal adaptive designs for binary response trials. Biometrics **57**(3), 909–913 (2001)
25. Rosenberger, W.F., Sverdlov, O., Hu, F.: Adaptive randomization for clinical trials. J. Biopharm. Stat. **22**(4), 719–736 (2012)
26. Skrivanek, Z., Gaydos, B.L., Chien, J.Y., Geiger, M.J., Heathman, M.A., Berry, S., Anderson, H., Forst, T., Milicevic, Z., Berry, D.: Dose-finding results in an adaptive, seamless, randomized trial of once-weekly dulaglutide combined with metformin in type 2 diabetes patients (AWARD-5). Diabetes Obes. Metab. **16**, 748–756 (2014)

27. Thall, P.F., Wathen, J.K.: Practical Bayesian adaptive randomization in clinical trials. Eur. J. Cancer **43**, 859–866 (2007)
28. Thall, P.F., Fox, P.S., Wathen, J.K.: Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. Ann. Oncol. **26**(8), 1621–1628 (2015)
29. Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**, 275–294 (1933)
30. Tymofyeyev, Y., Rosenberger, W.F., Hu, F.: Implementing optimal allocation in sequential binary response experiments. J. Am. Stat. Assoc. **102**, 224–234 (2007)
31. Villar, S.S., Wason, J., Bowden, J.: Response-adaptive randomization for multi-arm clinical trials using the forward looking Gittins index rule. Biometrics **71**, 969–978 (2015)
32. Wathen, J.K., Thall, P.F.: A simulation study of outcome adaptive randomization in multi-arm clinical trials. Clin. Trials **14**(5), 432–440 (2017)
33. Wei, L.J., Durham, S.: The randomized play-the-winner rule in medical trials. J. Am. Stat. Assoc. **73**, 840–843 (1978)
34. Yin, G., Chen, N., Lee, J.J.: Phase II trial design with Bayesian adaptive randomization and predictive probability. J. R. Stat. Soc. Ser. C **61**(2), 219–235 (2012)
35. Zang, Y., Lee, J.J.: Adaptive clinical trial designs in oncology. Chin. Clin. Oncol. **3**(4), 49 (2014)
36. Zelen, M.: Play the winner rule and the controlled clinical trial. J. Am. Stat. Assoc. **64**, 131–146 (1969)
37. Zhu, H., Hu, F.: Sequential monitoring of response-adaptive randomized clinical trials. Ann. Stat. **38**, 2218–2241 (2010)

Sample Size Determination Under Non-proportional Hazards

Miao Yang, Zhaowei Hua and Saran Vardhanabhuti

Abstract The proportional hazards assumption rarely holds in clinical trials of cancer immunotherapy. Specifically, delayed separation of the Kaplan-Meier survival curves and long-term survival have been observed. Routine practice in designing a randomized controlled two-arm clinical trial with a time-to-event endpoint assumes proportional hazards. If this assumption is violated, traditional methods could inaccurately estimate statistical power and study duration. This article addresses how to determine the sample size in the presence of nonproportional hazards (NPH) due to delayed separation, diminishing effects, etc. Simulations were performed to illustrate the relationship between power and the number of patients/events for different types of nonproportional hazards. Novel efficient algorithms are proposed to optimize the selection of a cost-effective sample size.

Keywords Non-proportional hazards · Sample size · Time-to-event endpoint · Log-rank test · Cancer immunotherapy · Power analysis

1 Introduction

Time-to-event endpoints such as overall survival and progression-free survival are commonly used as primary clinical endpoints in oncology clinical trials. The conventional way to design a randomized controlled two-arm clinical trial with time-to-event endpoints assumes proportional hazards between the two arms [6]. Using the log-rank test statistic to test the equality of two survival functions, the required number

M. Yang
Department of Statistics, Oregon State University, Corvallis, OR 97331, USA

Z. Hua (✉)
Alnylam Pharmaceuticals, Inc., Cambridge, MA 02142, USA
e-mail: zhua@alnylam.com

S. Vardhanabhuti
Takeda Pharmaceuticals, Cambridge, MA 02139, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_12

of events is driven by the assumption of hazard ratio (HR) at the pre-specified test significance level and the expected power level [5].

However, in clinical trials of cancer immunotherapy, the proportional-hazards assumption rarely holds presenting unique challenges for sample size determination. Cancer immunotherapy has achieved unprecedented milestones in treating life-threatening cancers such as melanoma and non-small cell lung cancer. These innovative therapies work by stimulating the immune system thereby imparting substantial benefits in tumor response and long term survival [4]. However, there is a lag in the translation of immune and anti-tumor response into a survival benefit [4] resulting in delayed separation of the Kaplan-Meier survival curves [1]. For example, the overall survival curves from CheckMate 141 trial targeting recurrent squamous-cell carcinoma of the head and neck demonstrate delayed separation around 4 months [2]. In addition, a subset of patients receiving cancer immunotherapy experience long-term survival and can be considered cured [1]. For example, the overall survival curves from CA184-024 trial targeting previously untreated metastatic melanoma show plateaus starting from approximately 3 years for both of the treatment arm of ipilimumab plus dacarbazine and the control arm of dacarbazine plus placebo [8].

In the presence of delayed separation of the survival curves or long-term survival, the assumption of proportional hazards no longer holds and conventional methods for determining sample size cannot be used. A potential consequence of ignoring delayed separation in survival curves is a potential loss of statistical power. Failing to account for long-term survival could lead to an underestimation of the study duration and, as a result, delayed access to effective therapies for patients [1]. Therefore, if nonproportional hazards are anticipated, it is important for trial design to take that into account when designing the trial.

Other sources of nonproportional hazards that have been observed in clinical trials include diminishing effects and crossing of survival curves. If diminishing effects are not considered in the trial design, statistical power could be reduced. Crossing of survival curves could also lead to power loss. More importantly, interpretation is challenging when the survival curves cross. For example, the progression-free survival (PFS) curves in the IPASS study cross around 6 months. The curves suggest a benefit of the control arm (carboplatin plus paclitaxel) before 6 months and a benefit of the treatment arm (gefitinib monotherapy) after 6 months [7]. However, the p-value based on log-rank test for the difference was less than 0.001, which is highly significant. But it is difficult to interpret that there is significant better treatment benefit for the combination of carboplatin plus paclitaxel over the gefitinib monotherapy because the PFS curves crosses around 6 months.

This article focuses on addressing how to determine the sample size using log-rank test statistic when nonproportional hazards are anticipated in designing a randomized 2-arm clinical trial with a time-to-event endpoint. Power patterns were evaluated for different types of nonproportional hazards. Novel computation-based efficient algorithms are proposed to search for cost-effective sample sizes.

2 Power Analyses Under Proportional Hazards

Sample size determination is a vital aspect of clinical trial design. If too few patients are enrolled, the study may turn out to lack statistical power to detect a clinically important treatment effect. In this section, we introduce traditional methods to determine sample size in a time-to-event trial.

In a time-to-event clinical trial where one wants to test the equality of two survival curves (treatment arm and control arm), the number of events required to achieve the power of $1-\beta$ at the significance level of α is given by

$$D = \frac{4\left(z_{\alpha/2} + z_{\beta}\right)^2}{\theta^2}, \tag{1}$$

where D is the number of events required; α, β are Type I and II errors, respectively; and θ is the log hazard ratio comparing the treatment and control arms. This equation is based on the log-rank test and requires the assumption of proportional hazards. Note that in Eq. (1), power depends on the number of events (D), but not on the sample size (N).

To show the relationship between power, the total number of events (D), and sample size (N), we ran a simple simulation. Proportional hazard assumption is made, such that the hazard ratio comparing treatment and control was set to 0.7, with the hazard rates for the control arm to be 0.05 from month 0–5, and 0.10 from month 5–10, and 0.15 otherwise. The Type I error was set to 0.05. We simultaneously varied D from 300 and 500 by 10 events and N from 500 to 1000 by 20 patients. For each combination of D and N, we determined power by Monte Carlo simulation using log-rank test. These powers are plotted in Fig. 1.

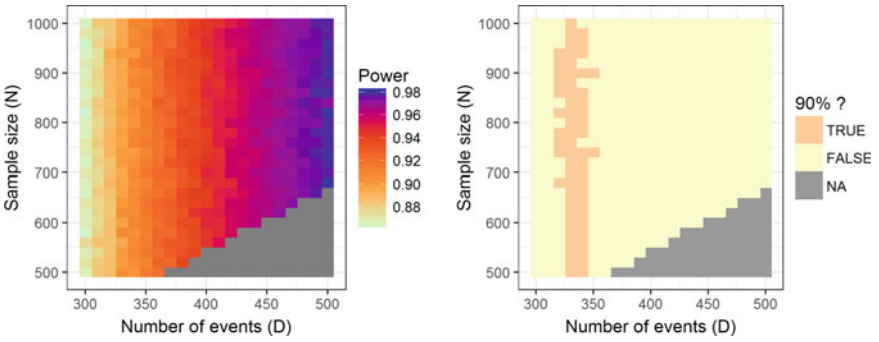


Fig. 1 Heatmap of relationsip between power and number of events (D)/sample size (N) under the proportional hazard assumption. The left panel plots the heatmap and the right panel uses the orange zone to highlight the combinations of D and N achieving the 90% power. The grey area in the lower right corner of both panels represent combinations of D and N that are infeasible under the clinical assumptions

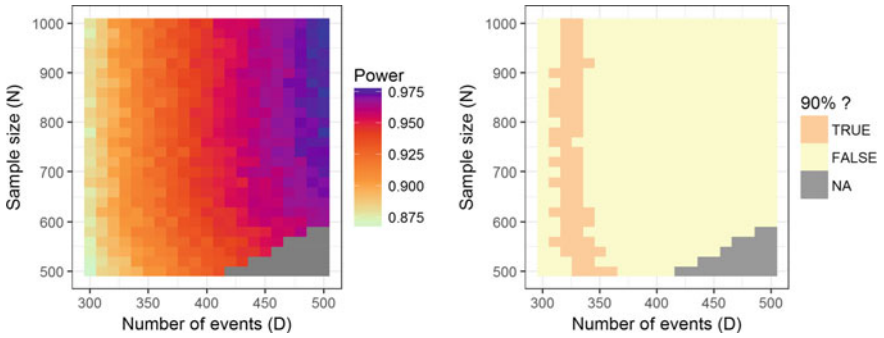


Fig. 2 Heatmap when PH assumption is slightly violated. The left panel plots the heatmap and the right panel uses the orange zone to highlight the combinations of (D, N) achieving 90% power. The grey areas in both panels are the infeasible (D, N)

Notice the vertical stripes of color in the left panel and the vertical highlighted stripe with 90% power in the right panel. In other words, for a fixed number of total events (D), change to the sample size (N) has little impact on power. In practice, with a target power (90%), we can design a study to target the same D with either larger N/shorter study duration or smaller N/longer study duration. These results support the use of Eq. (1) for power analyses under the proportional hazards assumption.

When the proportional hazards assumption is not severely violated, Eq. (1) can still be used. For example, we ran a second simulation where the hazard ratio comparing treatment and control varied over time and the survival distribution is piecewise exponential. Specifically, the hazard ratio was set to 0.65 in the time interval of months 0–5, 0.7 in months 5–10, and 0.75 in months $\geq$10. The heatmap of power for this simulation is plotted in Fig. 2. Note that the color patterns are similar to those of Fig. 1, suggesting Eq. (1) can be applied when the hazard ratio varies slightly over time. In this case, a reasonable guess of constant which represents the overall trend should be used.

3 Power Analyses Under Non-proportional Hazards

In many clinical trials, the PH assumption is severely violated making power analyses more challenging because Eq. (1) cannot be used. This is in part due to the fact that the relationship between power and D/N is more complex when the hazards are nonproportional. In this section, we explore this relationship via simulations under two sources of nonproportionality: delayed and diminishing treatment effects. Figure 3 shows sample survival curves for both scenarios. The left panel is a delayed effect model with hazard ratios to be 1 for the first 5 months and 0.7 otherwise, and the right panel is a diminishing effect model with hazard ratios increasing from 0.5 to 1 by 0.1 for every 1.2 months.

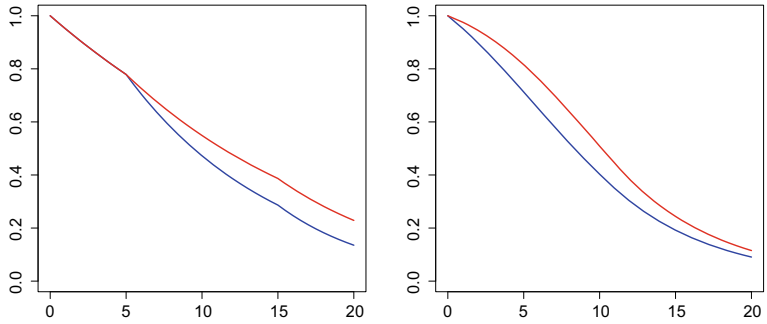


Fig. 3 Survival curves for NPH simulation

Under both cases, log-rank test is performed. Even though log-rank is no longer the most power test under these two scenarios, we still apply it in our simulations. This is because the relationship of power with D and N are similar regardless of which test we perform, and the only difference is that in order to achieve the same power, log-rank test may require larger combinations of (D, N) than other tests. Therefore 90% power under these two scenarios with log-rank tests will need much larger (D, N) and generate higher computing burden. To save computational time, we will set our target power to be lower (70%, 78% in delayed and diminishing effect models, respectively) in this section.

Figure 4 shows the power heatmap for the delayed treatment effect model with no effect in the first 5 months ($HR = 1$) and strong effect otherwise ($HR = 0.7$). The left panel shows that, for a fixed value of D, increasing N will cause power to decrease. The intuition for this finding is that, when there is a larger N, there will be higher proportion of early events occurring at early time points when there is no difference between the two arms, thus the test is less powered to detect a difference between the arms. As of the same reason of higher proportion of early events, a shorter study duration is required to observed D events with a larger N. The right panel uses the red band to highlight the (D, N) pairs which generate 70% power using log-rank test. Unlike Figs. 1 and 2, this band is no longer vertical and suggests that as either D or N increase, the other must also increase in order to maintain the same power. The practical implication is, we can no longer design a study to target the same D with either larger N/shorter study duration or smaller N/longer study duration to generate the same power. Instead, we need to run a simulation study to look for an optimal combination (D, N).

Figure 5 shows the power heatmap for the diminishing effects model with HR increasing from 0.5 to 1 by 0.1 for every 1.2 months. The left panel shows that, for a fixed value of D, increasing N will cause power to initially increase and then level off. When N increases, there will be higher proportion of early events occurring at early time points where there is larger treatment benefit, and the test is more powered to detect a difference between the arms. When N reaches to a limit such as that adding additional subjects will not contribute additional events to the fixed D, power will

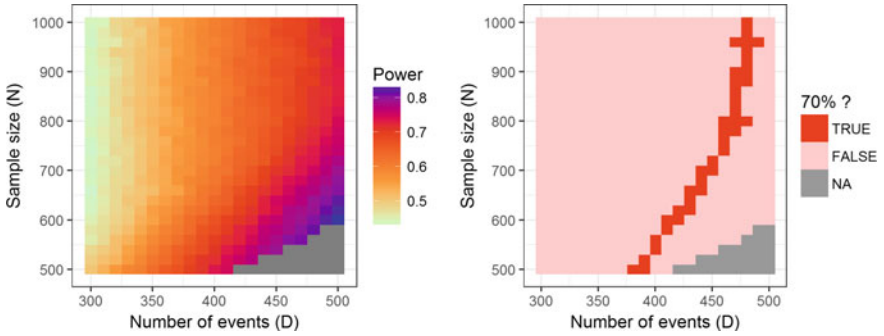


Fig. 4 Heatmap for a delayed effect model with hazard ratio 1 in (0, 5) and 0.7 otherwise

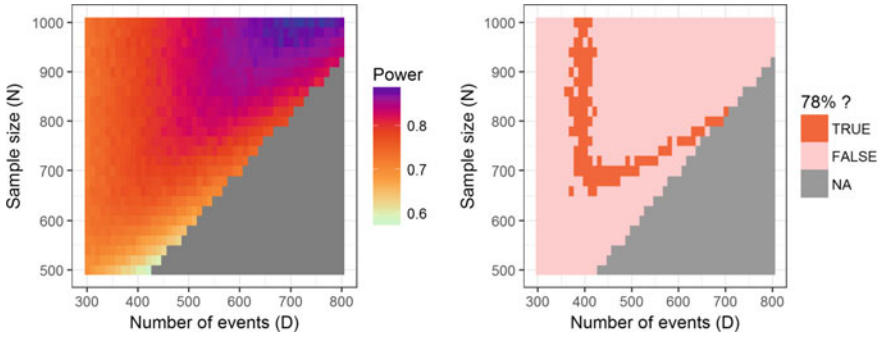


Fig. 5 Heatmap for a diminishing effect model with hazard ratio increasing from 0.5 to 1

stay stable. Again, with a larger sample size, a shorter study duration is required to observe D events. For a fixed value of N, increasing D will cause the power to initially increase and then decrease. New events that occur at early time points will increase the power to detect a difference between treatment and control. However, increasing D will also prolong the study and new events occurring at later times (when there is no difference between the treatment and control arms) will reduce the power. The right panel uses the orange band to display the (D, N) pairs with 78% power, which is made up of a vertical band and an increasing band. Regarding the increasing band, it is straightforward that increasing D without increasing N will cause higher proportion of late events at later time points where there is smaller treatment effect size, thus power will be lower than the target power. So increasing D will require larger N to maintain the appropriate proportion of early and late events to maintain the same power. Regarding the vertical band, the bottom point of minimal combination (D, N) corresponds to the limit of N such that adding additional subjects no longer contribute additional events to the minimal D. Thus when we choose the minimal D, increasing N will have no impact on power. Similarly in practice, we should run simulation to identify an optimal combination (D, N).

Therefore, in the presence of NPH, power depends on both the number of events and the number of subjects. When we design a study, a wide range of the combinations of (D, N) should be explored via simulation. The optimal combination of (D, N) will need to consider other factors such as cost and study duration.

4 Algorithm for Detecting Sample Size Under NPH

In the previous section, we demonstrated that power changes with both D and N in the presence of NPH. Unfortunately, there is no equation to determine the sample size and the number of events needed for a given power level when there is nonproportionality. In addition, we also showed that multiple combinations of sample size and the number of events yield the same power. In both Figs. 4 and 5, there is a $(D_{\min}, N_{\min})$ pair that requires the fewest patients and events for a given power level. While the $(D_{\min}, N_{\min})$ pair may not yield an optimal trial in terms of other factors, such as follow-up time and study duration, they do serve as a useful lower bound in finding the "best" (D, N). What is "best" is generally determined by the resources one has. In this section, we propose algorithms to search for $(D_{\min}, N_{\min})$ for both delayed and diminishing effects models.

The literature contains numerous search algorithms including the binary search algorithm. This algorithm is efficient. In order to target $(D_{\min}, N_{\min})$ using the binary search algorithm, power must be either monotone with D for fixed value of N or monotone with N for a fixed value of D, or both.

Figure 4 showed that, in a trial with delayed clinical effects, power monotonically decreases with increasing N for a fixed value of D. If our goal is to find $(D_{\min}, N_{\min})$, we would select the pair at the bottom left of the orange band in the right panel.

Note that for any D, one can always search for an N, such that power (D, N) is larger than any other N, where power (D, N) is the power of a test (for example log-rank) when there are N patients and D events. So we can search until we find a smallest $D_{\min}$, and for each D we can always get the corresponding N given the targeted power. In both search steps, binary search algorithm will be performed. We would take the following steps:

- Initialization a small D^0, tolerance parameter ϵ, moving windows Δ_D, Δ_N, type I and II errors α, β.
- Find $N^0 = \arg\max_N power\left(D^0, N\right)$;
- Use binary search to go through a series of D to update D_0; Initially start with increment of Δ_D, if returning to smaller D, using decrement of $\Delta_D/2$; Every time changing direction, set moving window $\Delta_D = \Delta_D/2$;
- Find $N_{\min} = \arg\max_N power\,(D_{\min}, N)$.

In the case of diminishing effects, power is no longer monotone with D for a fixed value of N, which suggests that a binary search may not be appropriate. However if we start with a huge N (for example 1, 500 for the model in Fig. 5), power

Table 1 Minimum numbers of events and sample sizes needed for 90% power in the case of delayed and diminishing effects using a binary search algorithm

Case	Test	$D_{\min}$	$N_{\min}$	$Power$	n_{iter}	Time (min)
Delayed effects	LR	611	731	0.9040	8	1.169
	FH(0,1)	531	651	0.9081	6	0.959
Diminishing effects	LR	751	1, 006	0.8915	8	4.189
	FH(1,0)	383	628	0.8950	6	3.014

Abbreviations: LR = log-rank; FH = Fleming-Harrington

is approximately monotone increasing with D. Therefore the searching algorithm contains two binary search steps:

- With a huge N, since a huge N will correspond to $D_{\min}$ based on the vertical band in the right panel of Fig. 5, use binary search algorithm to find $D_{\min}$ (that is, approximately 400 events for 78% power);
- With $D_{\min}$, go through the vertical band in the right panel of Fig. 5 to search $N_{\min}$ via binary search algorithm.

We apply the above algorithms to search for $(D_{\min}, N_{\min})$ for both the delayed and diminishing effects using both the log-rank and Fleming-Harrington (FH) tests [3]. The resulting $(D_{\min}, N_{\min})$'s are given listed in Table 1 along with the empirical power based on 10, 000 simulations, the number of iterations in the algorithm, and the computation time in minutes. For all four cases, the number of iterations was small and the computation time was short suggesting that the algorithm was efficient. In addition, all four of the empirical powers are close to the target power of 90%. Please note unlike generating heatmap that higher power (90%) will require much longer computational time, the binary search for $(D_{\min}, N_{\min})$ is very efficient. So we use 90% target power to illustrate efficiency of binary search algorithm. One more thing to note is that for the delayed effects case, the fact that FH(0,1) selected smaller values of $(D_{\min}, N_{\min})$ than LR indicates that FH(0,1) is more powerful than LR in this case. An even stronger trend is seen in the diminishing effects case indicating that FH(1,0) is much more powerful than LR.

5 Discussion

Nonproportional hazards have become a real challenge for biostatisticians tasked with determining sample sizes for clinical trials. In the presence of nonproportional hazards due to both delayed and diminishing effects, we first demonstrated how the power is related to the number of events and sample size. We then proposed computationally efficient algorithms for each case to search for the combined minimum number of events and patients $(D_{\min}, N_{\min})$ that meets the desired power level.

Our analysis focused on only two types of nonproportional hazards. We also did not consider other aspects of clinical trials which play important roles in power analyses such as accrual and dropout. Further research is necessary to determine sample sizes under more types of NPH while taking into account additional important factors.

We use heatmap to illustrate the power pattern under different combinations of (D, N) with using log-rank test. Such pattern of power depending on number of events and number of subjects maintains if other tests such as Fleming-Harrington weighted log rank test are used. For example, if using Fleming-Harrington(1, 0) with weight function $\widehat{S}(t)$ (pooled survival estimate) to allocate more weights to early time points, the power pattern is similar under diminishing effects with a shift to smaller D as Fleming-Harrington(1,0) is more powerful than log-rank test under diminishing effects; if using Fleming-Harrington(0,1) with weight function $1 - \widehat{S}(t)$ to allocate more weights to late time points, the power pattern is similar under delay separation with a shift to smaller D as well because Fleming-Harrington(0,1) is more powerful than log rank test under delay separation.

References

1. Chen, T.T.: Statistical issues and challenges in immuno-oncology. J. Immuno. Ther. Canser. **1**, 18 (2013). https://doi.org/10.1186/2051-1426-1-18
2. Ferris, R.L., Blumenschein Jr., G., Fayette, J., et al.: Nivolumab for recurrent squamous-cell carcinoma of the head and neck. N. Engl. J. Med. **375**, 1856–1867 (2016)
3. Fleming, T.R., Harrington, D.P., O'sullivan, M.: Supremum versions of the log-rank and generalized Wilcoxon statistics. J. Amer. Statist. Assoc. **82**(397), 312–320 (1987)
4. Hoos, A.: Evolution of end points for cancer immunotherapy trials. Ann. Oncol. **23**(8), viii47–viii52 (2012)
5. Lakatos, E.: Sample sizes based on the log-rank statistic in complex clinical trials. Biometrics, 229–241 (1988)
6. Lachin, J.M., Foulkes, M.A.: Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics, 507–519 (1986)
7. Mok, T.S., Wu, Y.L., Thongprasert, S., et al.: Gefitinib or carboplatin paclitaxel in pulmonary adenocarcinoma. N. Engl. J. Med. **361**(10), 947–957 (2009)
8. Robert, C., Thomas, L., Bondarenko, I., et al.: Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N. Engl. J. Med. **364**(26), 2517–2526 (2011)

Adaptive Three-Stage Clinical Trial Design for a Binary Endpoint in the Rare Disease Setting

Lingrui Gan and Zhaowei Hua

Abstract A fundamental challenge in developing therapeutic agents for rare diseases is the limited number of eligible patients. A conventional randomized clinical trial may not be adequately powered if the sample size is small and asymptotic assumptions needed to apply common test statistics are violated. This paper proposes an adaptive three-stage clinical trial design for a binary endpoint in the rare disease setting. It presents an exact unconditional test statistic to generally control Type I error when sample size is small while not sacrificing power. Adaptive randomization has the potential to increase power by allocating greater numbers of patients to a more effective treatment. Performance of the method is illustrated using simulation studies.

Keywords Rare disease · Small clinical trial · Z-pooled unconditional Test · Type I error · Combination method

1 Introduction

Rare diseases are often life threatening or chronically debilitating. Because of their low prevalences (e.g., a diseases is defined as rare if its prevalence is 5 per 10, 000 in Europe Union [1]), the fundamental challenge in conducting rare disease clinical trials is to make appropriate, conclusive inference with a small sample size.

Small sample size brings a caveat in use of asymptotic test statistics. Inappropriate use of asymptotic test statistics when the pillar of large sample is not present will lead to Type I error inflation or large Type II error. For example, the normal way

L. Gan
Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
e-mail: lgan6@illinois.edu

Z. Hua (✉)
Alnylam Pharmaceuticals, Inc., Cambridge, MA 02142, USA
e-mail: zhua@alnylam.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_13

of comparing the means in two groups is using two sample Z test. However, when samples sizes are less than 30, this approach becomes problematic [7]. Similarly, the normal way of comparing the proportion across two groups, chi-square statistic, becomes problematic when expected cell counts are less than 5 [8]. In addition, conventional randomized controlled designs require sufficiently large samples to maintain power and control Type I error. Therefore, alternative clinical trial designs are needed in the setting of rare diseases.

Alternative clinical trial designs have been proposed in the literature. (1) One strategy is to collect richer information from each individual patient, e.g., cross-over design and N of 1 design. Cross-over designs randomize patients to two treatment orders of treatment followed by control or vice versa [2]. N of 1 designs randomize a single patient sequentially to a sequence of treatment and control [3]. These designs, however, rely on every patient taking part in a placebo stage equal in length to the treatment stage. In addition, when expectations for the drug are high, the placebo stage may pose a recruitment problem. (2) Another strategy is to run a lead-in period to select a subset to patients that will be randomized to receive either treatment or control, e.g., randomized withdrawal design, the early escape design and response-adaptive randomization design. After the lead-in period, the randomized withdrawal designs enrich to the subset of high-sensitive patients to the investigated treatment [4]. The early escape designs drop early failures [5]. Response-adaptive randomization designs increase power by allocating patients with higher probability to the more effective treatment [6]. However, because only rather limited information is collected from each patient, using these designs is often not sufficient to achieve an acceptable statistical power when sample size is small.

Learned from both of the strategies, a three-stage design [9] combines an ordinary randomized trial with a randomized withdrawal trial to allow collecting richer information from each patient while also enriching to responders to treatment after the first stage. It increases the statistical power by collecting richer information from each patient, while avoiding the aforementioned recruitment problem.

It consists of an initial randomized-controlled stage, followed by a randomized withdrawal stage for responders to treatment, and another randomized stage of secondary responders to treatment who initially are non-responders to control. The chi-square statistics is used to test on the difference between the proportion across two groups. Although the statistical power is increased compared with the one stage randomized design, because asymptotic test is used, the Type I error of this design is not controlled, i.e., if we test the null hypothesis at the alpha level of α, the Type I error can not be controlled under the nominal level α.

In this paper, we introduce an adaptive three-stage clinical trial design which combine features of a three-stage design with a response-adaptive design. This design retains the benefits of the three-stage design and optimizes power of the design by allowing larger chances for responders to be randomized to more effective treatment after the first stage. In addition, instead of the chi-square statistics, Z-pooled unconditional test for a 2×2 contingency table [18] is recommended to use to generally control the study-wise Type I error when sample size is not adequately large to use chi-square test.

To introduce the adaptive three-stage design, the paper is organized as follows. The framework of the adaptive three-stage clinical trial design for a binary endpoint in the setting of rare diseases is described in Sect. 2. Performance of the proposed design in comparison with other approaches is illustrated in Sect. 3. Section 4 provides further discussion.

2 Methods

2.1 Adaptive Three Stage Randomization Design

The proposed design, illustrated in Fig. 1, embeds adaptive randomization within a three-stage design. The utility of a three-stage trial is to collect richer information from individual patient enrolled. This type of design is also attractive to patients because they have a better chance of receiving an efficacious treatment which benefits enrollment and retention.

The general procedure of the design is as follows. In Stage 1, patients are randomized to receive either treatment or control with equal probability. Treatment responders and control non-responders at Stage 1 proceed to the next stages. For ethical reasons, the follow-up of treatment non-responders and control responders ends. It would not be ethical to continually expose treatment non-responders to potentially toxic effects nor to require control responders to switch to treatment and incur additional expense.

In Stage 2, treatment responders from Stage 1 are re-randomized to either treatment or control. However, they are no longer randomized with equal probability. Instead, the randomization ratio is adaptive in that it depends on the response rates from Stage 1. Patients are randomized to the more efficacious treatment with greater probability.

In Stage 2.5, control non-responders from Stage 1 are assigned to receive treatment. In Stage 3, the treatment responders from Stage 2.5 are re-randomized to either treatment or control. Again, adaptive randomization (AR) based on the relative efficacy learned from Stage 1 and Stage 2 is used to determine the randomization ratio for Stage 3.

In Stages 1, 2 and 3, two-by-two contingency tables and their associated p-values p_1, p_2 and p_3 are computed. To appropriately control Type I error while also maximizing power, the Z-pooled unconditional test [18] is proposed to be used as the test statistics, instead of the conservative Fisher's exact test. Combination methods such as Stouffer's Z-score method [16] are used to combine these p-values and test whether there is a significant difference in response rates between patients assigned to treatment and control. Details on adaptive randomization (AR), test statistics and combinations methods are discussed in Sects. 2.2, 2.3 and 2.4.

2.2 Response-Adaptive Randomization

The main feature of response-adaptive randomization [10, 11] in clinical trials features is that the randomization ratio can adapt to favor a treatment regimen, using Bayesian framework, which has shown superior efficacy in earlier stages (i.e., higher response rate). In this section, we will briefly go through the adaptive randomization technique.

First, we will first define some notations. In Stage 2 and 3 of the designed trial, shown in Fig. 1, we use the adaptive randomization. For the outcomes gathered before each Stage (i.e., the outcomes in Stage 1 when adaptive randomizing for Stage 2 and the outcomes in Stage 1 and Stage 2 when adaptive randomizing for Stage 3), we denote n_j as the number of patients enrolled in treatment j, where $j = 1$ for control and 2 for treatment, $X_{i,j}$ as the outcome of patient i on treatment j, which takes 1 for response and 0 otherwise, and denote $X_j = \sum_i X_{i,j}$ as the total number of responders in treatment j. The adaptive randomization follows the following Bayesian structures:

$$\begin{aligned} X_{i,j}|\pi_j &\sim \mathsf{Bern}(\pi_j), j = 1, 2; \\ \pi_j &\sim \mathsf{Beta}(\alpha_j, \beta_j). \end{aligned} \tag{1}$$

where π_j is the response rate in the treatment j and α_j, β_j are the hyper-parameters summarized from the prior information.

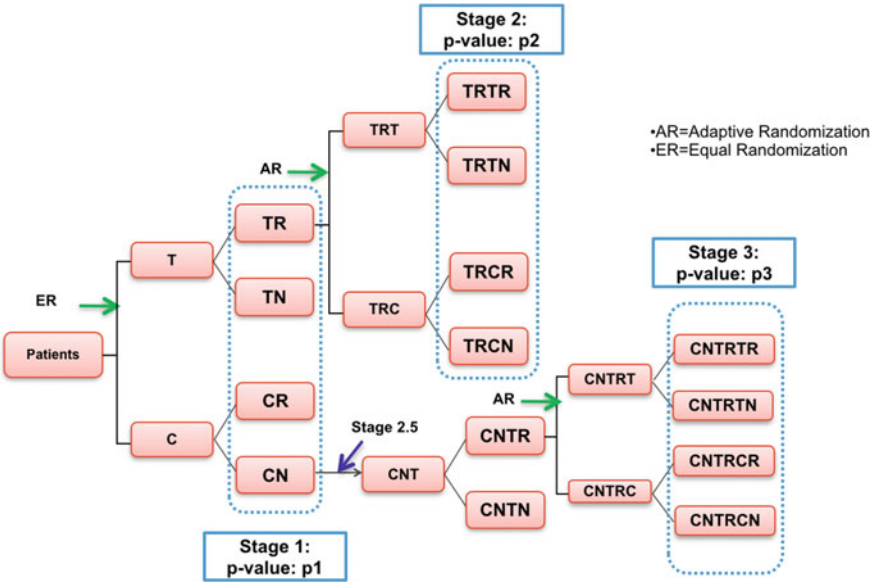


Fig. 1 Clinical Trial Design Diagram. Define T, C, R, N as treatment, control, responder and non-responder respectively, we use the combinations of these symbols to denote the sizes of every outcome groups in each stage. For example, TRCN in Stage 2 denotes the size of patients who respond in the treatment group in Stage 1, and then respond to control afterwards in Stage 2

The corresponding posterior distribution $\pi_j|X_j$, denoted as Y_j, then follows a Beta distribution, $\mathsf{Beta}(\alpha_j + X_j, \beta_j + n_j - X_j)$. According to the posteriors, we will allocate patients to arm j with the probability:

$$\rho_j = \frac{p_j^\lambda}{p_1^\lambda + p_2^\lambda}, \tag{2}$$

where $\lambda \geq 0$ and $p_j = P(Y_j > Y_{j'})$. The stochastic inequality $P(Y_j > Y_{j'})$ is the probability that treatment j is better than treatment j'. It could be computed efficiently and deterministically by calculating a hypergeometric function [10].

The parameter λ plays a role of shrinkage and controls how much information we want to learn from the previous information for the adaptive randomization. When $\lambda = 0$, the assignment is equivalent to equal randomization (ER). The shrinkage from assigning according to the posterior distributions from the previous observations to ER is controlled through changing λ from 1 to 0. Proper tuning of λ could lead to optimal adaptive randomization, i.e., maximizing power. For more guidances on how we choose λ in the adaptive randomization, we refer to Wathen and Cook (2006) [11].

To facilitate response-adaptive randomization, Jeffery's prior $\mathsf{Beta}(\frac{1}{2}, \frac{1}{2})$ [19] is combined with data from Stage 1 to determine the adaptive randomization ratio for Stage 2. To determine the randomization ratio for Stage 3, prior information from Stage 1 is combined with the data from Stage 2 to determine posterior assigning probabilities for Stage 3. Specifically, the hyperparameters α_j, β_j in the prior are set to be proportional to the numbers of responders and non-responders in the corresponding treatment arm from the previous stages and to make the prior only weakly informative.

2.3 Test Statistics

As previously mentioned, the commonly used asymptotic Pearson's chi-squared test may not be appropriate in the rare disease setting [12]. When the large sample requirement is not met, the use of Pearson's chi-squared test could lead to Type I error inflation. In order to obtain p-values from two-by-two contingency tables with small cell counts while maintaining Type I error control, we consider alternatives to Pearson's chi-square test: Fisher's exact test and the Z-pooled unconditional exact test.

It is feasible to calculate the exact power using both Fisher's exact test and the Z-pooled unconditional exact test. However, in the rare disease setting when sample sizes are small, we would prefer to use the test with higher power.

	Respond	Not respond
Treatment	x_1	$N_1 - x_1$
Control	x_2	$N_2 - x_2$

Suppose we have a two by two table as above, the exact power of a level α test is given by:

$$\text{Power}|\theta_1, \theta_2 = \sum_{(x_1,x_2)\in R} \binom{N_1}{x_1}\binom{N_2}{x_2}\theta_1^{x_1}(1-\theta_1)^{N_1-x_1}\theta_2^{x_2}(1-\theta_2)^{N_2-x_2}, \qquad (3)$$

where N_1, N_2 are the numbers of patients enrolled, x_1, x_2 are the numbers of responders, and θ_1, θ_2 are the true response rates for the treatment and control arms, respectively. The power function depends on the rejection region R for a test.

Both Fisher's exact test and the Z-pooled unconditional exact test assume a fixed total number of patients: $T = t$. Fisher's exact test also assumes fixed margins, i.e., row sums are fixed as N_1, $t - N_1$, and column sums are fixed as $x_1 + x_2$, and $t - (x_1 + x_2)$. Conditioning on the total number of patients $T = t$, the Fisher's exact test becomes a test statistic based on the one-dimensional distribution of X_1. The exact p-value for a one-sided Fisher's exact test is given by:

$$\text{P-value} = \sum_{i\in R} P_{H_0}(X_1 = i|T = t),$$

where

$$P_{H_0}(X_1 = i|T = t) = \frac{\binom{N_1}{i}\binom{t-N_1}{x_1+x_2-i}}{\sum_j \binom{N_1}{j}\binom{t-N_1}{x_1+x_2-j}}.$$

Fisher's exact test is conservative, because of the additional assumption of fixed margins. The Z-pooled unconditional test, on the other hand, only assumes fixed row sums. The exact p-value is calculated as the supremum of the exact probabilities for all possible θ under the null, assuming binomial distributions of X_1 and X_2:

$$\text{P-value} = \sup_{0\le\theta\le 1}\Big\{\sum_{i=0}^{N_1}\sum_{j=0}^{N_2} P_{H_0}(X_1 = i, X_2 = j|\theta) \times 1_{|Z_P(i,j)|\ge|Z_P(x_1,x_2)|}\Big\}, \qquad (4)$$

where

$$P_{H_0}(X_1 = x_1, X_2 = x_2|\theta) = \binom{N_1}{x_1}\binom{N_2}{x_2}\theta^{x_1+x_2}(1-\theta)^{N_1+N_2-x_1-x_2},$$

and the Z-pooled statistic is used to define the rejection region:

$$Z_P(x_1, x_2) = \frac{\hat\theta_1 - \hat\theta_2}{\sqrt{\tilde\theta(1-\tilde\theta)(\frac{1}{N_1}+\frac{1}{N_2})}},$$

where $\hat\theta_i = \frac{x_i}{N_i}$ and $\tilde\theta = \frac{x_1+x_2}{N_1+N_2}$.

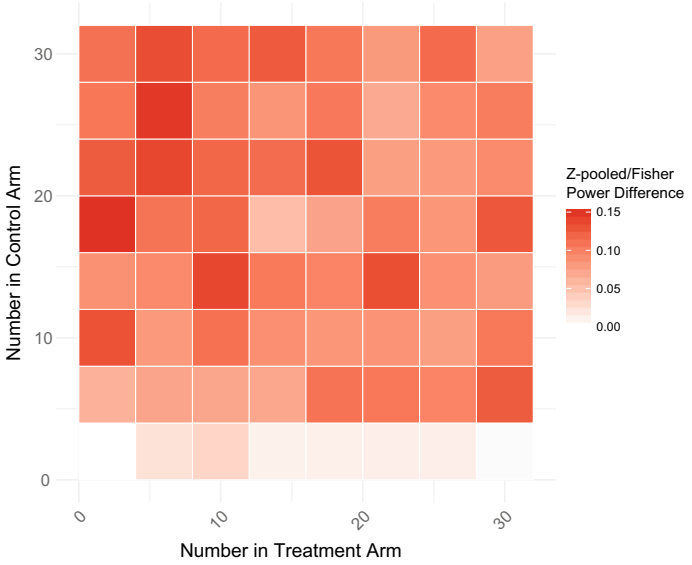


Fig. 2 The numeric difference between the exact powers of the Z-pooled unconditional exact test and the Fisher's conditional exact test at the 5% significance level for samples of patients ranging in size 2–30 per treatment arm. The statistical powers of Z-pooled unconditional exact test are uniformly greater than the ones of Fisher's exact test

The optimal θ is searched over the space [0, 1] to maximize the p-value for the two-sided test. A one-sided p-value for the Z-pooled unconditional test can also be obtained by summing the probabilities on one side of the rejection region. To gain computational efficiency in identifying the optimal θ, the interval method [13] is used by restricting optimization into a confidence set of highly likely regions of θ.

Mehrotra et al. [14] commented that Fishers exact test is generally less powerful than the Z-pooled unconditional test. We also conducted a study to compare power between the Fisher's exact test and the Z-pooled unconditional exact test. As a toy example, consider a sample of patients from 2 to 30 for each treatment arm with the treatment and control response rate at 0.4 and 0.2 respectively. Figure 2 shows that the Z-pooled unconditional exact test is uniformly more powerful than the Fisher exact test. Therefore, we chose to use the Z-pooled unconditional exact test for the calculations of p-values in our analysis.

2.4 Combination Methods

Recall from Sect. 2.1 that our three-stage design generates three p-values, p_1, p_2, and p_3. In order to combine these p-values into a single p-value for the difference between groups, we consider two combination methods: Fisher's combination method [15] and Stouffer's Z-score method [16].

Assuming p_1, p_2, and p_3 are independent, the two combination methods utilize different distributions to build the test statistics. Specifically, the Fisher's combination test combines the three individual p-values into a single p-value that follows a chi-square distribution (i.e., $-2\sum_{i=1}^{3}\log(p_i) \sim \chi^2_{2\times 3}$). The Stouffer's Z-score method first transforms each p-values p_i, $i = 1, 2, 3$ into Z score $Z_i = \Phi^{-1}(1 - p_i)$ where Φ is the cumulative distribution function for the standard normal distribution. It then combines them into be a single p-value that follows standard normal distribution, i.e., $\frac{\sum_{i=1}^{3} Z_i}{\sqrt{3}} \sim N(0, 1)$. If there is a stage having patients fewer than 4 (i.e., at least one cell of the contingency table is zero), the p-value calculated can only be at its extreme, i.e., 0 or 1. It hurts the correctness of decision making and often makes the combination method invalid. Thus, when a stage having patients fewer than 4, it will not be included when p-values are combined and decisions are made. For example, if there are more than 4 patients included in Stages 1 and 2, but fewer than 4 patients at Stage 3, decisions will be made based on the p-value combined from p_1 and p_2.

3 Simulation Studies

In this section, we compare our models with the competitive alternatives in a set of simulated rare-disease scenarios and demonstrate the performance of our approach in the Type I error control and the statistical power improvement.

We used a similar strategy as Honkanen et al. [9] to calculate the exact Type I error and power. Pseudo-code for the calculation is provided in Algorithm 1. Clinical trials are simulated for every possible case according to the proposed design in Fig. 1. Patients are first equally randomized to treatment and control in Stage 1. We calculate the probability P for each downstream possibility in terms of numbers of responses at each stage.

Define T, C, R, N as treatment, control, responder and non-responder respectively, we use the combinations of these symbols to denote the sizes of every outcome groups in each stage. Specifically, we denote the numbers of patients assigned to treatment and control as n_T and n_C respectively. We further assume a clinical trail has n_{TR} responders to treatment among n_T patients randomized to treatment and n_{CR} responders to control among n_C patients randomized to control in Stage 1. At Stage 2, n_{TRT} and n_{TRC} ($n_{TR} = n_{TRT} + n_{TRC}$) patients are re-randomized to either treatment or control, and n_{TRTR} respond to treatment and n_{TRCR} respond to control. At Stage 2.5, n_{CNTR} respond to treatment among n_{CNT} patients. At Stage 3, n_{CNTRT} and n_{CNTRC} ($n_{CNTR} = n_{CNTRT} + n_{CNTRC}$) patients are re-randomized to either treatment or control, and n_{CNTRTR} respond to treatment and n_{CNTRCR} respond to control. The response rates for each of these groups is denoted as $p_T, p_C, p_{TT}, p_{TC}, p_{CT}, p_{CTT}$ and p_{CTC} according to the assignments each of these groups received. The probability P for this scenario is calculated as:

$$\binom{n_T}{n_{TR}} p_T^{n_{TR}} (1-p_T)^{n_T - n_{TR}} \binom{n_C}{n_{CR}} p_C^{n_{CR}} (1-p_C)^{n_C - n_{CR}}$$
$$\binom{n_{TRT}}{n_{TRTR}} p_{TT}^{n_{TRTR}} (1-p_{TT})^{n_{TRT} - n_{TRTR}} \binom{n_{TRC}}{n_{TRCR}} p_{TC}^{n_{TRCR}} (1-p_{TC})^{n_{TRC} - n_{TRCR}}$$
$$\binom{n_{CNT}}{n_{CNTR}} p_{CT}^{n_{CNTR}} (1-p_{CT})^{n_{CNT} - n_{CNTR}} \binom{n_{CNTRT}}{n_{CNTRTR}} p_{CTT}^{n_{CNTRTR}} (1-p_{CTT})^{n_{CNTRT} - n_{CNTRTR}}$$
$$\binom{n_{CNTRC}}{n_{CNTRCR}} p_{CTC}^{n_{CNTRCR}} (1-p_{CTC})^{n_{CNTRC} - n_{CNTRCR}}, \tag{5}$$

We analyze each simulation scenario from the adaptive three stage randomization design by using exact numeration as discussed in Algorithm 1.

Algorithm 1 Calculating Exact Power

```
Initialize Power=0
for TR in 0:[n/2] do
  for CR in 0:n-[n/2] do
    TN=[n/2]-TR; CN=n-[n/2]-CR
    Assign treatment responders to TRT and TRC by AR.
    for TRTR in 0:TRT do
      for TRCR in 0:TRC do
        TRTN=TRT-TRTR; TRCN=TRC-TRCR;
        CNT=CN;
        for CNTR in 0:CNT do
          CNTN=CNT-CNTR;
          Assign to CNTRT and CNTRC by AR.
          for CNTRTR in 0:CNTRT do
            for CNTRCR in 0:CNTRC do
              CNTRTN=CNTRT-CNTRTR;
              CNTRCN=CNTRC-CNTRCR;
              Calculate probability P according to (5)
              If the null hypothesis is rejected, Power← Power+P
            end for
          end for
        end for
      end for
    end for
  end for
end for
Return
```

We considered scenarios where the sample size n is set to be 20, 25, 30, 35, 40, 45, 50 and 55. To compute the 3 p-values, i.e., p_1, p_2, p_3, for Stages 1, 2 and 3, the Z-pooled unconditional test is used. In our studies, both Fisher's combination method and Stouffer's Z-score method are considered to combine the 3 p-values and to test the null hypothesis at the alpha level of 5%. We compare these results to those of a one-stage equal randomization design with the same sample sizes.

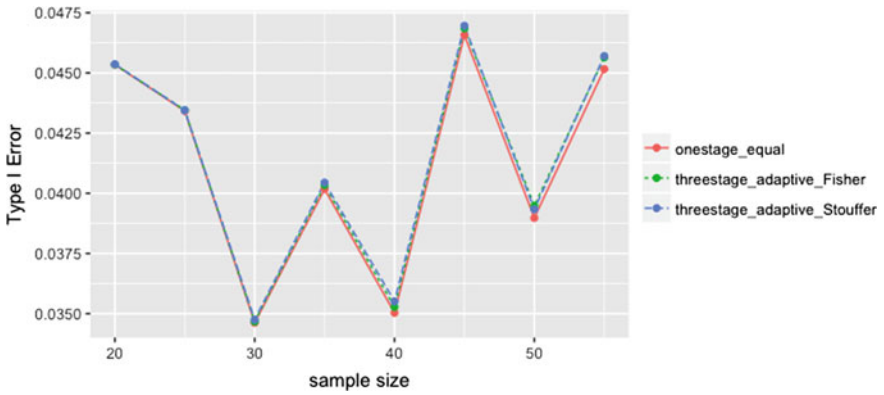


Fig. 3 Null case 1—Type I error when the response rate is uniformly 20% across stages for both treatment and control

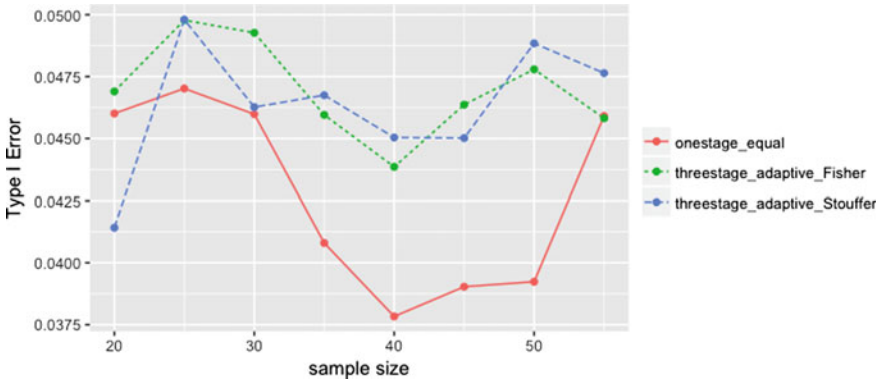


Fig. 4 Null case 2—Type I error when the response rate is uniformly 40% across stages for both treatment and control

Type I error control is illustrated via 2 null cases: (1) $p_T = p_{TT} = p_{CT} = p_{CTT} = p_C = p_{TC} = p_{CTC} = 0.2$; (2) $p_T = p_{TT} = p_{CT} = p_{CTT} = p_C = p_{TC} = p_{CTC} = 0.4$. Results are shown in Figs. 3 and 4, and adaptive three stage randomization designs with both Fisher's combination method and Stouffer's Z-score method have Type I error within the nominal 5% level. Compared with the one stage randomization design, these two methods have Type I error closer to the nominal 5% level.

Three alternative cases are presented to show the statistical power of detecting a significant treatment effect. In the first two cases, treatment is beneficial (i.e., the response rate is higher for treatment than control) and the response rates are constant across stages. Case 1: $p_T = p_{TT} = p_{CT} = p_{CTT} = 0.4$; $p_C = p_{TC} = p_{CTC} = 0.2$; Case 2: $p_T = p_{TT} = p_{CT} = p_{CTT} = 0.5$; $p_C = p_{TC} = p_{CTC} = 0.2$. The third case assumes a mixed population in which 20% of subjects always respond to treatment and never respond to control, and the remaining 80% of subjects have a 25% chance

of responding to either treatment or control. This population composition results in higher response rates for treatment and lower response rates for control at later stages:

- Stage 1: $p_T = 0.4, p_C = 0.2$;
- Stage 2: $p_{TT} = 0.625, p_{TC} = 0.125$;
- Stage 2.5: $p_{CT} = 0.4375$;
- Stage 3: $p_{CTT} = 0.6786, p_{CTC} = 0.1071$.

Results of these alternative studies are shown in Figs. 5, 6 and 7. In the uniform response rate cases (Figs. 5, 6), both Fisher's combination method and Stouffer's Z-score method for the adaptive three-stage randomization design produce higher powers than the one-stage randomization design. The larger the sample size and the larger the underlying treatment effect size, the greater the improvement in power. In

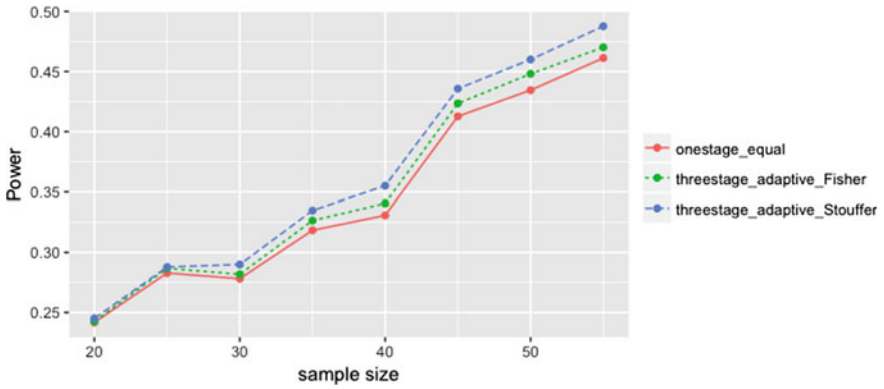


Fig. 5 Alternative hypothesis case 1—Power comparison when response rate is uniformly 40% across stages for treatment and uniformly 20% across stages for control

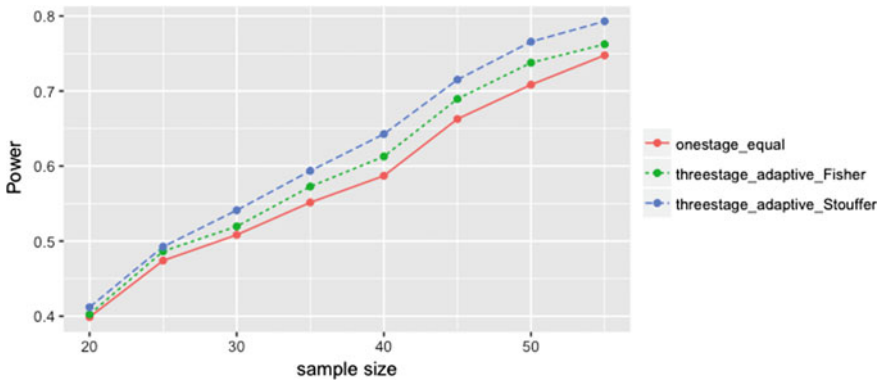


Fig. 6 Alternative hypothesis case 2—Power comparison when response rate is uniformly 50% across stages for treatment and uniformly 20% across stages for control

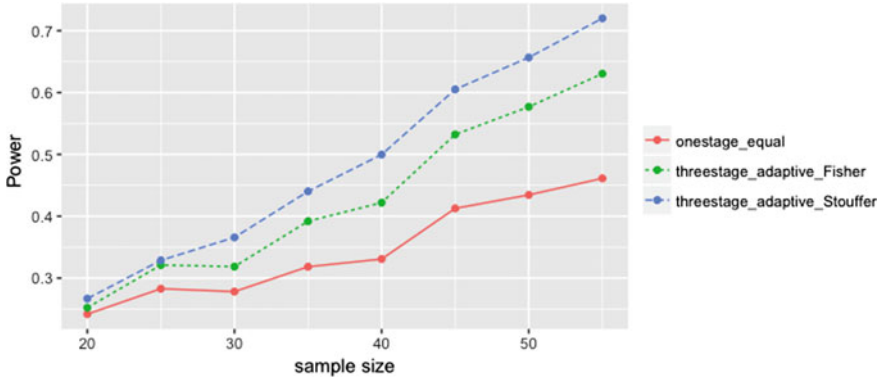


Fig. 7 Alternative hypothesis case 3—Power comparison when response rates are non-uniform across stages for both treatment and control, and come from a mixture model that 20% subjects always respond to treatment and never respond to control, and the remaining 80% subjects have 25% chance equally responding to either treatment or control. The response rates are: Stage 1, $p_T = 0.4$; $p_C = 0.2$; Stage 2, $p_{TT} = 0.625$; $p_{TC} = 0.125$; Stage 2.5, $p_{CT} = 0.4375$; Stage 3 $p_{CTT} = 0.6786$; $p_{CTC} = 0.1071$

the mixed population case (Fig. 7), the relationship between sample size, underlying treatment effect size, and power improvement is even more pronounced. This is in part because the difference in response rates between treatment and control is larger at later stages.

In all of the scenarios consider, the Stouffer's Z-score method generated higher power than the Fisher's combination method. The superiority of the Stouffer's Z-score method over the Fisher's combination method for the design coincides with the claims in Abelson [17]: "the Stouffer's test statistic is sensitive to consistent, even if mild, departures from the null hypothesis, whereas the Fisher procedure is most sensitive to occasional, extreme departures".

4 Discussion

In this paper, we introduced an adaptive three-stage design for small clinical trials with a binary endpoint and demonstrated its superior performance to one-stage equal randomization designs. The use of a Z-pooled unconditional test together with combination methods for three-stage designs controlled the Type I error when the sample size was small and asymptotic statistics were not appropriate. We considered both Fisher's combination method and Stouffer's Z-score method to combine p-values from the three stages and discovered that Stouffer's Z-score method can generate higher powers than Fisher's combination method.

If the study is designed to start with blinding from stage 1, Investigators should be aware that it is not possible to conduct a trial with this adaptive three-stage design in

a completely double-blinded manner as patients, investigators, and statisticians will learn the treatment assignments from stage 1 when the study proceeds to later stages. This unblinding has the potential to introduce bias. Alternatively, stage 1 can be run as an open-label induction phase with blinding only in later stages. In this case, the primary analysis will focus on patients from stages 2 and 3. The overall p-value then is a combination of the p-values from stages 2 and 3.

In this paper, we focused on binary endpoints. Future work will extend the design to non-binary endpoints (e.g., continuous and time-to-event endpoints). We are also considering using a model-based approach or a non-parametric approach to avoid the need to combine p-values. The combination methods do not consider the correlation in response rates between stage 1 and later stages. If the response rate at stage 1 is highly correlated with the response rate at later stages, our calculations may overestimate power. A model-based approach or a non-parametric approach can account for the correlation.

Clinical trials for rare diseases are challenging and the crucial challenges lie in how to reach acceptable statistical power with a small sample size. We believe the proposed adaptive three-stage design provides a decent option in optimizing statistical powers with limited small sample size. Its success through our extensive simulation studies will motivate further interest in this direction of rare-disease trial design.

References

1. European Commission: Useful information on rare diseases from an EU perspective. Accessed 19 May 2009
2. Piantadosi, S.: Cross-over designs. In: Clinical Trials, A Methodologic Perspective. Wiley, Toronto (1997)
3. Guyatt, G.H., Heyting, A., Jaeschke, R., Keller, J., Adachi, J.D., Roberts, R.S.: N of 1 randomized trials for investigating new drugs. Control. Clin. Trials **11**(2), 88–100 (1990)
4. Temple, R.J.: Special study designs: early escape, enrichment, studies in non-responders. Commun. Stat. Theory Methods **23**(2), 499–530 (1994)
5. Temple, R.J.: Problems in interpreting active control equivalence trials. Account. Res. **4**(3–4), 267–275 (1996)
6. Rosenberger, W.: Randomized play-the-winner clinical trials: review and recommendations. Control. Clin. Trials **20**, 328–342 (1999)
7. Cook, J.D.: Error in the normal approximation to the t distribution. https://www.johndcook.com/blog/normal_approx_to_t/. Accessed 18 Jan 2017
8. Rao, C.R.: Linear Statistical Inference and its Applications. Wiley, New York (1965)
9. Honkanen, V.E., Siegel, A.F., Szalai, J.P., Berger, V., Feldman, B.M., Siegel, J.N.: A three-stage clinical trial design for rare disorders. Stat. Med. **20**, 3009–3021 (2001)
10. Cook, J.D., Nadarajah, S.: Stochastic inequality probabilities for adaptively randomized clinical trials. Biom. J. **48**, 356–365 (2006)
11. Wathen, J.K., Cook, J.D.: Power and bias in adaptively randomized clinical trials. Technical Report UTMDABTR-002-06. Accessed 7 Mar 2006
12. Lydersen, S., Fagerland, M.W., Laake, P.: Recommended tests for association in 2 × 2 tables. Stat. Med. **28**, 1159–1175 (2009)

13. Berger, R.L., Boos, D.D.: P values maximized over a confidence set for the nuisance parameter. J. Am. Stat. Assoc. **89**, 1012–1016 (1994)
14. Mehrotra, D.V., Chan, I.S., Berger, R.L.: A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. Biometrics **59**, 441–450 (2003)
15. Bauer, P., Kohne, K.: Evaluation of experiments with adaptive interim analyses. Biometrics **50**, 1029–1041 (1994)
16. Stouffer, S.A., Lumsdaine, A.A., Lumsdaine, M.H., Williams, R.M., Smith, M.B., Janis, I.L., Star, S.A., Cottrell, L.S.: The American soldier: combat and its aftermath. Studies in Social Psychology in World War II., vol. 2. Princeton University Press, Princeton (1949)
17. Abelson, R.P.: Statistics as Principled Argument. Psychology Press, New York (1995)
18. Berger, R.L., Boos, D.D.: P values maximized over a confidence set for the nuisance parameter. J. Am. Stat. Assoc. **89**(427), 1012–1016 (1994)
19. Jeffreys, H.: An invariant form for the prior probability in estimation problems. In: Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences, pp. 453–461 (1946)

Part V
Biomarker-Driven Trial Design

Clinical Trial Designs to Evaluate Predictive Biomarkers: What's Being Estimated?

Gene Pennello and Jingjing Ye

Abstract Predictive biomarkers are used to predict whether a patient is likely to receive benefits from a therapy that outweigh its risks. In practice, a predictive biomarker is measured with a diagnostic assay or test kit. Usually the test has some potential for measuring the biomarker with error. For qualitative tests indicating presence or absence of a biomarker, the probability of misclassification is usually not zero. Study designs to evaluate predictive biomarkers include the biomarker-stratified design, the biomarker-strategy design, the enrichment (or targeted) design, and the discordant risk randomization design. Many authors have reviewed the main strengths and weaknesses of these study designs. However, the estimand being used to evaluate the performance of the predictive biomarker is usually not provided explicitly. In this chapter, we provide explicit formulas for the estimands used in common study designs assuming that the misclassification error of the biomarker test is non-differential to outcome. The estimands are expressed as terms of the biomarker's predictive capacity (differential in treatment effect between biomarker positive and negative patients when the biomarker is never misclassified) and the test's predictive accuracy (e.g., positive and negative predictive values of the test for the biomarker). Upon inspection, the estimands reveal not only well-known strengths and weaknesses of the study designs, but other insights. In particular, for the biomarker-stratified design, the estimand is the product of the biomarker predictive capacity and an attenuation factor between 0 and 1 that increases with the test's predictive accuracy. For other designs, the estimands illuminate important limitations in evaluating the clinical utility of the biomarker test. After presenting the theoretical estimands, we present and discuss estimand values for a hypothetical case study of Procalcitonin (PCT) as a biomarker in Procalcitonin-guided evaluation and management of subjects suspected of lower respiratory tract infection.

G. Pennello
Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD, USA

J. Ye (✉)
Division of Biometrics V, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA
e-mail: jingjing.ye@fda.hhs.gov

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_14

Keywords Estimand · Predictive biomarkers · Clinical trial design · Clinical performance · Attenuation factor

1 Introduction

Predictive biomarkers have become essential for individualizing safe and effective treatments of cancer and other diseases, thereby improving health care through precision medicine. Formally, a predictive biomarker is "used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent" [1].

Predictive biomarkers are typically measured with a commercially developed, in vitro diagnostic device (IVD) or test kit. According to the Food and Drug Administration (FDA), an IVD is a "companion diagnostic" to a specific therapy, if it is "essential for the safe and effective use of [the] therapeutic product" [2]. For a binary-valued test to be considered a companion diagnostic, the treatment effect may need to be positive (beneficial) and large in patients who are test positive, but close to zero or negative (harmful) in patients who are test negative [3].

To evaluate binary-valued predictive marker tests, including companion diagnostics, common Phase III trial designs include the biomarker-stratified design, the biomarker-strategy design, and the enrichment design [4–9]. The biomarker-stratified design has also been called the all-comers or interaction design. The enrichment design has also been called the targeted design. These designs have been reviewed for practical and statistical considerations in many publications, e.g., [10–13]. Alternatively, in a discordant risk randomization design, a new biomarker test is evaluated only in subjects with a test result that is discordant with a standard evaluation [8, 12].

While the literature is generally excellent in listing considerations for choosing a trial design to evaluate the predictive biomarker test as a companion to a treatment, the target estimand being utilized is generally not stated precisely as a mathematical expression. Furthermore, the propensity of the test to misclassify biomarker status is frequently not considered. In this paper, we identify precisely the target estimand of the four trial designs mentioned. To identify the estimand, we assume that test misclassification error is non-differential to the outcome under study [14]. We show that the precise estimand for a particular trial design can reveal not only well-known strengths and weaknesses, but other useful insights.

For the four trial designs mentioned, we identify the target estimand for a mean difference between randomization arms of either a binary or a continuous outcome (e.g., objective response, time-to-event, etc.). For each trial design, we use the estimand to interpret the design's strengths and weaknesses in evaluating the predictive biomarker test for clinical utility. To facilitate this interpretation, we express the estimand in terms of the biomarker's *predictive capacity* (difference in treatment effect between biomarker positive and negative patients when never misclassified) and the test's predictive accuracy (negative and positive predictive values for the

biomarker). To illustrate, we discuss the results of a hypothetical case study of Procalcitonin (PCT) as a biomarker in Procalcitonin-guided evaluation and management of subjects suspected of lower respiratory tract infection.

We consider an estimand to be simply what is being estimated in a statistical analysis. In a clinical trial, the estimand is ideally a direct reflection of the scientific question of interest. For trials of predictive biomarkers, the estimand should link biomarker test results with expected outcomes of treatment decisions. In a recent addendum to the ICH E9 guideline, E9 (R1), a main concern is "absence of a clear relationship between the apparent target estimand and the trial design and analysis in terms of aspects such as patient follow-up after discontinuation of randomised treatment, decisions around which data to exclude from the statistical analysis and handling of missing data" [15]. While aspects of the design, conduct, and analysis of a trial should be consistent with its target estimand, the focus of our paper is not trial implementation, but simply to identify the estimand for the purpose of evaluating the trial design.

2 Clinical Performance of Predictive Biomarkers and Biomarker Tests

Consider a binary biomarker $B = 0$ or 1 whose absence or presence is measured with a binary-valued test $T = 0$ or 1 indicating if a subject is negative or positive for the biomarker. In general, the biomarker test may measure biomarker status with some probability of misclassification. That is, the test result has a non-zero probability of misclassifying the biomarker status. The predictive value of test result $T = t$ for biomarker presence is

$$p_t = \Pr(B = 1|T = t), t = 0, 1$$

That is, $1 - p_0 = \Pr(B = 0|T = 0)$ is the negative predictive value (NPV) of the test and $p_1 = \Pr(B = 1|T = 1)$ is its positive predictive value (PPV). The overall probability of testing positive is $\tau = \Pr(T = 1)$.

Given the biomarker status and test result, the probability of true negative (TN) is $\Pr(\boldsymbol{TN}) = \Pr(\boldsymbol{B = 0, T = 0}) = \Pr(\boldsymbol{B = 0|T = 0})\mathbf{Pr}(\boldsymbol{T = 0}) = (\mathbf{1} - \boldsymbol{p_0})(\mathbf{1} - \boldsymbol{\tau})$. The probability of false positive (FP), false negative (FN) and true positive (TP) can be defined similarly (Table 1).

For the general presentation, we assume $\delta_b > 0$ confers treatment benefit, consistent with many outcomes Y such as binary response status and time to an untoward event. However, for other outcomes, such as duration of hospital stay in the case study presented later, $\delta_b < 0$ confers treatment benefit. The problem can then be fit into the general presentation by considering $\delta_b' = -\delta_b$, with $\delta_b' > 0$ conferring treatment benefit.

The *predictive capacity* of the biomarker is defined as the difference in treatment effects

Table 1 Joint probabilities of biomarker status and biomarker test result

Biomarker status			
Test result	$B = 0$	$B = 1$	Test probability
$T = 0$	$\Pr(TN) = (1-\tau)(1-p_0)$	$\Pr(FN) = (1-\tau)p_0$	$1-\tau$
$T = 1$	$\Pr(FP) = \tau(1-p_1)$	$\Pr(TP) = \tau p_1$	τ

Table 2 Clinical trial mean outcome by treatment assignment and biomarker status

Biomarker status		
Treatment	$B = 0$	$B = 1$
$A = 0$	θ_{00}	θ_{01}
$A = 1$	θ_{10}	θ_{11}
Effect	$\delta_0 = \theta_{10} - \theta_{00}$	$\delta_1 = \theta_{11} - \theta_{01}$

$$\Delta = \delta_1 - \delta_0,$$

i.e., the interaction between biomarker and treatment on outcome. For quantitative interactions, the predictive capacity of the biomarker depends on how the outcome is scaled. For example, if biomarker and treatment have multiplicative effects on Y, then they have additive effects on log Y. Thus the predictive capacity of the biomarker is non-zero on the original scale of Y but zero on the log scale. However, choice of scale cannot eliminate a qualitative interaction, which is often observed for predictive biomarkers especially if, e.g., the biomarker is the molecular target of an effective cancer treatment and is well-measured. A qualitative interaction of a biomarker with a treatment, that is, $\delta_1 > 0 \geq \delta_0$, is generally considered to confer clinical utility (Table 2).

Given treatment $A = a$ and biomarker statust $B = b, a, b = 0, 1$, we denote the mean outcome by $\theta_{ab} = E(Y|A = a, B = b) = E_{ab}(Y)$. The true status of the biomarker is typically not known. We measure it with biomarker test result T. To distinguish mean outcome given true biomarker status from mean outcome by biomarker test value, we use star notation. Given treatment $A = a$ and biomarker test result $T = t$, we denote the expected outcome by $\theta^*_{at} = E(Y|A = a, T = t) = E_{at}(Y), a, t = 0, 1$ (Table 3). The treatment effect given test result $T = t$ is the difference in mean outcome

$$\delta^*_t = \theta^*_{1t} - \theta^*_{0t},$$

$t = 0, 1$. That is, $\delta^*_0 = \theta^*_{10} - \theta^*_{00}$ and $\delta^*_1 = \theta^*_{11} - \theta^*_{01}$ are the treatment effects in biomarker test negative and positive subjects, The difference in treatment effect between test positive and negative subjects (treatment by test interaction) is

Table 3 Clinical trial mean outcome by treatment assignment and test result

Biomarker test result		
Treatment	$T = 0$	$T = 1$
$A = 0$	θ^*_{00}	θ^*_{01}
$A = 1$	θ^*_{10}	θ^*_{11}
Effect	$\delta^*_0 = \theta^*_{10} - \theta^*_{00}$	$\delta^*_1 = \theta^*_{11} - \theta^*_{01}$

$$\Delta^* = \delta^*_1 - \delta^*_0,$$

which links biomarker test results to the outcomes of treatment decisions.

The connection between θ^*_{at} and θ_{ab} can be derived using conditional probability:

$$\theta^*_{at} = E_{at}(Y) = \Sigma^1_{b=0} E_{at}(Y|B = b)\,\Pr(B = b|T = t)$$

This expression simplifies if T is assumed to exhibit *non-differential misclassification error* (NDME) [14]. Under NDME, T is conditionally independent of clinical outcome Y given biomarker status B :

$$T|Y, B = T|B$$

That is, the probability of misclassification of B by T is non-differential to (independent of) clinical outcome Y. Equivalently,

$$Y|T, B = Y|B$$

That is, test result T provides no additional predictive information about outcome Y if the biomarker value B is known, an assumption that is often plausible.

Given NDME, the term $E_{at}(Y|B = b)$ in the above expression reduces to

$$E_{at}(Y|B = b) = E_a(Y|B = b) = E_{ab}(Y) = \theta_{ab}$$

Therefore,

$$\theta^*_{at} = \Sigma^1_{b=0}\theta_{ab}\Pr(B = b|T = t)$$

$$= \theta_{a0} + p_t(\theta_{a1} - \theta_{a0})$$

Thus given test result $T = t$ the treatment effect is

$$\delta^*_t = \theta^*_{1t} - \theta^*_{0t}$$

$$= \delta_0 + p_t(\delta_1 - \delta_0)$$

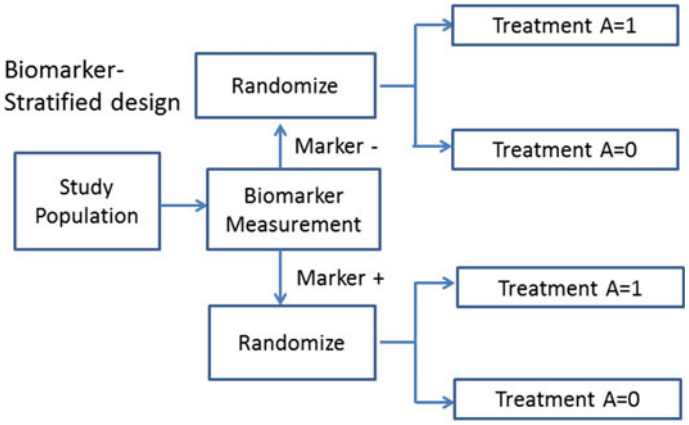


Fig. 1 Biomarker-stratified design

The difference in treatment effect between test positives and negatives is

$$\begin{aligned} \Delta^* &= \delta_1^* - \delta_0^* \\ &= (p_1 - p_0)(\delta_1 - \delta_0) \\ &= (NPV + PPV - 1)\Delta. \end{aligned} \tag{1}$$

That is, Δ^* is the product of the predictive capacity Δ of the biomarker and an attenuation factor $NPV + PPV - 1$, which quantifies the ability of the test to predict the correct biomarker value over and above a random test.

3 Study Designs and Estimands

3.1 Biomarker-Stratified Design

In a biomarker-stratified design, all patients are enrolled regardless of biomarker test value. The enrolled patients are randomized to the treatment or a control. The treatment effect (mean difference between treatment and control) is stratified by biomarker test value. The study design is illustrated in Fig. 1. If the test is available at time of randomization, then randomization may be stratified by test value to avoid imbalance in numbers of subjects assigned to treatment or control within biomarker test subsets.

In the biomarker stratified design, the estimand is Δ^*, the interaction between treatment and biomarker test. From Eq. (1), Δ^* is non-zero if

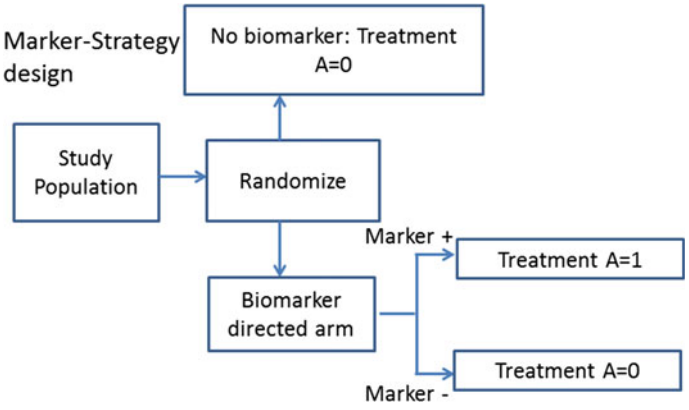


Fig. 2 Biomarker-strategy design

(1) the treatment effect is not homogeneous (e.g., $\delta_1 > \delta_0$), and
(2) the test is informative, i.e., is not a random test for biomarker status ($p_1 > p_0$).

This estimand is ideal in that it factors out diagnostic accuracy, as measured by $p_1 - p_0$, from biomarker capacity, i.e., treatment by biomarker interaction $\delta_1 - \delta_0$. The treatment by test interaction $\delta_1^* - \delta_0^*$ is the treatment by biomarker interaction $\delta_1 - \delta_0$ attenuated by the accuracy of the test $p_1 - p_0$, which is the positive predictive value ($PPV = p_1$) of the test plus its negative predictive value ($NPV = 1 - p_0$) minus 1.

3.2 Biomarker-Strategy Design

In a biomarker strategy design, patients are randomized to either a test strategy arm in which the biomarker test value is utilized to determine treatment received or a control arm in which it is not utilized. In its simplest version, patients assigned to the control arm receive the control (standard) treatment ($A = 0$), while patients in the experimental arm receive the experimental treatment ($A = 1$) or the control treatment ($A = 0$), depending on whether they test positive or negative for the biomarker, respectively. The design is an attempt to evaluate if test-directed treatment selection can improve clinical outcomes. A diagram of the design is shown in Fig. 2.

For the biomarker-strategy design, the estimand is the difference in average outcome between test-strategy and control arms. In the notation given above, the estimand for biomarker-strategy design can be expressed as

$$\Delta_S^* = \tau\theta_{11}^* + (1-\tau)\theta_{00}^* - \left[\tau\theta_{01}^* + (1-\tau)\theta_{00}^*\right] = \tau[\theta_{11}^* - \theta_{01}^*]$$

$$= \tau\delta_1^* = \tau[\delta_0 + p_1(\delta_1 - \delta_0)] = \tau[\delta_0 + p_1\Delta]$$

where recall $\delta_1^* = \delta_0 + p_1(\delta_1 - \delta_0)$ is the treatment effect in patients who test positive. Note that patients who test negative are given the control treatment $A = 0$, regardless of whether they are randomized to the test strategy arm or the control arm. In patients who test negative, the difference in expected outcome between the arms is zero. As a result, the expected difference between the arms is $\Delta_S^* = \tau\delta_1^*$, the treatment effect δ_1^* in test positive subjects diluted by the probability of testing positive τ. The inefficiency of diluting the treatment effect in test positives by τ is a well-known limitation of the biomarker-strategy design [10–13].

Some special cases of the estimand for the biomarker-strategy design are worth noting:

(1) When the treatment effect is homogeneous, $\delta_0 = \delta_1 = \delta$, and the predictive capacity of the biomarker is $\Delta = 0$. In this case, the estimand of the biomarker-strategy design reduces to $\Delta_S^* = \tau\delta$. That is, the estimand is positive even when the biomarker has no predictive capacity but the homogeneous treatment effect is positive. Moreover, the estimand depends not on the diagnostic accuracy of test for biomarker status, but only on the probability of testing positive τ, the factor by which the homogeneous treatment effect is diluted. Thus, the difference between arms is the same for a perfectly accurate test ($p_1 = 1$, $p_0 = 0$) or a random test ($p_0 = p_1$), if the probability of testing positive is the same.

(2) When the treatment only has an effect among biomarker positive subjects, i.e. $\delta_1 > \delta_0 = 0$, the difference between arms is $\Delta_S^* = \tau p_1 \delta_1 = \Pr(TP)\delta_1$, the treatment effect δ_1 among biomarker positive subjects diluted by the probability of a true positive test result, i.e., correct identification of biomarker positive subjects.

(3) In scenario (2), if the probability of a true positive test result is random, then $p_0 = p_1 \equiv p$, where p is the prevalence of being biomarker positive, and $\Delta_S^* = \tau p \delta_1$. In this case, the treatment only has an effect in biomarker positive subjects, the test selects some of these subjects at random, and the result is a positive difference between study arms. This case shows that estimand for the biomarker-strategy arm can be positive, even when the test is merely selecting at random some of the subjects in the subset who benefit from the treatment, another well-known limitation of the design [10–13, 17–21].

3.3 Enrichment Design

If biological rationale and early-phase evidence is strong enough to expect that patients who are biomarker negative are likely to not benefit from the therapy, then an enrichment design is often employed in which only test positive subjects are enrolled into a confirmatory trial of the investigational treatment. Biomarker test positive subjects are randomized to receive either the treatment or a control, but test negative subjects are screened out, not eligible for trial enrollment. The diagram of the study design is shown in Fig. 3.

Because only patients who test positive are enrolled in the trial, the estimand for the enrichment design is simply

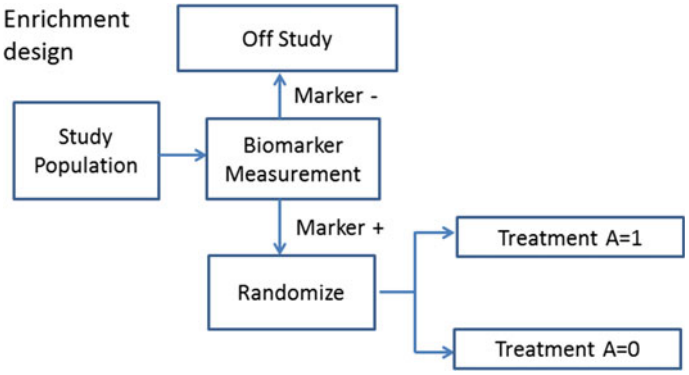


Fig. 3 Enrichment design

$$\Delta_E^* = \delta_1^* = \delta_0 + p_1(\delta_1 - \delta_0)$$

Some special cases in enrichment design are worth noting:

(1) When the treatment effect is homogeneous, $\delta_0 = \delta_1 = \delta$, and the estimand reduces to $\Delta_E^* = \delta$, the treatment effect. As with the biomarker-strategy design, the difference between arms does not depend on the diagnostic accuracy of test for biomarker status. Thus the difference between arms is the same for a perfectly accurate test ($p_1 = 1$, $p_0 = 0$) or a random test ($p_0 = p_1$). However, unlike biomarker-strategy design, the enrichment design is not inefficient in that the homogeneous treatment effect δ is not diluted by τ.

(2) When the treatment only has an effect among biomarker positive subjects, i.e. $\delta_1 > \delta_0 = 0$ the estimand is $\Delta_E^* = p_1\delta_1 = PPV * \delta_1$, the treatment effect δ_1 in biomarker positive subjects diluted by the positive predictive value of the test.

(3) In scenario (2), if the test is random, $p_0 = p_1 \equiv p$, and the estimand reduces to $\Delta^* = p\delta_1$ the treatment effect δ_1 in biomarker positive subjects diluted by their prevalence p.

In sum, for special cases (2) and (3), in which the treatment only has an effect in biomarker positive subjects, the estimand is the treatment effect in biomarker positive subjects is diluted proportionally by the predictive value of test in identifying these subjects, which could be random but still result in a positive difference between study arms.

While the enrichment design can validate that a treatment is effective in a subset of subjects selected by the test (test positives), a well-known drawback of the design is that it obviously can't be used to evaluate if the treatment is effective in the complement of that subset.

Comparing the estimands for the enrichment and biomarker-strategy designs, the biomarker-strategy design can be interpreted as an inefficient version of the

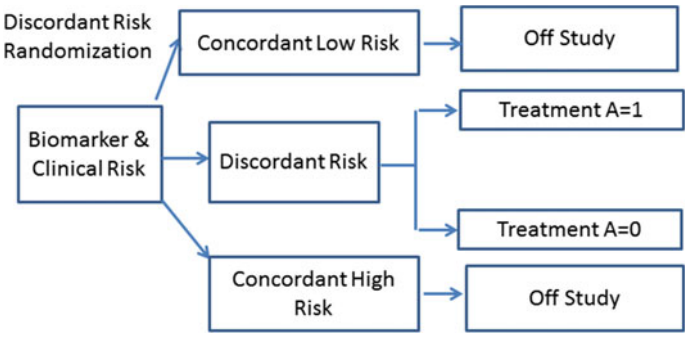


Fig. 4 Discordant risk randomization design

enrichment design in which treatment effect is diluted by biomarker test-positive probability τ.

3.4 Discordant Risk Randomization Design

In a discordant risk randomization design [8], patients are randomized to either investigational treatment or control when there is discordance between biomarker test result and the clinical risk, as evaluated by traditional clinical variables in routine use, which could include one or more standard biomarker assessments. For example, in the case study of the next section, based on her clinical assessment, a clinician may intend to give antibiotic therapy to a subject with suspected lower respiratory tract infection. However, this assessment may be discordant with a low level of the biomarker procalcitonin (PCT), indicating low risk of infection. This subject would be eligible for randomization to receive the therapy or not, in a discordant risk randomization trial design. A version of this design is being implemented in the ongoing trial, Targeted Reduction of Antibiotics using Procalcitonin in a multi-center, randomized, blinded, placebo-controlled, non-inferiority study of azithromycin treatment in outpatient adults with suspect Lower Respiratory Tract Infections (LRTI) and a procalcitonin level of <0.1 ng/mL (TRAP-LRTI, http://arlg.org/studies-in-progress). Another example of discordant risk randomization is given in [12], Fig. 2. A virtue of the discordant risk randomization design is that the predictive biomarker is evaluated only in subjects for whom it may change the treatment decision (Fig. 4).

Biomarker test value $T = 0$ or 1 and standard clinical assessment $S = 0$ or 1 both attempt to classify correctly a latent subset of subjects $B = 0$ or 1 in whom, respectively, the investigational treatment $A = 1$ is less or more effective, relative to standard treatment $A = 0$. In the most inclusive discordant risk randomization design, subjects eligible for enrollment may have a test value that is discordant with

either a positive or negative standard clinical assessment. Let δ^*_{st} be the treatment effect given $S = s$ and $T = t$. An estimand is

$$\Delta^*_D = \delta^*_{01} - \delta^*_{10},$$

the difference in treatment effect between subjects with test results positively discordant with the clinical assessment and those negatively discordant. If the measurement errors in S and T are both non-differential to outcome Y, and we denote $p_{st} = \Pr(B = 1|S = s, T = t)$. Then the estimand can be written as

$$\Delta^*_D = (p_{01} - p_{10})(\delta_1 - \delta_0)$$

Note if T is not associated with B given S, then it adds no incremental value over S in predicting B, and $p_{01} - p_{10} = p_{0.} - p_{1.} < 0$ if S is better than random at predicting B. Thus if B has predictive capacity ($\delta_1 > \delta_0$), then the estimand is positive only if T adds sufficient incremental value over non-random S at predicting B.

In sum, test T would be useful to decide treatment within subsets defined by standard test S if

(1) the effect of experimental treatment is not homogeneous, with B having predictive capacity ($\delta_1 > \delta_0$), and
(2) the predictive value for $B = 1$ is greater if the standard assessment is negative ($S = 0$) but the test is positive ($T = 1$) than if the standard is positive ($S = 1$) but the test is negative ($T = 0$), i.e., $p_{10} > p_{01}$.

4 Case Study

We consider a hypothetical case study for Procalcitonin (PCT) as a biomarker in Procalcitonin-guided evaluation and management of subjects suspected of lower respiratory tract infection. The clinical outcome of interest in this case study is the number of days of hospital stay.

Using the notation in the paper, B is the biomarker status, i.e., the indicator of bacterial infection. Treatment indicator value $A = 0$ or 1 indicates that antibiotic (AB) therapy was or was not initiated, respectively. Test measurement of PCT level (ng/mL) is dichotomized, with $T = 0$ or 1indicating whether or not PCT <= 0.25 ng/ml.

Hypothetical values for the expected number of days of hospital stay stratified by biomarker value, test value, and treatment assignment are summarized in Table 4. For example, for the no antibiotic therapy treatment arm ($A = 1$), the mean length of stay in the hospital is 6 days if bacterial infection is absent ($B = 0$), but 18 days if bacterial infection is present ($B = 1$). Meanwhile for antibiotic therapy treatment arm ($A = 0$), the mean length of stay is 6 days regardless of whether bacterial

Table 4 Hypothetical case study on PCT for management of antibiotic (AB) use, length of hospital stay (Days) by AB use and bacterial infection status

	No Bact (B = 0)	Bact (B = 1)	Test probability
PCT <=0.25 (T = 0)	TN	FN	$1-\tau$
PCT > 0.25 (T = 1)	FP	TP	τ
Prevalence	$1-p$	p	
AB (A = 0)	6	6	
No AB (A = 1)	6	18	

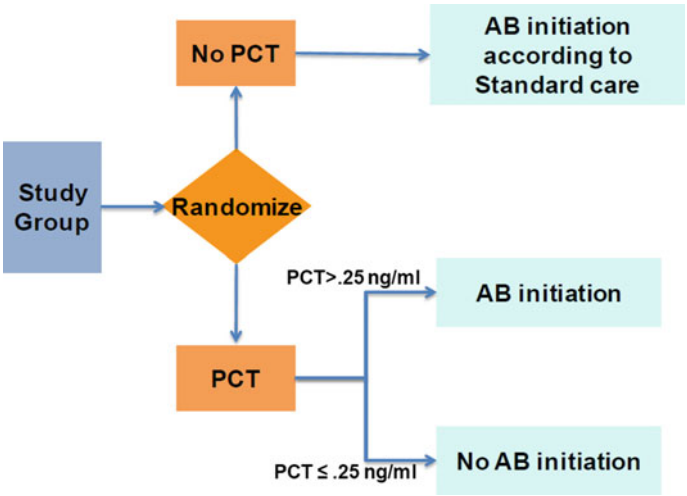


Fig. 5 Biomarker-strategy design for procalcitonin-guided strategy of antibiotic (AB) initiation

infection is absent or present. Thus the effects of not initiating AB therapy on length of hospital stay are $\delta_1 = \theta_{11} - \theta_{01} = 18 - 6 = 12$ and $\delta_0 = \theta_{10} - \theta_{00} = 6 - 6 = 0$ days for subjects with and without bacterial infection. Note that in the notation, a positive treatment effect confers not benefit, but detriment. To translate to effects that confer benefit, in the subsequent discussion we use $\delta'_b = -\delta_b$.

In the biomarker strategy design, subjects are randomized to have the decision to initiate antibiotic therapy guided by PCT level or not (Fig. 5). All subjects in the nonPCT guided group receive, antibiotic (AB) therapy ($A = 0$), the standard of care (SOC). Subjects in the PCT-guided group will receive AB therapy unless PCT $\leq$ 0.25 ng/ml, indicating low risk of bacterial infection.

Because in this biomarker-strategy design (Fig. 5), standard AB therapy is withheld in the test-strategy arm for subjects testing negative ($T = 0$), the estimand is

$$\Delta_S^* = (1-\tau)\delta_0'^* = (1-\tau)\left[\delta'_0 + p_0(\delta'_1 - \delta'_0)\right],$$

a modification of the estimand derived in Sect. 3.2. It is a comparison between nonPCT guided- and PCT guided-therapy groups.

In this case study, suppose the prevalence of bacterial infection is $p = \Pr(B = 1) = 0.1$. We examine the estimand for three different scenarios of test performance: as a random test, a perfect test, or a decent test.

Suppose binary-valued PCT is a random test for predicting bacterial infection, i.e., $p_1 = p_0 = p = 0.1 =$ prevalence, with probability of testing positive $\tau = \tau_0 = \tau_1 = 0.25$, where $\tau_b = \Pr(T = 1|B = b)$ denotes test classification accuracy, i.e., τ_0 is the test's false positive fraction (1 – specificity) and τ_1 is its sensitivity. Then the estimand value is

$$\Delta_S^* = (1-\tau)\delta_0'^* = (1-\tau)\left[\delta_0' + p_0\left(\delta_1' - \delta_0'\right)\right]$$

$$= (1-\tau)p\left(\delta_1' - \delta_0'\right) = 0.75(0.1)(-12-0) = -0.9$$

days. According to this estimand value, the detrimental effect of using PCT to guide decisions to initiate AB therapy or not is just 0.9 days, on average, which could be mis-interpreted to mean that PCT is providing adequate stewardship in determining who should, and who should not, receive AB therapy, despite that it is a random test! In fact, this estimand value is a dilution of the 12 extra days of hospital stay occurring on average when not providing antibiotic (AB) therapy to subjects with bacterial infection.

Alternatively, suppose the PCT test is better than random with decent performance, specifically specificity $= 1 - \tau_0 = 0.75$ and sensitivity $\tau_1 = 0.9$. Because prevalence $p = 0.1$ for bacterial infection, the probability of testing positive is $\tau = 0.315$. By Bayes theorem the predictive values of the test are $NPV = 1 - p_0 = 0.9854$ and $PPV = p_1 = 0.2857$. Therefore, the estimand value is

$$\Delta_S^* = (1-\tau)\delta_0'^* = (1-\tau)\left[\delta_0' + p_0\left(\delta_1' - \delta_0'\right)\right]$$

$$= (1-\tau)p_0\left(\delta_1' - \delta_0'\right) = 0.685(0.0146)(-12-0) = -0.12$$

days. The negative predictive value of the test $NPV = 0.9854$ has mitigated the effect of not providing antibiotic (AB) therapy to subjects with bacterial infection from 12 extra days of hospital stay to 0.1752 days on average. However, this effect is diluted from 0.1752 days to 0.12 days by the probability 0.685 of testing negative, illustrating an inadequacy of the biomarker strategy design.

For a perfect test, $p_0 = 0$ and $p_1 = 1$, and the estimand is $\Delta_S^* = 0$, the best possible result, given the configuration of expected outcomes in Table 4.

For the enrichment design (Fig. 6), the estimand is

$$\Delta_E^* = \delta_0'^*,$$

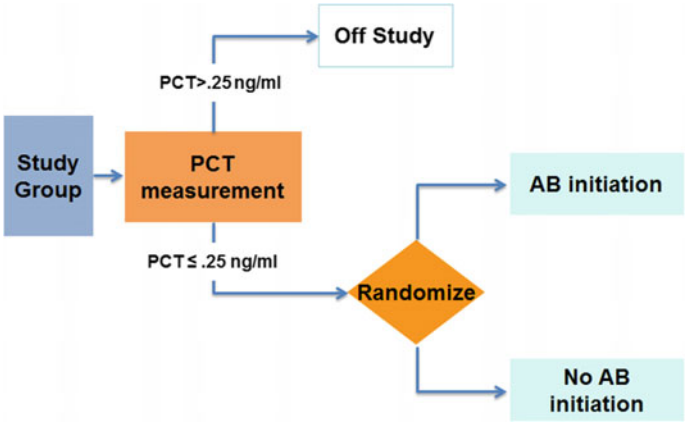


Fig. 6 Enrichment design for procalcitonin-guided strategy of antibiotic initiation

which is equal to −1.2, −0.1752, and 0 days for the random, decent, and perfect tests. Unlike the biomarker-strategy design, the treatment effect is not subject to dilution by the probability of testing negative because the estimand is conditional on enrollment of only test negative subjects into the trial.

For ethical reasons, a biomarker-stratified design should not be conducted, but if it were, the estimand would be:

$$\Delta^* = \delta_1'^* - \delta_0'^*$$

$$= (p_1 - p_0)\left(\delta_1' - \delta_0'\right),$$

which is 0, −3.25, and −12 days for the random, decent and perfect tests. In this case, a large, negative estimand value confers that the biomarker test has clinical utility. For a perfect test, the estimand value is –12 days, the predictive capacity of the biomarker to discriminate treatment effects when biomarker status is never misclassified. The results for all of the designs and tests are summarized in Table 5.

Discordant risk randomization is an attractive option for this example. In a discordant risk randomization design, only those whose PCT test result disagrees with the clinician's assessment on whether AB should be initiated or not are randomized to these treatment options (Table 6, Fig. 7). By comparison, in the biomarker strategy design (Fig. 5), a randomized comparison is made of PCT + SOC guided therapy with SOC on the whole population. In contrast, in the discordant randomization design, randomization is restricted to the subsets of subjects for whom PCT would change the treatment decision (Table 6). The point of randomization has changed from whether to use the test (biomarker strategy), to what to do with the test result (discordant risk randomization), which intuitively can be seen to increase trial efficiency [12].

In the aforementioned TRAP trial (TRAP-LRTI, http://arlg.org/studies-in-progress), subjects suspected of LRTI are being enrolled who (most likely) would

Table 5 Estimand of PCT case study given different biomarker test results*. Biomarker predictive capacity is defined as $\delta_1' - \delta_0'$, the differential in treatment effect between biomarker positive and negative subjects when biomarker status is never misclassified

		Random Test	Decent Test	Perfect Test
		$\tau = 0.25$ $\tau_0 = 0.25$ $\tau_1 = 0.25$	$\tau = 0.315$ $\tau_0 = 0.25$ $\tau_1 = 0.90$	$\tau = 0.10$ $\tau_0 = 0$ $\tau_1 = 1$
Trial Design	**Estimand**		$p_0 = .0146$ $p_1 = .2857$	$p_0 = 0$ $p_1 = 1$
Biomarker Stratified	$\Delta^* = (p_1 - p_0)(\delta_1' - \delta_0')$	0	−3.25	−12
Biomarker Strategy	$\Delta_S^* = (1-\tau)[\delta_0' + p_0(\delta_1' - \delta_0')]$	−0.9	−0.12	0
Enrichment	$\Delta_E^* = \delta_0' + p_0(\delta_1' - \delta_0')$	−1.2	−0.1752	0

*Prevalence of bacterial infection $p = \Pr(B = 1) = 0.1$

Table 6 Discordant Randomization Treatment Decisions

	SOC + PCT	
SOC[a]	**no AB**	**AB**
no AB	**No change**	**Change**
AB	**Change**	**No change**

[a]SOC = standard of care

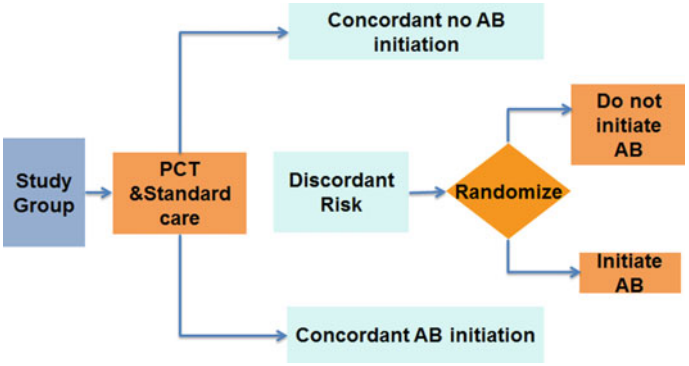


Fig. 7 Discordant risk randomization design for procalcitonin-guided strategy of antibiotics initiation

have received AB therapy but whose PCT value indicated low risk of bacterial infection. Thus, only one of the two types of discordant risk subsets is being enrolled (Table 6, lower left off-diagonal cell.)

For the TRAP design of enrollment of subjects with low PCT value for whom AB therapy would otherwise have been initiated (Fig. 5), the estimand for discordant risk randomization is modified from the formula in Sect. 3.4 to

$$\Delta_D^* = \delta_{10}'^*,$$

the treatment effect in subjects who are standard clinical assessment positive, (would receive AB therapy) but test negative (low risk of bacterial infection according to PCT low value). Under NDME,

$$\Delta_D^* = \delta_0' + p_{10}\left(\delta_1' - \delta_0'\right),$$

where $p_{st} = \Pr(B = 1|S = s, T = t)$ is the predictive value of standard clinical assessment $S = s$ and PCT test result $T = t$, $s, t = 0, 1$. Suppose the false and true positive fractions (1 – specificity and sensitivity) are $\tau_0 = 0.4$ and $\tau_1 = 0.9$ for standard assessment and $\tau_0 = 0.25$ and $\tau_1 = 0.9$ for PCT. Suppose also that the correlation of PCT with standard assessment is 0.5556 for subjects with and 0 for subjects without bacterial infection. Then calculations show that $p_{10} = 0.0146$, and the estimand value is $\Delta_D^* = \delta_0' + p_{10}\left(\delta_1' - \delta_0'\right) = 0 + 0.0146(-12 - 0) = -0.1752$ days. Therefore, using PCT low value ($T = 0$) as the basis for deciding to withhold initiation of AB therapy among those who would have received it according to standard assessment leads to 0.1752 extra days of hospital stay, on average, which can be compared with 12 and 0 extra days on average in subjects with and without bacterial infection. By comparison, if PCT were a random test, $p_{10} = 0.2$, the PPV of standard assessment, and the estimand value is -2.4 days. If PCT were a perfect test, then $p_{10} = 0$, because PCT is never false negative ($NPV = 1$), and the estimand value is 0 days.

5 Discussion and Conclusions

Clinical trials are conducted in efforts to translate trial results to clinical practice. A predictive biomarker test has direct clinical consequences because it is essential for the safe and effective use of a corresponding therapeutic product. A predictive biomarker test for a therapeutic product should be evaluated using an estimand that provides a clear link between test results and the outcomes of treatment decisions.

Four commonly used study designs were discussed in the paper, namely, the biomarker-stratified, biomarker-strategy, enrichment and discordant risk randomization designs. Even though strengths and limitations of the designs have been described by many authors, among the papers of which we are aware, the estimand being used

in each design to evaluate performance of the predictive biomarker is not given precisely. We developed a general framework for deriving the explicit estimand of each design. These estimands permitted us to draw connections between the designs in greater detail than perhaps other publications.

In the biomarker-stratified design, the predictive capacity of biomarker to differentiate treatment effects is attenuated by the factor $PPV + NPV - 1$, which is intuitively appealing, being equal to the sum of the two predictive values of the test minus 1. We reiterate that our general framework is predicated on the assumption of NDME (non-differential misclassification error), i.e., that the biomarker test result is conditionally independent of clinical outcome, given the true biomarker status. This assumption could be relaxed by assuming that NDME holds conditional on covariates that have effects on outcome.

Our derived estimands for the study designs hold true for clinical outcomes of either binary or continuous endpoints in which the treatment effect is defined as the difference in mean outcome between the treatments. For a time-to-event endpoint, the treatment effect could be defined, not as a mean difference, but by the hazard ratio, which requires a different form of estimand, not discussed here. Derivation of this estimand could be the subject of future research.

Identifying the estimand is important because it defines what is to be estimated to address the scientific question of interest. As discussed in ICH E11, an estimand is often confused with estimator or estimate. An estimator is a method for obtaining an estimate, a likely value of the estimand given the sample data. Once the estimand has been identified, it may suggest an efficient estimator. Appropriate estimators for estimands used in trial designs involving predictive biomarkers can be a focus of future research. Sensitivity of an estimate to such factors as treatment non-compliance, missing data, modeling assumptions, (e.g., NDME), etc., can generally be used to assess robustness of study results. In this paper, we ignored aspects of trial conduct or analysis assumptions that could have introduced bias into an estimate for a study.

Besides the four commonly used study designs discussed in the paper, Eng [22] proposed a design called reverse biomarker- based strategy design. In this design, patients are randomly assigned to one of the two treatment strategies. In the first arm, biomarker-positive patients receive the experimental treatment, whereas biomarker-negative patients are allocated to receive the control. By contrast, in the second arm, biomarker-positive patients receive the control and biomarker-negative patients receive the treatment [22]. As Ondra et al. [23] point out, the reverse biomarker-based design cannot address the question of whether a treatment strategy that does not require the determination of the biomarker status would be superior to the biomarker-guided treatment strategy. This design has not been commonly used in practice for evaluation of predictive biomarkers. Therefore, evaluation of its estimand is not provided here. However, in our general framework, the estimand for this design could be derived.

We discussed commonly used study designs based on randomized clinical trials. However, a, single-arm trial based on biomarker status is sometimes conducted. For example, a trial could be conducted in which patients who are biomarker-positive are enrolled and only treated with the active treatment. Rather than the treatment

difference defined based on δ_0^* or δ_1^*, the estimand in this trial can be derived based on θ_{at}^*, a, t = 0, 1. In the case of patients only treated when they are biomarker-positive, the estimand for the single-arm trial would be $\theta_{11}^* = \theta_{10} + p_1(\theta_{11} - \theta_{10})$, which reduces to θ_{11} for a perfect test.

We restricted the biomarkers considered in this paper to those with binary values, e.g. positive or negative. However, the attenuation factor we introduced can be extended to biomarkers having more than two levels. For example, for a gene with two different alleles A and B, the biomarker value may be one of the three genotypes: homozygous in A, homozygous in B, or heterozygous. For a biomarker with more than two ordered categories, misclassification error attenuates the difference in treatment effects between the two most extreme categories, but may not attenuate the difference between other pairs of categories (Kuha and Skinner [24]).

Acknowledgements The authors gratefully thank Drs. Qin Li from and Thomas Gwise from the Food and Drug Administration for the helpful discussions and the reviewers of our draft manuscript for their comments and suggestions.

References

1. US FDA-NIH Biomarker Working Group: BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US) (2016)
2. Us, F.D.A.: Principles for codevelopment of an in vitro companion diagnostic device with a therapeutic product. Silver Spring MD, US FDA (2016)
3. US FDA: In vitro companion diagnostic devices; US FDA: Silver Spring, 2014. http://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm262327.pdf. Accessed March 2017
4. US FDA: Guidance on enrichment strategies for clinical trials to support approval of human drugs and biological products. US FDA: Silver Spring, MD, 2012. US FDA. In Vitro Companion Diagnostic Devices, US FDA: Silver Spring MD (2014)
5. Us, F.D.A.: Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product. Silver Spring MD, US FDA (2016)
6. Beaver, J.A., Tzou, A., Blumenthal, G.M., McKee, A.E., Kim, G., Pazdur, R., Philip, R.: An FDA perspective on the regulatory implications of complex signatures to predict response to targeted therapies. Clin. Cancer Res. **23**(6), 1368–1372 (2017)
7. Polley, M.C., Freidlin, B., Korn, E.L., Conley, B.A., Abrams, J.S., McShane, L.M.: Statistical and practical considerations for clinical evaluation of predictive biomarkers. J. Natl. Cancer Inst. **105**, 1677–1683 (2013)
8. Buyse, M., Michiels, S., Sargent, D.J., Grothey, A., Matheson, A., de Gramont, A.: Integrating biomarkers in clinical trials. Expert Rev. Mol. Diagn. **11**(2), 171–182 (2011)
9. Baker, S.G., Kramer, B.S., Sargent, D.J., Bonetti, M.: Biomarkers, subgroup evaluation, and clinical trial design. Discov. Med. **13**(70), 187–192 (2012)
10. Freidlin, B., McShane, L.M., Korn, E.L.: Randomized clinical trials with biomarkers: design issues. J. Natl. Cancer I. **102**(3), 152–160 (2010)
11. Simon, R.: Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Personalized Med. **7**(1), 33–47 (2010)
12. Bossuyt, P.M., Lijmer, J.G., Mol, B.W.: Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet **356**(9244), 1844–1847 (2000)
13. Mandrekar, S.J., Sargent, D.J.: Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J. Clin. Oncol. **27**(24), 4027–4034 (2009)

14. Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: Measurement Error in Nonlinear Models: A Modern Perspective. Chapman Hall/CRC, Boca Raton, FL (2006)
15. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Final concept paper: E9(R1): Addendum to statistical principles for clinical trials on choosing appropriate estimands and defining sensitivity analyses in clinical trials. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9__R1__Final_Concept_Paper_October_23_2014.pdf. Accessed 1 Aug 2017
16. Simon, R.: Stratification and partial ascertainment of biomarker value in biomarker driven clinical trials. J. Biopharm. Stat. **24**(5), 1011–1021 (2014)
17. Buyse, M., Michiels, S.: Omics-based clinical trial designs. Curr. Opin. Oncol. **25**(3), 289–295 (2013)
18. Simon, R., Maitournam, A.: Evaluating the efficiency of targeted designs for randomized clinical trials. Clin. Cancer Res. **10**(20), 6759–6763 (2004)
19. Pennello, G.A.: Analytical and clinical evaluation of biomarkers assays: when are biomarkers ready for prime time? Clin Trials. **10**(5), 666–676 (2013)
20. Pennello, G.A., Ye, J.: Companion diagnostics. In: Chow, S.-C (ed.) Encyclopedia of Biopharmaceutical Statistics 3rd edn. CRC Press (2017). https://doi.org/10.1081/e-ebs3-140000151
21. Sharma, A., Zhang, G., Aslam, S., Yu, K., Chee, M., Palma, J.F.: Novel approach for clinical validation of the cobas kras mutation test in advanced colorectal cancer. Mol. Diagn. Ther. **20**(3), 231–240 (2016)
22. Eng, K.H.: Randomized reverse marker strategy design for prospective biomarker validation. Stat. Med. **33**, 3089–3099 (2014)
23. Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., Posch, M.: Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. J. Biopharm. Stat. **26**(1), 99–119 (2016)
24. Kuha, J., Skinner, C., Palmgren, J.: Misclassification error. In: Armitage, P., Colton, T (eds.) Encyclopedia of Biostatistics. Wiley (2005)

Biomarker Enrichment Design Considerations in Oncology Single Arm Studies

Hong Tian and Kevin Liu

Abstract Oncology drug development has been increasingly shaped by molecularly targeted agents (MTAs), which often demonstrate differential effectiveness driven by the biomarker expression levels on tumors. Innovative statistical designs have been proposed to tackle this challenge, e.g., Freidlin et al. [3, 4], Jiang et al. [7]. All of these are essentially adaptive confirmatory Phase 3 designs that combine the testing of treatment effectiveness in the overall population with an alternative pathway for a more restrictive efficacy claim in a sensitive subpopulation. We believe that, in cases that there are strong biological rationales to support that a MTA may provide differential benefit in a general patient population; proof-of-concept (POC) is likely intertwined with predictive enrichment. Therefore, it is imperative that early phase POC studies be designed to specifically address biomarker-related questions to improve the efficiency of development. In this paper, we propose three strategies for detecting efficacy signals in single-arm studies that allow claiming statistical significance either in the overall population or in a biomarker enriched subpopulation. None of the three methods requires pre-specification of biomarker thresholds, but still maintains statistical rigor in the presence of multiplicity. The performance of these proposed methods are evaluated with simulation studies.

Keywords Biomarker thresholds · Enrichment design · Proof of concept · Single arm · Binary outcome

1 Introduction

In the past two decade, the landscape of oncology drug development has witnessed a dramatic shift to molecularly targeted agents, which attack cancer cells with more specificity, from traditional cytotoxic chemotherapeutic drugs. These novel agents tend to demonstrate differential effectiveness driven by heterogeneous expression

H. Tian (✉) · K. Liu
Janssen Research & Development, 920 Route 202, Raritan, NJ, USA
e-mail: htian@its.jnj.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_15

profile of their molecular targets on tumors. Among all therapies which were approved by FDA between July 2013 to December 2015 for non-small cell lung cancer, 8 out of 12 were approved in biomarker enriched subpopulations using companion diagnostic devices to identify patients who are more likely to benefit from the therapy, such as Xalkori (approved in 2013), Keytruda (approved in 2015), and Tagrisso (approved in 2015).

It is imperative for researchers to properly design clinical experiments by providing flexibility to identify and confirm efficacy signal in either the overall patient population or a biomarker defined subgroup. Innovative statistical designs have been proposed to tackle this new challenge, e.g., Freidlin et al. [3, 4], Jiang et al. [7]. All of these are essentially adaptive confirmatory Phase 3 designs that combine the testing of treatment effectiveness in the overall population with an alternative pathway to identify a sensitive subpopulation for label claim. We argue that, in cases that there are strong biological rationales and preclinical evidence to support that a MTA may provide differential benefit in a general patient population, early phase POC studies can be designed to specifically address biomarker-related questions, e.g., subgroup selection, biomarker threshold evaluation, in order to improve the efficiency of development. For example, a companion diagnostic may have to be developed for a more restrictive label claim, and this should not wait until the completion of confirmatory Phase 3 trials. Due to the nature of oncology drug development, POC is often studied in relapsed and/or refractory patient populations on the ground that spontaneous tumor regression is rare, therefore leading to single-arm POC study designs including only the experimental therapy. None of the aforementioned designs are specifically designed for single arm studies. Freidlin et al. [3, 4] used modeling of the treatment by subgroup interactions for subgroup identification, which requires a competitor arm. The statistical inference for Jiang et al. [7] relies on the permutation tests to evaluate the strength of statistical evidence, which again only applies to randomized settings.

In this paper, we will present our proposed designs. In Sect. 2, we provide motivations for our research and a brief review of current methods proposed for Phase 3 confirmatory settings with a competitor arm; In Sect. 3, we propose three new approaches, namely, single-arm adaptive signature design (ASD), cross-validated version of single-arm adaptive signature design (CV-ASD) and single-arm biomarker-adaptive threshold (BAT) design. In Sect. 4, we present our simulation studies to evaluate the performance of our proposed methods in comparison to the strategy for testing the overall population only. We will conclude the paper with a brief discussion including directions for future work.

2 Motivations

For novel oncologic agents that target specific molecular pathways, it is likely to benefit only a subset of patient population based on certain predictive biomarkers. For example, the molecular targets may only express in a subset of tumor cells, often

with varying intensity. After a recommend Phase 2 dose (RP2D) is established in a dose escalation study in which toxicity is typically the outcome of primary interest, a Phase 2 proof-of-concept (POC) study, often in a single-arm setting, provides first meaningful opportunity to explore anti-tumor activity and its association with biomarker expression levels. A typical occurrence is to treat an unselected patient population with the experimental therapy and evaluate the objective response rate (ORR) when the data is sufficiently mature. If the ORR is sufficiently high, all is good; otherwise, exploratory analyses will be performed to evaluate ORR in subgroups defined by biomarker levels. In case that a promising subgroup is identified, the finding should be considered as part of hypothesis-generating, which need be confirmed by another study in order for POC to be established. Obviously this is not an efficient way for designing POC studies. An alternative is to conduct the POC study in a more restrictive population based on the best guess about what the subset should be. However, given that it is not rare to have a targeted therapy to show effectiveness in a broad population, this approach can lead to missed opportunities. Furthermore, with available accelerated approval pathways (e.g., accelerated approval in the U.S. and conditional approval in EU) for unmet medical needs in serious or life-threatening diseases such as refractory cancer, it is even possible for transformational therapies to gain marketing authorizations based on data from single-arm studies using objective response rate (or some other endpoint that can be reliably assessed in uncontrolled studies) as the primary endpoint. Therefore, it is both scientifically important and strategically imperative to evaluate the efficacy of a drug with statistical rigor using pre-planned analysis in POC studies.

While researchers have provided some useful tools to address these issues in a randomized Phase 3 setting, there is still a void for specific and reasonably rigorous designs that can be used to evaluate POC via a single-arm Phase 2 study simultaneously in an overall population and in an enriched subpopulation driven by the data.

2.1 A Literature Review of Current Methods

In case that a well-defined subpopulation of interest has already been established prior to the initiation of a confirmatory study, a variety of methodologies have been proposed to test treatment efficacy in the overall population as well as in the subpopulation while maintaining the overall alpha at the 5% (two-sided) level. This is commonly done by Bonferroni adjustment, which allocates a small portion of the overall alpha (say, 1%) for the subpopulation while reserving most of the alpha for the overall population, or some other possibly adaptive methods (e.g., [1, 2, 9] that take advantage of the correlation structure of the test statistics between the overall population and the pre-specified subpopulations, and/or use alpha recycling methodologies. However, in practice, these approaches often face challenges, such as how to precisely define the subpopulation of interest before the trial starts. This issue becomes especially pronounced in the settings where the underlying biomarker level

is continuous but an optimal cutoff is yet to be selected, or in the cases where multiple proteomic/genomic signatures are possibly associated with the mechanism of action, which in turn may impact the treatment efficacy.

Freidlin et al. [3, 4] proposed an adaptive signature design (ASD) and later a cross-validated version of the adaptive signature design for randomized Phase 3 confirmatory trials. Both designs aim to build an alternative pathway, in addition to the testing in the overall population, which enables the identification of a subpopulation based on genomic/proteomic signatures and testing of efficacy in this population, while still controlling the overall Type I error rate. The general idea is to split the study population in two complementary cohorts: (1) a development cohort, which is used to build the statistical model to identify potential biomarker(s) by examining potential treatment and biomarker interactions; (2) a validation cohort, in which the biomarker signature developed using the development cohort is employed to define the subpopulation. The treatment effect can then be tested either in the overall population including all subjects from both development and validation cohorts, or in the biomarker-tailored subpopulation within the validation cohort. Freidlin et al. [3] concluded that the ASD controls the overall Type I error rate since the method for identifying subpopulation is established using data external to the validation cohort. Freidlin et al. [4] further improved the original ASD by incorporating a cross-validation component. This enhanced the efficiency of the original design by allowing every subject enrolled to be used for the signature development as well as for the signature validation, but also introduced a drawback in that the biomarker signatures employed to classify biomarker sensitive status may lack consistency due to the difference in the subset of patients used to develop the signatures. It is noted that Freidlin et al. [3, 4] focused on qualitative biomarkers and the work is developed for large randomized studies with a comparator arm, and therefore cannot be applied to single-arm studies, in which all subjects receive the experimental therapy. In addition, both variants used Bonferroni adjustment to deal with multiplicity in efficacy testing that occurs in both the overall population and the data-dependent biomarker-sensitive subpopulation, which is straightforward to implement but may not have the optimal operating characteristics.

Jiang et al. [7] proposed biomarker-adaptive threshold (BAT) design to identify biomarker tailored subpopulation. The assumptions are that a primary, quantitative biomarker is clearly established based on MOA, and can be reliably quantified with a validated assay. The BAT design is developed to simultaneously achieve two objectives: (1) to establish treatment effectiveness either in the overall population or in an enriched subpopulation based on pre-specified primary biomarker; and (2) to estimate the biomarker threshold if the experimental treatment is efficacious only in the enriched subpopulation. The method (Procedure B in their article) combines the test of treatment effect in the overall population along with those tests in nested sets of subpopulations based on a finite set of candidate biomarker threshold values, which is done via a maximally selected chi-square test. The authors proposed to add a constant to the test statistic for the overall population in order to ensure a reasonable power when the experimental treatment is effective for the overall population. Since the standard asymptotic theory doesn't apply, the authors proposed

to use the permutation test for the hypothesis testing. On a side note, given that their focus was on the setting with a time-to-event endpoint, we caution on the use of permutation tests as it may inflate the Type I error rate since the null would be the equality of both the survival distribution for the event of interest and the censoring distribution. Again, their method was proposed in a randomized phase 3 setting with a comparator arm, and we would like to focus on the setting for single-arm studies.

2.2 Potential Application in Single Arm Studies

All the methods mentioned in Sect. 2.1 are limited to studies with a comparator arm. As stated earlier, POC for cytotoxic oncology drugs are commonly evaluated in single-arm studies, especially in relapsed/refractory settings in which spontaneous tumor regression is rare if not completely impossible. As such, in the following section, we will discuss our proposals to extend these ideas to single-arm POC studies. It should be noted that we implicitly assume that the quantitative biomarker of interest may be predictive of antitumor efficacy, but not prognostic. In other words, any difference in efficacy at different biomarker levels is attributed to differential effectiveness rather than the prognostic value of the biomarker. In case that the biomarker of interest is also prognostic, a randomized study should be used, or alternatively, a Bayesian design may be used if the prognostic value of the biomarker is well-understood.

3 Methods

The most widely used endpoint in oncology POC studies is objective tumor response, which enjoys a rare advantage that it may be reliably evaluated in single-arm studies. Denoting p as the objective response rate for the experimental treatment, the main statistical inference is then about the testing of the null hypothesis (H_0) that $p \leq p_0$ against an alternative hypothesis (H_a) that $p \geq p_a$. Here, p_0 represents a response rate that is of no interest for further development and p_a is a promising response rate which may warrant future development or even likely to predict clinical benefit. The study is considered as positive if the null hypothesis can be rejected in the overall study population or a biomarker enriched subpopulation to be identified in the same study. We assume that there is a strong scientific rationale to support that a biomarker may be predictive of clinical activity, and that an assay has been developed to reliably measure the biomarker of interest. Without loss of generality, we also assume that higher biomarker values is associated with more active antitumor activity.

3.1 Single-Arm Adaptive Signature Design (ASD)

For a quantitatively measured biomarker Z ($Z \geq 0$), we can specify a set of cut points $\{C_1,\ldots, C_J\}$ as candidate threshold values. We then divide the enrolled subjects into two non-overlapping cohorts: a development cohort and a validation cohort (i.e., a training set and a validation set). The separation of two cohorts can be done by a pre-specified randomization scheme according to a prefixed ratio based on the size of the development and validation sets (e.g., 2–3). As in Freidlin et al. [3, 4], we also employ an alpha splitting strategy for multiplicity (e.g., assign 80% of the total α to the testing in the overall population and the remaining 20% α to the biomarker-enriched subpopulation).

Step 1: Test H_0: $p \leq p_0$ in the overall population using its allocated α. If the null hypothesis is rejected, then stop and claim satisfactory antitumor activity for the overall population, otherwise proceed to the next step.

Step 2: Select the biomarker threshold from the set of candidate values using the development cohort. Since a biomarker threshold is assumed predictive, meaning that the objective response rate differs between those subjects whose biomarker values are greater or below the threshold, we can then calculate the likelihood given the biomarker threshold value C_j.

Assume there are a total of n subjects in the development cohort. For the subject i, let Y_i be the binary response ($Y_i = 1$, responder; $Y_i = 0$, non-responder) and Z_i be his/her biomarker expression level (continuous, without loss of generality assuming between 0 and 1, which can be achieved by empirical cumulative distribution transformation). For each candidate threshold value C_j, we can separate the patient population into two subpopulations. Let n_1 denote the number of subjects with biomarker level below C_j, and $n_2 = n - n_1$ for those at or above C_j; also let k_1 denote the total number of responders among those n_1 subjects below the threshold, and k_2 the total number of responders among the n_2 subjects at or above the threshold. In other words, $n_1 = \sum_{i=1}^{n} I(Z_i < C_j)$, $k_1 = \sum_{i=1}^{n} I\left(Z_i < C_j and Y_i = 1\right)$, $n_2 = \sum_{i=1}^{n} I(Z_i \geq C_j)$ and $k_2 = \sum_{i=1}^{n} I\left(Z_i \geq C_j and Y_i = 1\right)$. Let θ_1 be the objective response rate below the threshold and θ_2 be the objective response rate at or above the threshold. The likelihood function, given the threshold C_j, can then be written as,

$$\binom{n_1}{k_1}\theta_1^{k_1}(1-\theta_1)^{(n_1-k_1)}\binom{n_2}{k_2}\theta_2^{k_2}(1-\theta_2)^{(n_2-k_2)}$$

With the likelihood values computed for all the candidate threshold values, the optimal threshold can then be identified by selecting the one corresponding to the maximum likelihood value, which can be used to define the biomarker-enriched population in the validation cohort.

Step 3: Test H_0 in the biomarker enriched population using the threshold identified in Step 2. The population in which the test is to be performed will be the subjects in the validation cohort with biomarker value greater than or equal to the threshold found in Step 2. The allocated type I error rate will be the remaining α.

The POC study is considered as successful if the hypothesis can be rejected either in the overall population or in the biomarker enriched subpopulation.

3.2 Cross Validated Single-Arm Adaptive Signature Design (CV-ASD)

The idea of cross-validation can be similarly adopted as in Freidlin et al. [4]. The only modification to the single arm adaptive threshold design in Sect. 3.1 is limited to STEP2. A K-way cross validation can be utilized. First randomly divide the entire cohort into K approximately equal sized subpopulations. For each one of the subpopulation, use the complimentary set including $(K - 1)$ subpopulations to estimate the biomarker threshold as discussed in Sect. 3.1 STEP2; use the estimated threshold to identify "sensitive patients" for each subpopulation. Pool all the "sensitive patients" together and perform hypothesis testing as stated in STEP3. Please notice the estimated threshold used to identify sensitive group for each subpopulation can be different. A summary measure (e.g., median or mean) may be used as estimate for the threshold.

3.3 Single Arm Biomarker-Adaptive Threshold (BAT) Design

For a quantitatively measured biomarker Z ($Z \geq 0$), we again assume that a set of candidate threshold values $\{C_0, C_1, \ldots, C_J\}$, with $C_0 = 0$ (corresponding to the overall population). For each cutpoint C_j ($j = 1, \ldots, J$), a log likelihood ratio statistics S_i can then be constructed against the null hypothesis H_0: $P \leq P_0$ using subjects with biomarker value at or above this threshold. (An alternative is to obtain the tail probability using binomial tabulation, and then transform it into a quantile of either a standard normal or a chi-square distribution). Similar to Jiang et al. [7], we prespecify a constant R that will be used to weight up the test statistic for the overall population, S_0.

The selection of the threshold is then based on the maximum of the test statistics

$$T = \max\{S_0 + R, S_1, \ldots S_J\}$$

The selected threshold value will be the one corresponding to the component S_i statistic that gives T the maximum value. The response rate p for the selected

population can be estimated using those subjects whose biomarker values are at or above the selected threshold.

The hypothesis testing using T is not straightforward. Here we propose a resampling approach. Incorporating the two guidelines for bootstrap hypothesis testing according to Hall et al. [6], the bootstrap hypothesis testing can be carried out using $\hat{p}$, which is the observed response rate in the population with biomarker value greater or equal to the selected threshold. Denoting $\hat{p}^*$ as the value obtained following the above maximum T-based procedure using a bootstrap sample, and $\hat{\sigma}^*$ as the bootstrap standard deviation. The test is then based on the bootstrap distribution $\left(\hat{p}^* - \hat{p}\right)/\hat{\sigma}^*$ and the critical value $\widehat{cr}$ can be found by

$$\Pr\left(\frac{\hat{p}^* - \hat{p}}{\hat{\sigma}^*} > \widehat{cr}\right) > \alpha$$

The statistical significance can be claimed when $\frac{\hat{p}-p_0}{\hat{\sigma}} > \widehat{cr}$, where $\hat{\sigma}$ is the standard deviation of $\hat{p}$.

In line with the recommendation by Jiang et al. [7], a possible choice of R can be 2.2, which is equal to the difference between the 95-th and 80-th percentiles from the chi-squared distribution with 1 degree of freedom.

4 Simulation

4.1 Simulation Setup

We consider the setting in which biomarker values are obtained by an immunohistochemistry (IHC) staining assay on tumor tissues, and the potential outcomes are 0, 1+, 2+, 3+ according to the intensity of staining.

The outcome of interest is overall response rate. The response rate and biomarker value follows a monotonically non-decreasing relationship.

The prevalence of a biomarker positive population is the percentage of subjects whose biomarker values are at or above a threshold.

In our simulation, we assume the prevalence of 0, 1+ , 2+ and 3+ in the population is 20%, 30%, 30% and 20% respectively, and the true response rates for the positive and negative populations may differ. Therefore, the prevalence of biomarker positive population can be 100, 80, 50 or 20% depending on where the true biomarker threshold is.

We analyzed each simulation dataset using four methods testing H_0: $p < 20\%$ versus H_a: $p > 35\%$ at a total alpha of 5%. The total number of subjects is 80, which can provide 90% power under $H_{a,}$ assuming a homogenous population (i.e., biomarker threshold equals to 0) using the exact binominal test for the overall population.

(1) Overall Test: Test the overall population at the 5% alpha level; if successful, claim efficacy signal in the overall population (i.e., biomarker threshold equals to 0).
(2) ASD: Allocate an alpha of 4% to the overall population and using the remaining 1% for the biomarker positive population to be identified using the method specified in Sect. 3.2. Forty percent of the randomly selected population is assigned to the development cohort (i.e., the training set) and the rest is in the validation cohort (i.e., the test set).
(3) CV-ASD: Alpha allocation stays the same as the ASD and 4 fold cross validation is added to evaluate potential performance.
(4) BAT: Set $R = 2.2$ as the constant to weight up the likelihood ratio test for the overall population. The number of bootstrap sampling runs is set at 10,000 times.

To evaluate the performance of these methods, we used the following operating characteristics measures:

(1) Probability of claiming statistical significance in the overall population or in any subpopulation;
(2) Probability of identifying the correct threshold.

Simulation Results

Table 1 presents the simulation results. The left panel presents the simulation setup for 18 scenarios with varying threshold values, true response rates for the biomarker negative and the biomarker positive populations, and the prevalence of the biomarker positive population. For example, Scenario 18 is for the setting in which the true response rate is 40% for the 3+ group, and 20% for the rest of the population, the true threshold value is 3+, and the prevalence of the biomarker positive population (3+) is 20% of the overall population.

The right panel reports the simulation results in terms of probability of claiming success using each of the four methods, and the probability of identifying the correct threshold using these methods.

In Scenarios 1–3, the biomarker is assumed to have no predictive value, i.e., the biomarker threshold is zero. In this sense, Scenario 1 can be considered as the setting in which the null hypothesis is true, where the true response rate is 20% for all subjects regardless of what biomarker values they have. On the other hand, Scenarios 2 and 3 are intended to evaluate the operating characteristics of the proposed methods in the situation where the experimental therapy is active regardless of the biomarker value. In other words, we can assess potential tradeoff by building in an alternative pathway for an enriched subpopulation.

In Scenarios 4–12, the response rate for the biomarker negative population is assumed to be zero; while in Scenarios 13–19, the response rate for the biomarker negative population is assumed to be 20% (the same as in the null hypothesis).

All methods appear to have satisfactory Type I error rate control as intended according to the simulation results in Scenario 1.

Table 1 Simulation setup and results

Scenario setup (Truth)					Simulation results						
#	Response rate		Prevalence of biomarker + (%)	Biomarker threshold	Probability of claiming stat significance in any subpopulation or overall				Probability of identifying the correct threshold		
	Biomarker +	Biomarker −			Overall test	ASD	CV-ASD	BAT	ASD	CV-ASD	BAT
1	0.2	0	100	0	0.04	0.04	0.04	0.06	0.04	0.04	0.05
2	0.3	0	100	0	0.65	0.65	0.65	0.57	0.65	0.65	0.55
3	0.4	0	100	0	0.99	0.99	0.99	0.96	0.99	0.99	0.93
4	0.2	0	80	1	0.00	0.00	0.00	0.02	0.00	0.00	0.00
5	0.3	0	80	1	0.19	0.21	0.19	0.31	0.01	0.00	0.19
6	0.4	0	80	1	0.78	0.79	0.78	0.85	0.00	0.00	0.70
7	0.2	0	50	2	0.00	0.00	0.00	0.02	0.00	0.00	0.01
8	0.3	0	50	2	0.00	0.01	0.00	0.27	0.00	0.00	0.23
9	0.4	0	50	2	0.04	0.07	0.07	0.75	0.01	0.00	0.69
10	0.2	0	20	3	0.00	0.00	0.00	0.04	0.00	0.00	0.00
11	0.3	0	20	3	0.00	0.00	0.00	0.03	0.00	0.00	0.02
12	0.4	0	20	3	0.00	0.00	0.00	0.14	0.00	0.00	0.14
13	0.3	0.2	80	1	0.47	0.48	0.49	0.38	0.00	0.00	0.02
14	0.4	0.2	80	1	0.93	0.93	0.93	0.85	0.00	0.00	0.22
15	0.3	0.2	50	2	0.27	0.29	0.29	0.25	0.01	0.00	0.09
16	0.4	0.2	50	2	0.62	0.64	0.65	0.62	0.01	0.00	0.37
17	0.3	0.2	20	3	0.11	0.12	0.14	0.14	0.01	0.03	0.07
18	0.4	0.2	20	3	0.19	0.24	0.31	0.34	0.04	0.12	0.26

[a]For each scenario set up, the total number of simulation runs is 2000 and the total sample size in each simulation run is 80

In Scenarios 2 and 3, where the experimental treatment works uniformly well for the overall population, the testing in the overall population gives highest power, as expected, while the other methods, which allow an alternative pathway for an enriched subpopulation, resulted in mild to moderate power loss.

Scenarios 6, 9 and 12 assumes that the response rate is 0.4 for the biomarker positive population and 0 for the biomarker negative population, with the prevalence for the biomarker positive population varying from 80, 50, to 20%. The performance of ASD and CV-ASD is close to the overall testing, while the BAT method outperforms the other three by an impressive margin when the prevalence of biomarker positive population is relatively low. In addition, the BAT method appears to perform better in identifying the correct threshold most of the cases. More research is needed to shed more insight into this finding.

Scenarios 14, 16 and 18 mirror the settings in Scenarios 6, 9 and 12, except that the response rate for the biomarker negative population is assumed to be 20% (the null) instead of 0. Similarly, the benefit of employing the BAT Approach starts to manifest as the prevalence of biomarker becomes lower. However, the advantage appears to be less pronounced when the response rate for the biomarker negative population increases from zero to 20%.

In conclusion, the ASD and CV-ASD method performs similarly to the overall testing but does have an alternative pathway for an enriched subpopulation. Adding cross validation component to ASD does not provide additional benefit at least in the scenarios we evaluated. On the other hand, the BAT method appears to have better operating characteristics in terms of establishing an efficacy signal as well as identifying the correct biomarker threshold, especially when the prevalence of biomarker positive population is relatively low and the difference of true response rates between the biomarker positive and negative populations is pronounced.

5 Discussions

In this paper, we proposed three strategies to detect efficacy signals in single-arm POC studies which allow simultaneous statistical significance testing either in the overall population or in a biomarker enriched subpopulation to be identified using the same study. None of the three methods require the pre-specification of a biomarker threshold to identify the sensitive population. Instead they allow the threshold to be estimated using the data but still maintain reasonable statistical rigor. Among the proposed methods, the BAT appears to have notably better operating characteristics, particularly when the difference in response rate between biomarker positive and negative is pronounced and the prevalence of biomarker positive population is relatively low. The ASD and CV-ASD perform similarly to the overall test, which may be due to the relatively small size of the biomarker enriched subpopulation, which is only a part of the validation cohort, as well as the relatively conservative Bonferroni adjustment. In contrast, the implicit alpha sharing in the BAT test via a maximal test statistic is theoretically more efficient. The choice of the value for R needs to be

further elucidated. Intuitively, a very large value of R will let the overall test dominate the maximally selected test statistic, hence approximating the performance of BAT closer to the overall testing. The performance of ASD was not improved by adding cross-validated component in our evaluation, in spite of the complication of biomarker threshold identification when different development cohorts are used.

In the POC stage, a companion diagnosis assay may still be in development, meaning that the quantification of the biomarker of interest may have yet to be perfected. There are two implications: (1) the threshold estimated in the POC study should still be further evaluated and validated in future development studies, and (2) it is important to evaluate the impact of measurement error in biomarker quantification on the operating characteristics of the proposed strategies. The preliminary results (not presented here) of our ongoing research shows that the adverse effect of measurement error may be mitigated by adjusting for misclassifications utilizing a predictive value weighting method [8].

References

1. Alosh, M., Huque, M.F.: A flexible strategy for testing subgroups and overall population. Stat. Med. **28**, 3–23 (2009)
2. Alosh, M., Huque, M.F.: A consistency-adjusted alpha-adaptive strategy for sequential testing. Stat. Med. **29**, 1559–1571 (2010)
3. Freidlin, B., Simon, R.: Adaptive signature design: an adaptive clinical trial design for generating 1519 and prospectively testing a gene expression signature for sensitive patients. Clin. Cancer 1520 Res. 11, 7872–7878 (2005)
4. Freidlin, B, Jiang, W., Simon, R.: The cross-validated adaptive signature design. Clin. Cancer 1522 Res. 16, 691–698 (2010)
5. Food and Drug Administration. Guidance for Industry: Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products, 25 May 2005
6. Hall, P., Wilson, S.: Two guidelines for bootstrap hypothesis testing. Biometrics **47**(2), 757–762 (1991)
7. Jiang, W., Freidlin, B., Simon, R.: Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J. Natl. Cancer Inst. 99(13), 1036–1043 (2007)
8. Lyles, R., Lin, J.: Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. Stat. Med. **29**(22), 2297–2309 (2010)
9. Song, Y., Chi, G.Y.: A method for testing a prespecified subgroup in clinical trials. Stat. Med. **26**, 3535–3549 (2007)

Challenges of Bridging Studies in Biomarker Driven Clinical Trials: The Impact of Companion Diagnostic Device Performance on Clinical Efficacy

Szu-Yu Tang and Bonnie LaFleur

Abstract Personalized medicine involves the co-development of both the therapeutic agent (Rx) and a companion diagnostic device (CDx), which directs a group of patients to a particular treatment. There are instances, however, when there are competing, or multiple CDx products for a given Rx. Drivers for multiple CDx products can be driven by improved efficiency, cost, novel technologies, or updated techniques over time. In these instances, concordance between the old assay (e.g., the assay used in the clinical trial or comparator companion diagnostic device in this paper) and a new assay (follow-on companion diagnostic device) needs to be assessed. Discrepancies between the old and new assays, and specifically the impact of discordance on clinical efficacy, need to be evaluated. Studies that establish similarity between two or more CDx products are called bridging studies. We provide a statistical framework for method comparison studies where there is bias in measurement of one or both assessments. We then present a simulation study to evaluate the statistical impact of an imperfect CDx on the sensitivity and specificity of the follow-on companion diagnostic device. Further, we demonstrate the influence of the CDx accuracy on clinical efficacy in the context of an enrichment clinical trial.

Keywords Bridging studies · Companion diagnostic device (CDx) · Comparator companion diagnostic device · Follow-on companion diagnostic device · Personalized medicine

1 Introduction

Personalized medicine is the practice of prescribing treatments that are tailored for groups of patients who will benefit from a specific therapy. Patients are identified

S.-Y. Tang (✉)
Ventana Medical Systems, Inc., Tucson, AZ, USA
e-mail: tang.142@buckeyemail.osu.edu

B. LaFleur
HTG Molecular Diagnostics, Tucson, AZ, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_16

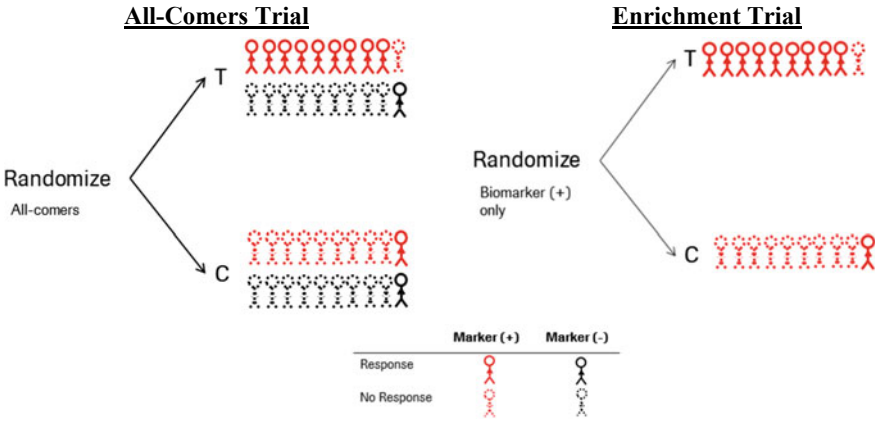


Fig. 1 Schematic illustrating the difference between an all-comers clinical trial and an enrichment clinical trial design. T is treatment arm and C is control arm in the clinical trial. Red indicates CDx positive patients and black indicates CDx negative patients. Solid symbols represent responders and dotted symbols represent non-responders to the drug. The proportion of responders versus non-responders depicted above is hypothetical

by one or more biomarkers [1]. The successful implementation of personalized medicine relies on both an accurate companion diagnostic device (CDx) that can correctly identify the patients who will benefit from a specific treatment, and the efficacious treatment of the identified patients. Therefore, personalized medicine is a co-development process, where both drug efficacy and device accuracy determine the success of treatment.

Two types of strategy for biomarker trials are enrichment trials or all-comers trials (Fig. 1). An all-comers trial enrolls all patients meeting the eligibility criteria (regardless of a particular biomarker status). An enrichment design prospectively selects a study population in which detection of a drug effect is more likely than it would be in an all-comers trial. In this paper, an enrichment trial refers to the clinical trial that only enrolls patients who are biomarker positive. Because enrichment trials randomize patients within a known biomarker status, as identified by a companion diagnostic, it is important to know how the accuracy of a CDx assay impacts evaluation of drug efficacy. For the purposes of the methods described here, we assume the same drug therapy is paired with multiple diagnostic devices. For example, after having received FDA-approval with a diagnostic device (i.e., a comparator companion diagnostic, CCD), a drug company seeks another less expensive, more effective, or updated device (i.e., a follow-on companion diagnostic device (FCD) [2]. An FCD should demonstrate similar safety and efficacy as the CCD and a common approach is to evaluate this expectation is by way of a bridging study. The difficulties and challenges for bridging studies have been discussed in several papers [2–4]. In this article, we focus on evaluation of the impact of FCD accuracy on the establishment of clinical efficacy under FCD stratification.

Sensitivity and specificity are measures of performance for a diagnostic device. Sensitivity of an assay is defined as the probability that the CDx result is positive when the patient's true biomarker status is positive. Specificity is defined as the probability that the CDx tested result is negative when the patient's true biomarker status is negative. Both sensitivity and specificity provide distinct, and equally important, pieces of information; and the FDA recommends optimization of metrics unless clinically justifiable otherwise. In this paper, CDx accuracy (or assay performance) refers to both sensitivity and specificity.

In Sect. 2, we establish the relationship between clinical efficacy and CDx accuracy. We use simulation results to demonstrate how CDx accuracy impacts clinical efficacy in an enrichment trial design. In Sect. 3, we extend the relationship between clinical efficacy and CDx accuracy to include two assays (FCD and CCD) for the same therapy. We use simulation results to explain how FCD accuracy impacts efficacy conditioned on positive and negative test results from the CCD, assuming both FCD and CCD are independent and correlated scenarios.

2 The Impact of Diagnostic Accuracy on Clinical Efficacy in an Enrichment Trial (Single Assay)

2.1 Assumptions and Notation

A simple case scenario is the impact of diagnostic accuracy on clinical efficacy in an enrichment trial using a single CDx assay. It is assumed that the enrichment trial is designed to enroll biomarker positive patients using a CDx assay. Notation and parameters include: total number of patients screened (n); true biomarker positive or negative status (G+ or G−); biomarker positive prevalence (π); CDx sensitivity (S), and assay specificity (C). The number of patients enrolled and not-enrolled, based on results from the CDx assay, can be calculated as shown in Table 1.

For this disposition table, we assume a 1:1 randomization such that half of the patients are assigned to the treatment arm and the other half are assigned to the control arm. Therefore, ½*[n* π * S + n*(1 − π)*(1 − C)] patients receive treatment drug and the same number of patients receive control drug.

Table 1 Disposition table of patients enrolled or not enrolled in an enrichment trial, expressed by CDx sensitivity and specificity

		CDx		Total
		CDx + (Enrolled)	CDx – (Not Enrolled)	
True biomarker status	G+	n ∗ π ∗ S	n ∗ π ∗ (1 − S)	n ∗ π
	G−	n ∗ (1 − π) ∗ (1 − C)	n ∗ (1 − π) ∗ C	n ∗ (1 − π)

Let the true response rate (RR) from true biomarker positive patients (G+) and who receive the treatment equal $\gamma + \delta(\delta > 0)$ and let the response rate equal γ for all other groups (control arm patients or G− patients). Therefore, the expected numbers of responders in the treatment arm are $^1/_2 * \left[n * \pi * S * (\gamma + \delta) + n * (1 - \pi) * (1 - C) * \gamma \right]$. In the control arm, the expected numbers of responders are $^1/_2 * \left[n * \pi * S * \gamma + n * (1 - \pi) * (1 - C) * \gamma \right]$. The clinical efficacy in an enrichment trial (r+) is measured by way of the difference between response rate in the treatment and control arms. It follows that the relationship between clinical efficacy (r+) and CDx accuracy (S and C) is defined as follows:

$$r_+ = \frac{^1/_2 * \left[n^* \pi * S^* (\gamma + \delta) + n^* (1 - \pi) * (1 - C) * \gamma \right]}{^1/_2 * [n^* \pi * S + n^* (1 - \pi) * (1 - C)]} - \gamma$$
$$= \delta * \frac{\pi s}{\pi s + (1 - \pi)(1 - c)}$$

2.2 Simulation

From the relationship established above, clinical efficacy in the enrichment trial is impacted by both CDx sensitivity and specificity. We conduct a simple simulation to express the association between accuracy and efficacy. This simulation assumes the prevalence of biomarker (π) is 50%, RR for G+ and treatment group is $\gamma + \delta = 60\%$, and RR for all other groups (G− or control group) is $\gamma = 40\%$. Therefore, in an enrichment trial which enrolls patients by a CDx, the expected clinical efficacy (r+) would be between 0% and 20%.

2.3 Simulation Results

A CDx accuracy or assay performance (i.e., specificity or sensitivity) of less than 50% means that the chance to correctly identity biomarker positive or negative patients is less than 50% and it is unlikely that an assay of such low accuracy would be on market. Therefore, this discussion is focused on the assay performance above 50% (the gray area in Fig. 2).

Figure 2 shows the relationship between assay performance (either specificity or sensitivity) on the x-axis and clinical efficacy (r+) on the y-axis when either sensitivity (S) or specificity (C) is fixed at 50, 70, and 90% (i.e., panels 1, 2, and 3). If C is fixed, sensitivity (S) varies (solid curve) and if S is fixed, specificity (C) varies (dotted curve).

At 50% assay performance (white-gray boundary on Fig. 2), sensitivity is associated with higher (or equal) clinical efficacy (r+) relative to specificity. However,

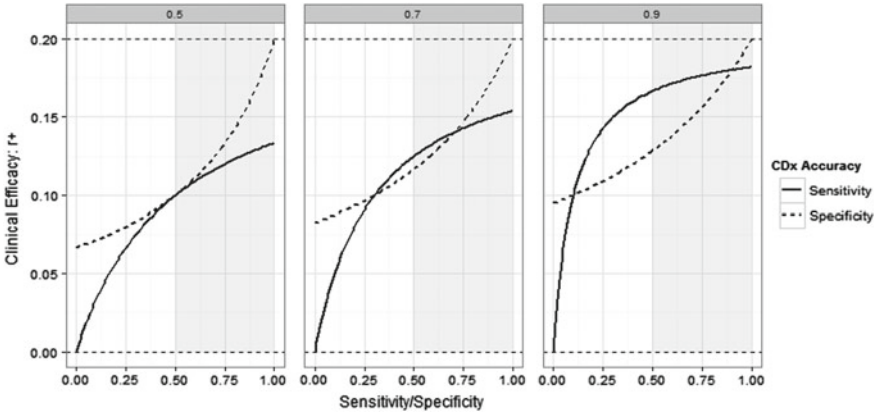


Fig. 2 Graph depicting the impact of CDx accuracy on clinical efficacy in an enrichment trial (single assay). The solid curves are the relationship between sensitivity (S) and clinical efficacy in an enrichment trial (r+) when specificity (C) is fixed at 50%, 70% and 90% (panels 1, 2, and 3 respectively). The dotted curves are the relationship between specificity (C) and clinical efficacy in an enrichment trial (r+) when sensitivity (S) is fixed at 50%, 70% and 90% (panels 1, 2, and 3 respectively)

as assay performance increases (approaching 100%), specificity is associated with higher clinical efficacy compared to sensitivity.

When specificity equals 100%, clinical efficacy (r+) reaches a maximum of 20%, regardless of sensitivity (i.e., specificity reaches 20% at each panel shown in Fig. 2). On the other hand, clinical efficacy (r+) does not reach a 20% maximum when sensitivity equals 100%. This implies that the upper limit of clinical efficacy (r+) is determined by specificity and not sensitivity. In addition, dotted curves (specificity) have a steeper slope than solid curves (sensitivity) when assay performance is high. This suggests that CDx assay specificity is a better indicator of improved drug efficacy than sensitivity, especially when the diagnostic accuracy is high.

3 The Impact of FCD Accuracy on Clinical Efficacy Conditioned on a CCD in an Enrichment Trial (Two Assays)

3.1 Assumption and Notation

A more complicated situation occurs when an FCD is developed following the enrollment of patients in a clinical trial using a CCD. In these cases, a bridging study may be needed to bridge the efficacy from the CCD to the FCD. The relationship between the FCD and clinical efficacy is influenced by the accuracy of the CCD used to enroll

patients. In this section, we extend the simulation from one assay to two assays (CCD and FCD) and establish the relationship between clinical efficacy and FCD assay performance conditional on CCD assay performance.

Following the notation and assumptions from 2.1, let n be the total number of patients screened, G+ or G− be true biomarker positive or negative status and π be the biomarker positive prevalence. Furthermore, let sensitivity of the CCD and the FCD be S1 and S2 respectively, and let specificity of the CCD and the FCD be C1 and C2, respectively.

First, we describe the number of patients under CCD and FCD assay performance by a 2 by 2 by 2 ($2 \times 2 \times 2$) table (true marker status by CCD status by FCD status) which is like Table 1 in Sect. 2. The disposition table can be constructed based on two different assumptions: under the conditional independent assumption (CIA), where CCD and FCD are conditionally independent; and without the CIA, where CCD and FCD are correlated.

3.1.1 Under CIA: CCD and FCD are Conditionally Independent

The conditional independence assumption (CIA) assumes that CCD test results and FCD test results are independent given the true marker status (G+ or G−). This means CCD and FCD do not tend to misdiagnose the same patient [6]. Under CIA, the joint probability of test results from CCD and FCD, conditioned on the true biomarker status, can be calculated as follows:

$$\begin{aligned} p^{ind}_{1+\cap 2+|G+} &= \Pr(\text{FCD} = + \cap \text{CCD} = +|\text{G}+) \\ &= \Pr(\text{FCD} = +|\text{G}+) * \Pr(\text{CCD} = +|\text{G}+) \qquad \text{(CIA)} \\ &= S_1 S_2 \end{aligned}$$

Similarly,

$$p^{ind}_{1-\cap 2+|G+} = (1 - S1) * S2$$

$$p^{ind}_{1+\cap 2+|G-} = (1 - C1) * (1 - C2)$$

$$p^{ind}_{1-\cap 2+|G-} = C1 * (1 - C2)$$

$$p^{ind}_{1+\cap 2-|G+} = S1 * (1 - S2)$$

$$p^{ind}_{1-\cap 2-|G+} = (1 - S1) * (1 - S2)$$

$$p^{ind}_{1+\cap 2-|G-} = (1 - C1) * C2$$

$$p^{ind}_{1-\cap 2-|G-} = C1 * C2$$

Table 2 Using CIA, Disposition table of patients grouped by true biomarker status (G+ or G−) and test results status (+ or −) for both CCD and FCD

	G+			G-		
	CCD+	CCD-	Total	CCD+	CCD-	Total
FCD+	$n * \pi * S1 * S2$	$n * \pi * (1 - S1) * S2$	$n * \pi * S2$	$n * (1 - \pi) * (1 - C1) * (1 - C2)$	$n * (1 - \pi) * C1 * (1 - C2)$	$n * (1 - \pi) * (1 - C2)$
FCD-	$n * \pi * S1 * (1 - S2)$	$n * \pi * (1 - S1) * (1 - S2)$	$n * \pi * (1 - S2)$	$n * (1 - \pi) * (1 - C1) * C2$	$n * (1 - \pi) * C1 * C2$	$n * (1 - \pi) * C2$
Total	$n * \pi * S1$	$n * \pi * (1 - S1)$	$n * \pi$	$n * (1 - \pi) * (1 - C1)$	$n * (1 - \pi) * C1$	$n * (1 - \pi)$

Applying the probabilities calculated under CIA, the sampling distribution of results can be described in a $2 \times 2 \times 2$ table based on FCD and CCD accuracy (S1, S2, C1 and C2) (Table 2).

The relationship between clinical efficacy and both CCD and FCD accuracy using sampling distributions is provided in Table 2. As discussed in Sect. 1, patients in the clinical trial would be tested and enrolled using CCD; therefore, the bridging study examines the drug efficacy that would likely have occurred if the FCD had been used to enroll patients. Drug efficacy in FCD positive patients is established for CCD positive and negative patients separately.

We assume that patients are 1:1 randomized into the treatment arm and control arm within stratified FCD and CCD test results. We again assume that the response rate for patients that are both G+ and within the treatment group equals $\gamma + \delta$ ($\delta > 0$) and for all other patients (control arm patients or G- patients) response rate equals γ.

Conditioned on the CCD tested positive patients, the number of FCD positive patients is equal to $^1/_2 * [n * \pi * S1 * S2 + n * (1 - \pi) * (1 - C1) * (1 - C2)]$. The expected numbers of responders is equal to $^1/_2 * \left[n * \pi * S1 * S2 * (\gamma + \delta) + n * (1 - \pi) * (1 - C1) * (1 - C2) * \gamma\right]$. The clinical efficacy for FCD positive patients, conditioned on CCD positive ($r^{ind}_{2+|1+}$), is derived as below:

$$r^{ind}_{2+|1+} = \delta * \frac{\pi S_1 S_2}{\pi S_1 S_2 + (1 - \pi)(1 - C_1)(1 - C_2)}$$

Similarly, the clinical efficacy for FCD positive patients, conditioned on CCD negative ($r^{ind}_{2+|1-}$), is derived as below:

$$r^{ind}_{2+|1-} = \delta * \frac{\pi(1 - S_1) S_2}{\pi(1 - S_1) S_2 + (1 - \pi) C_1 (1 - C_2)}$$

3.1.2 Without CIA: CCD and FCD are Correlated

In the previous Sect. 3.1.1, we assume that CCD and FCD are conditionally independent given the true biomarker status (G+ or G−). In reality, CCD and FCD are highly correlated for a given patient. For example, if a patient's marker expression is highly positive, there is higher chance that CCD and FCD test results are also positive. To account for conditional dependence between the CCD and the FCD, the covariance (covp or covn) is included among those who have positive or negative latent true biomarker status where 0 <= covp <= (min(S1, S2) − S1 ∗ S2) and 0 <= covn <= (min(C1, C2) − C1 ∗ C2) [5]. Table 2 can be revised as Table 3 by adding covariance parameters.

By adding the covariance, we refine the relationship in an enrichment trial between clinical efficacy and FCD accuracy, conditioned on CCD positive ($r^{corr}_{2+|1+}$) and CCD negative ($r^{corr}_{2+|1-}$), without CIA as follows:

$$r^{corr}_{2+|1+} = \delta * \frac{\pi(S_1 S_2 + covp)}{\pi(S_1 S_2 + covp) + (1-\pi)((1-C_1)(1-C_2) + covn)}$$

$$r^{corr}_{2+|1-} = \delta * \frac{\pi((1-S_1)S_2 - covp)}{\pi((1-S_1)S_2 - covp) + (1-\pi)(C_1(1-C_2) - covn)}$$

3.2 Simulation

Similar to that performed in Sect. 2 (the impact of diagnostic accuracy on clinical efficacy in an enrichment trial: single assay), in this simulation we assess the impact of diagnostic accuracy on clinical efficacy in an enrichment trial for FCD accuracy conditioned on CCD positive or negative results. Fixed assumptions for the simulations include prevalence of biomarker (π) fixed at 50%. RR for G+ and treatment group is 60% and RR for G− or control group is 40%. For the "without CIA scenario", covp and covn are added using ((min (S1, S2) − S1 ∗ S2) ∗ 0.8) and ((min (C1, C2) − C1 ∗ C2) ∗ 0.8), respectively, to represent an 80% correlation between two assays.

Figures 3 and 4 show FCD assay performance (either specificity or sensitivity) on the x-axis and clinical efficacy under CIA ($r^{ind}_{2+|1+}$ and $r^{ind}_{2+|1-}$) on the y-axis when either FCD sensitivity (S2) or FCD specificity (C2) is fixed at 50, 70, and 90% (i.e., panels 1, 2, and 3). Figures 5 and 6 show the relationship between FCD assay performance (either specificity or sensitivity) and clinical efficacy without using CIA ($r^{corr}_{2+|1+}$ and $r^{corr}_{2+|1-}$). If C2 is fixed, sensitivity (S2) varies (solid curve) and if S2 is fixed, specificity (C2) varies (dotted curve). The different colored curves represent the different combinations of CCD accuracy (S_1, C_1) at 50, 70, and 90%. Curves representing a specificity of C1 = 90% are shown in light-dark green, curves representing a specificity of C1 = 70% are shown in purple-pink, and curves representing a specificity of C1 = 50% are shown in light-dark blue.

Table 3 Without using CIA, disposition table of patients grouped by true biomarker status (G+ or G−) and test results status (+ or −) for both CCD and FCD

	G = +			G = −		
	CCD+	CCD-	Total	CCD+	CCD-	Total
FCD+	$n * \pi * (S1 * S2 + covp)$	$n * \pi * ((1-S1) * S2 - covp)$	$n * \pi * S2$	$n * (1-\pi) * ((1-C1) * (1-C2) + covn)$	$n * (1-\pi) * (C1 * (1-C2) - covn))$	$n * (1-\pi) * (1-C2)$
FCD-	$n * \pi * (S1 * (1-S2) - covp)$	$n * \pi * ((1-S1) * (1-S2) + covp)$	$n * \pi * (1-S2)$	$n * (1-\pi) * ((1-C1) * C2 - covn)$	$n * (1-\pi) * (C1 * C2 + covn)$	$n * (1-\pi) * C2$
Total	$n * \pi * S1$	$n * \pi * (1-S1)$	$n * \pi$	$n * (1-\pi) * (1-C1)$	$n * (1-\pi) * C1$	$n * (1-\pi)$

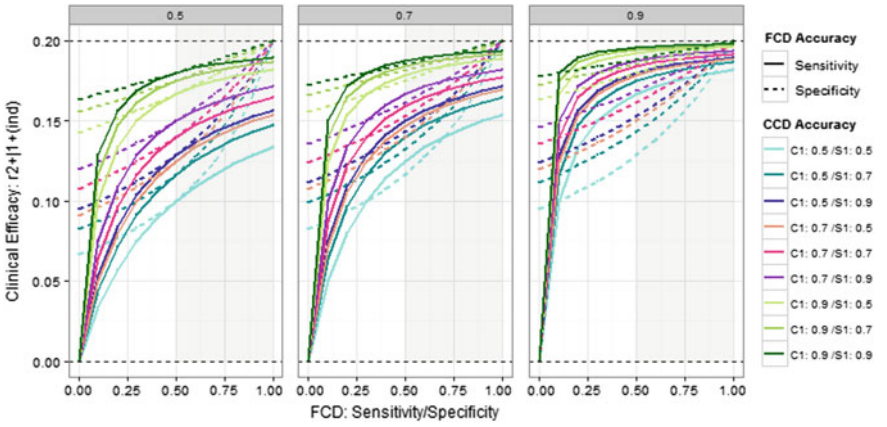


Fig. 3 Graph depicting the impact of FCD accuracy on clinical efficacy in an enrichment trial conditioned on CCD tested result is positive ($r^{ind}_{2+|1+}$) using CIA. Solid curves represent the relationship between FCD sensitivity (S2) versus $r^{ind}_{2+|1+}$ when specificity (C2) is fixed at 50%, 70% and 90% (panels 1, 2, and 3, respectively). The dotted curves represent the relationship between specificity (C2) versus $r^{ind}_{2+|1+}$ when sensitivity (S2) is fixed at 50%, 70% and 90% (panels 1, 2, and 3, respectively). Different color of curves indicates different combination of CCD sensitivity and specificity

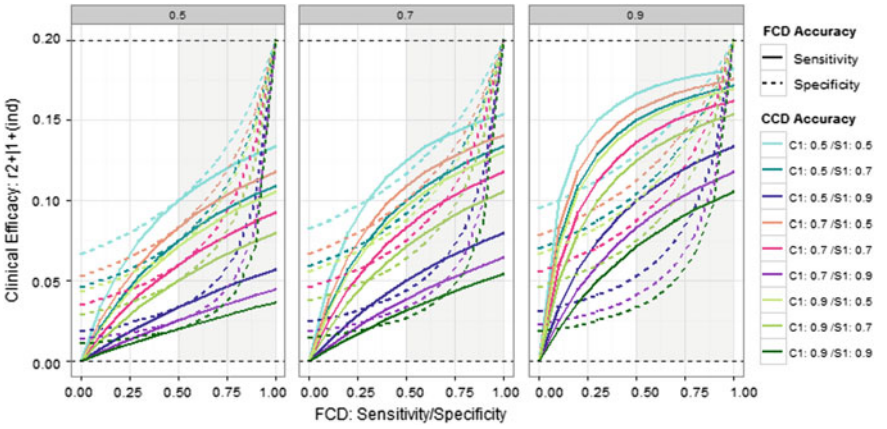


Fig. 4 Graph depicting the impact of FCD accuracy on clinical efficacy in an enrichment trial conditioned on CCD tested result is negative ($r^{ind}_{2+|1-}$) using CIA. Solid curves represent the relationship between FCD sensitivity (S2) versus $r^{ind}_{2+|1-}$ when specificity (C2) is fixed at 50%, 70% and 90% (panels 1, 2, and 3, respectively). The dotted curves represent the relationship between specificity (C2) versus $r^{ind}_{2+|1-}$ when sensitivity (S2) is fixed at 50%, 70% and 90% (panels 1, 2, and 3, respectively). Different color of curves indicates different combination of CCD sensitivity and specificity

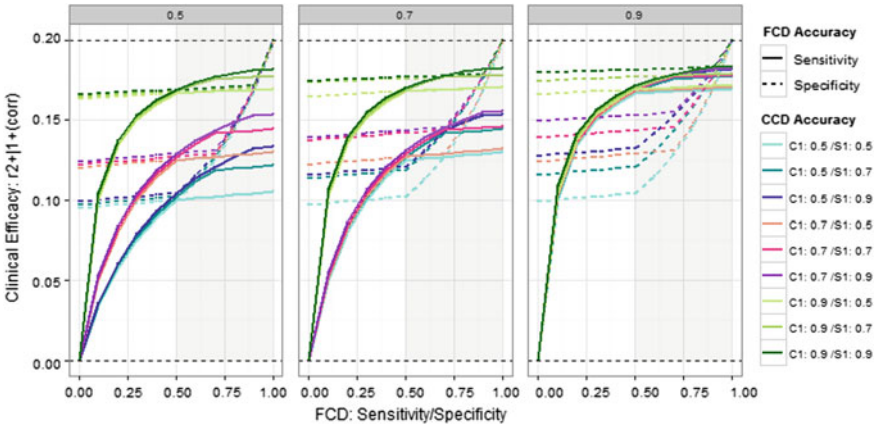


Fig. 5 Graph depicting the impact of FCD accuracy on clinical efficacy in an enrichment trial conditioned on CCD tested result is positive ($r^{corr}_{2+|1+}$) when CCD and FCD are 80% correlated. Solid curves represent the relationship between FCD sensitivity (S2) versus $r^{corr}_{2+|1+}$ when specificity (C2) is fixed at 50%, 70% and 90% (panels 1, 2, and 3, respectively). The dotted curves represent the relationship between specificity (C2) versus $r^{corr}_{2+|1+}$ when sensitivity (S2) is fixed at 50%, 70% and 90% (panels 1, 2, and 3, respectively). Different color of curves indicates different combination of CCD sensitivity and specificity

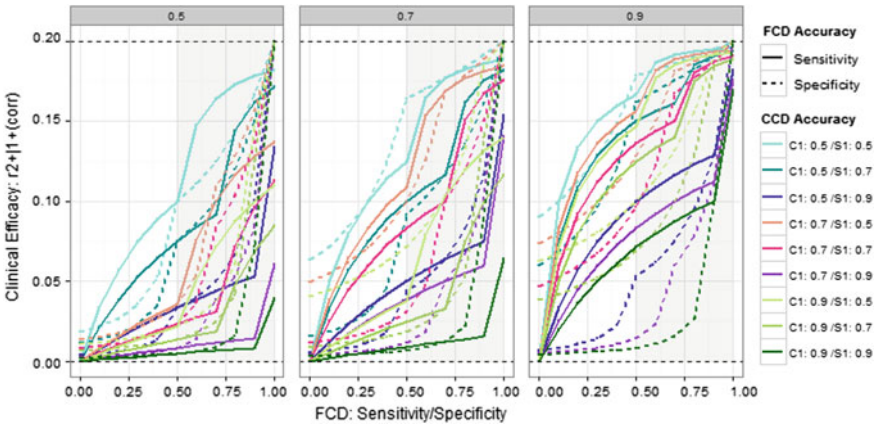


Fig. 6 Graph depicting the impact of FCD accuracy on clinical efficacy in an enrichment trial conditioned on CCD tested result is negative ($r^{corr}_{2+|1-}$) when CCD and FCD are 80% correlated. Solid curves represent the relationship between FCD sensitivity (S2) versus $r^{corr}_{2+|1-}$ when specificity (C2) is fixed at 50%, 70% and 90% (panels 1, 2, and 3, respectively). The dotted curves represent the relationship between specificity (C2) versus $r^{corr}_{2+|1-}$ when sensitivity (S2) is fixed at 50%, 70% and 90% (panels 1, 2, and 3, respectively). Different color of curves indicates different combination of CCD sensitivity and specificity

3.3 Simulation Results

3.3.1 With CIA and Conditioned on CCD Positive Results

Figure 3 shows the impact of FCD accuracy on clinical efficacy in the enrichment trial given that the CCD test result is positive ($r^{ind}_{2+|1+}$) under CIA. Regardless of CCD sensitivity and specificity (i.e., shown in different colored curves), the impact of the FCD on efficacy is similar to observations in the single assay scenario described in Sect. 2. In other words, as assay performance increases (approaching 100%), (1) specificity is associated with greater clinical efficacy than is sensitivity, and (2) specificity can reach a maximum clinical efficacy of 20% whereas sensitivity cannot.

When considering CCD sensitivity and specificity (i.e., colored curves), the simulation results demonstrate highest clinical efficacy occurs when CCD specificity is equal to 90% (C1 = 0.9, 3 light-dark green curves) and lowest clinical efficacy occurs when CCD specificity is equal to 50% (C1 = 0.5, two light-dark blue curves). This indicates that, for CCD tested positive patients; higher CCD specificity predicts increased clinical efficacy. This is because there are fewer false positive results with higher CCD specificity. Therefore, we can expect more power to detect clinical efficacy if both FCD and CCD specificities are high.

3.3.2 CIA and Conditioned on CCD Negative Results

Figure 4 shows the impact of FCD accuracy on clinical efficacy in the enrichment trial given that the CCD test result is negative ($r^{ind}_{2+|1-}$) using CIA. Similar to that observed for CCD positive (Sect. 3.3.1), FCD specificity is associated with greater clinical efficacy than is sensitivity, and specificity reaches a maximum clinical efficacy of 20%.

However, unlike that observed for CCD positive, when considering CCD sensitivity and specificity, the simulation results show that the higher clinical efficacy occurs when CCD sensitivity is equal to 50% (S1 = 0.5) and the lower clinical efficacy occurs when CCD sensitivity is equal to 90% (S1 = 0.9). This suggests lower CCD sensitivity can increase $r^{ind}_{2+|1-}$ to a greater degree than higher CCD sensitivity. This is likely because true positive patients (i.e., G+) were misclassified by the CCD as negative, because of bad CCD sensitivity, but then identified as positive by the FCD as positive, thereby increasing efficacy.

To summarize the CIA results, when CCD test result is positive, increased specificity of both CCD and FCD is associated with increased clinical efficacy. When CCD test result is negative, decreased sensitivity of CCD and increased specificity of FCD is associated with increased clinical efficacy.

3.3.3 Without CIA and Conditioned on CCD Positive Results

Figure 5 shows the impact of FCD accuracy on clinical efficacy in an enrichment trial given that the CCD test result is positive ($r^{corr}_{2+|1+}$) and when FCD and CCD have an 80% correlation. Regardless of CCD sensitivity and specificity (i.e., shown in Fig. 4 in different colored curves), the FCD impact on efficacy without CIA is similar to the impact observed for both the single assay and CIA scenarios, except that the shapes of the curves are slightly different. The clinical efficacy increases quickly at lower values of FCD sensitivity and then stays at similar level for the higher values of sensitivity (i.e., the solid curves in Fig. 5). On the contrary, clinical efficacy stays at similar level for lower values of FCD specificity but quickly increases at the higher values of FCD sensitivity (i.e., the dotted curves in Fig. 5). This suggests there is some impact plateau of improving clinical efficacy corresponding to correlation between FCD and CCD.

Similar to the results for CCD positive under CIA (Sect. 3.3.1), simulation results show that clinical efficacy is demonstrated when CCD specificity is equal to 90% (C1 = 0.9, 3 light-dark green curves) and diminishes when CCD specificity is equal to 50% (C1 = 0.5, two light-dark blue curves). This again suggests that higher CCD specificity results in increased clinical efficacy. This is especially true when sensitivity or specificity of the FCD is low (i.e., Figure 5, Panel 1). The impact of FCD accuracy on clinical efficacy are similar when CCD specificity is at the same level regardless of CCD sensitivity (i.e., solid curves or dotted curves are grouped together by the same level of C1 in Fig. 5). In other words, the degree of which the CCD specificity impacts efficacy is greatest with lower accuracy based on the FCD.

3.3.4 Without CIA and Conditioned on CCD Negative Results

Figure 6 shows the impact of FCD accuracy on clinical efficacy in an enrichment trial given that the CCD test result is negative ($r^{corr}_{2+|1-}$) and when FCD and CCD have an 80% correlation. The consistent message is that FCD specificity is more associated with clinical efficacy than is sensitivity, and specificity reaches a maximum clinical efficacy of 20%. However, comparing using CIA versus without using CIA and conditioned on CCD negative results (Fig. 4 versus Fig. 6), we observed the different shape of the solid curve (sensitivity of FCD vs. $r^{corr}_{2+|1-}$ when specificity of FCD is fixed) and the dotted curve (specificity of FCD vs. $r^{corr}_{2+|1-}$ when sensitivity of FCD is fixed). The solid curves are very steep when FCD sensitivity value is high. This indicates small improvements of FCD sensitivity can quickly increase clinical efficacy.

When considering CCD sensitivity and specificity (i.e., different colored curves in Fig. 6), the maximum clinical efficacy occurs when sensitivity is equal to 50% (S1 = 0.5) and the minimum clinical efficacy occurs when sensitivity is equal to 90% (S1 = 0.9). This is also a consistent message from under CIA conditioned on CCD negative results (Sect. 3.3.2) which indicates that lower CCD sensitivity can increase $r^{corr}_{2+|1-}$ to a greater degree than higher CCD sensitivity.

In summary, simulations based on CIA show similar results to simulations without CIA; specifically, for cases where the CCD test result is positive, increased specificity of both CCD and FCD is associated with increased clinical efficacy. And when the CCD test result is negative, decreased sensitivity of CCD and increased specificity of FCD is associated with increased clinical efficacy. However, the shape of sensitivity and specificity curves are different under these two assumptions (with and without CIA).

4 Discussion and Summary

We demonstrate the impact that CDx accuracy has on clinical efficacy in an enrichment trial. For a single assay, clinical efficacy of a targeted treatment is a function of the sensitivity and specificity of the CDx used to identify patients. CDx assays with poor assay performance diminish the chance of demonstrating clinical efficacy and may cause the enrichment trial to fail. Moreover, specificity, rather than sensitivity, is more correlated with improvements in clinical efficacy. In an enrichment trial, only diagnostic positive patients will be enrolled into clinical trial. An assay with better specificity would ensure that there are fewer false positive patients enrolled, which would dilute the true efficacy.

We also have the impact of a diagnostic device accuracy of a follow-on companion device (FCD) on clinical efficacy. Specifically, accuracy of the FCD conditioned on the test results from a comparator companion diagnostic device (CCD). The scenarios with and without considering the conditional independence assumption (CIA) between CCD and FCD are both evaluated. Simulations show how increased specificity of a FCD can enhance clinical efficacy regardless of CCD. These results are consistent with those found for the single assay scenario. Furthermore, the simulations under CIA or without CIA suggest: (1) when the CCD test result is positive, increased specificity of both CCD and FCD is associated with increased clinical efficacy; and (2) when the CCD test result is negative, decreased sensitivity of CCD and increased specificity of FCD is associated with increased clinical efficacy. In actual enrichment trial; however, CCD negative patients are not enrolled; therefore, efficacy will not be evaluated for this subgroup. However, the simulation results of conditioned on CCD native patients can provide theoretically insight about how the efficacy is impacted if the performance of CCD and FCD are known.

Acknowledgements The authors gratefully thank Dr. Chang Xu from Qiagen and Dominic LaRoche from HTG Molecular Diagnostics.

References

1. Maitournam, A., Simon, R.: On the efficiency of targeted clinical trials. Stat. Med. **24**(3), 329–339 (2005)
2. Li, M.J.: Statistical methods for clinical validation of follow-on companion diagnostic devices via an external concordance study. Stat. Biopharm. Res. **8**(3), 355–363 (2016)
3. Li, M.J.: Statistical consideration and challenges in bridging study of personalized medicine. J. Biopharm. Stat. **25**, 397–407 (2015)
4. Li, M.J., Yu, T.H., Hu, Y.F.: The impact of companion diagnostic device measurement performance on clinical validation of personalized medicine. Stat. Med. **34**, 2222–2234 (2015)
5. Lu, Y., Dendukuri, N., Schiller, I., Joseph, L.: A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. Stat. Med. **29**(24), 2532–2543 (2010)
6. Zhou, X.-H., Obuchowski, N.A., McClish, D.K.: Statistical methods in diagnostic medicine, 2nd edn. Wiley, Hoboken (2011)

Part VI
Application of Novel Data Modality

Parallel-Tempered Feature Allocation for Large-Scale Tumor Heterogeneity with Deep Sequencing Data

Yang Ni, Peter Müller, Max Shpak and Yuan Ji

Abstract We developed a parallel-tempered feature allocation algorithm to infer tumor heterogeneity from deep DNA sequencing data. The feature allocation model is based on a binomial likelihood and an Indian Buffet process prior on the latent haplotypes. A variation of parallel tempering technique is introduced to flatten peaked local modes of the posterior distribution, and yields a more efficient Markov chain Monte Carlo algorithm. Simulation studies provide empirical evidence that the proposed method is superior to competing methods at a high read depth. In our application to Glioblastoma multiforme data, we found several distinctive haplotypes that indicate the presence of multiple subclones in the tumor sample.

Keywords Haplotype deconvolution · Single nucleotide variants · Next-generation sequencing data · Indian buffet process · Glioblastoma multiforme

Y. Ni
Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX, USA

P. Müller
Department of Mathematics, The University of Texas at Austin, Austin, TX, USA

M. Shpak
Sarah Cannon Research Institute, Nashville, TN, USA

Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, TX, USA

Fresh Pond Research Institute, Cambridge, MA, USA

Y. Ji (✉)
Program of Computational Genomics & Medicine, NorthShore University HealthSystem, Evanston, IL, USA
e-mail: koaeraser@gmail.com

Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_17

1 Introduction

We propose a computational strategy for statistical inference aiming to reconstruct subclone diversity and infer intra-tumor heterogeneity using deep DNA sequencing. Increased read depth provides a larger sample size per site, and consequently more information for subclone phasing. However, it also poses challenging computational problems, as the likelihood is essentially a point mass at empirical allele frequencies (assuming a binomial sampling model). Inference methods based on optimization or sampling algorithms are inefficient, due to peaked local modes. In this paper, we use a latent feature allocation model with the Indian buffet process (IBP, [1]) prior, and incorporate a variant of parallel tempering technique [2] to facilitate the sampling algorithm.

Tumors are genetically heterogeneous, often containing a diverse population of subclonal variants [3]. While somatic (or clonal) mutations associated with the early stages of tumorigenesis are typically shared across all subclones, individual subclonal lineages are defined by having somatic mutations that are not shared by other cell lines in the tumor. This is the basis of within-tumor heterogeneity, which allows cancer cells to adapt to the host environment and to therapies, and introduces substantial challenges in designing effective treatment regimes. Understanding tumor heterogeneity allows for better understanding of the causes and progression of cancer, and is critical to stratify the patient population that is likely to benefit from specific targeted therapies [4]. Many personalized treatments have been developed; some of these targeted therapies have proven to be successful, (e.g., [5, 6]).

Clonal mutations are shared by all cells and are expected to occur in the tumor at approximately 0.5 frequency, assuming no copy number variation and no loss of heterozygosity through gene conversion (we note that both assumptions are major caveats for cancer genomes), while mutations that are specific to a subclone will occur at lower frequencies. A complete characterization of genetic variation within a tumor requires the identification of the lower frequency, subclone-specific mutations. Tumor sequencing experiments using lower coverage (e.g. $<100x$) are only capable of reliably identifying the most common clonal mutations shared by most tumor cells, while deep sequencing (e.g. 500–$1000x$) facilitates the identification of low-frequency variants unique to particular subclonal lineages, provided that sufficiently large regions of tumor are sampled. Furthermore, deep sequencing allows us to accurately estimate the frequency of variant nucleotides within the tumor sample, rather than their presence or absence alone. Estimated allele frequency across sites is critical for the reconstruction of multilocus subclone genotypes and their relative frequencies.

Because next-generation sequencing (NGS) of tumor samples generates short reads from the genomes of multiple cells, and variant sites at individual reads are separately genotyped, there is the problem of "phasing" variants at different loci to reconstruct subclone genotypes. Even when the majority of somatic mutations are identified in a tumor through deep sequencing, subclonality is not known with

independent genotyping of reads. For the clonal (fixed) mutations, this problem is trivial, but for subclonal mutations, the phasing problem poses a significant challenge for cancer genomics.

The identification of subclones through phasing allows us to fully characterize tumor heterogeneity at the level of haplotypes, rather than variants at individual loci. This information is critical for the reconstruction of the evolutionary history of subclonal lineages, using phylogenetic methods [7], as well as in the application of tests of natural selection versus neutral evolution [8] that leverage multilocus genotype frequency data.

Feature allocation models have been proposed in the literature to model tumor heterogeneity with NGS data. Reference [9] studied tumor heterogeneity, in terms of subclones, which are defined as a set of single nucleotide variants (SNVs) on the same homologous genome. Since the subclones are not directly observable, they modeled the latent subclones with a finite feature allocation model, using observed SNVs read counts. The inference was implemented by reversible jump Markov chain Monte Carlo [10], using a variation, inspired by fractional Bayes factors [11]. Recently, [12] considered an infinite feature allocation model, with IBP prior, and proposed an optimization-based algorithm for maximum a posteriori estimation and scalable tumor heterogeneity inference.

However, neither of these two approaches is suitable for our deep sequencing data where read depths often approach or exceed ∼1000, with a depth of up to 4437 for some sites. In [12] and earlier applications, the read depth is typically around 50. Deep sequencing data are more informative in identifying subclones, but also pose challenges in computation because of extremely peaked likelihood and posterior distribution. We propose a parallel-tempered feature allocation algorithm (PTFA) which flattens the posterior landscape while targeting the correct posterior distribution. With parallel tempering, the Markov chain transits more smoothly in its state space and is less likely to be trapped in local modes. Empirical studies show superior performance of our methods against competing methods.

We remark that the approach presented here is not restricted to cancer genomic data. Rather, it is applicable to any data set where variant alleles and their frequencies are estimated from NGS data with reads sampled from a large number of pooled genomes, such as microbial cultures.

2 Model

2.1 Sampling Model

We briefly review the model from [9]. Let n_s and N_s denote the number of short reads that bear a variant sequence and the total number of reads at the loci of single nucleotide variant (SNV) $s = 1, \ldots, S$ in a given tumor sample. The ratio $f_s = n_s/N_s$, termed *variant allele fraction* (VAF), is the proportion of short reads bearing

a variant sequence. Given N_s, we model n_s as an independent binomial random variable $n_s \sim \text{Bin}(N_s, p_s)$ where $p_s = E(f_s)$ is the expected VAF. The likelihood is then given by

$$p(\boldsymbol{n}|\boldsymbol{N}, \boldsymbol{p}) = \prod_{s=1}^{S} p(n_s|N_s, p_s) = \prod_{s=1}^{S} \binom{N_s}{n_s} p_s^{n_s}(1-p_s)^{N_s-n_s}, \tag{1}$$

with $\boldsymbol{n} = (n_1, \ldots, n_S)$, $\boldsymbol{N} = (N_1, \ldots, N_S)$ and $\boldsymbol{p} = (p_1, \ldots, p_S)$. We assume the tumor tissue is composed of a mixture of C haplotypes, with each haplotype characterized by a different configuration of SNVs. Let $Z_{sc} \in \{0, 1\}$ be the latent binary variable that indicates whether SNV s bears a variant sequence for haplotype c and let w_c be the proportion of haplotype c in the tumor tissue for $c = 1, \ldots, C$. Then we deconvolute p_s with respect to latent haplotypes.

$$p_s = w_0\rho + \sum_{c=1}^{C} w_c Z_{sc}, \tag{2}$$

with $\sum_{c=0}^{C} w_c = 1$. The first term is added to allow for background VAF's.

2.2 Prior

The model defined by (1) and (2) involves a latent feature matrix $\boldsymbol{Z} = (Z_{sc})$, a weight vector $\boldsymbol{w} = (w_0, w_1, \ldots, w_C)$ and a background SNV frequency parameter ρ. In this section, we discuss the prior distribution for each set of parameters in turn.

In describing the prior for $\boldsymbol{Z}$, we start with a fixed number of features (haplotypes) C but will later relax it. Given C, we assume each entry of $\boldsymbol{Z}$ is an independent Bernoulli random variable $Z_{sc}|\pi_c \sim \text{Bernoulli}(\pi_c)$ with success probability π_c following a conjugate beta prior $\pi_c \sim \text{Beta}(\alpha/C, 1)$ where α is a fixed hyperparameter. Integrating out π_c, the marginal prior for $\boldsymbol{Z}$ is given by $p(\boldsymbol{Z}) = \prod_{c=1}^{C} \frac{\frac{\alpha}{C}\Gamma(m_c+\frac{\alpha}{C})\Gamma(S-m_c+1)}{\Gamma(S+1+\frac{\alpha}{C})}$, where $m_c = \sum_{s=1}^{S} Z_{sc}$.

However, in practice, the number C of latent features is unknown a priori and inference on C is often of key interest by itself. Taking the limit as $C \to \infty$ and removing columns of $\boldsymbol{Z}$ with all zeros, we obtain the IBP prior for $\boldsymbol{Z} \sim \text{IBP}(\alpha)$

$$p(\boldsymbol{Z}) = \frac{\alpha^{C_+}\exp\{-\alpha H_S\}}{C_+!} \prod_{c=1}^{C_+} \frac{\Gamma(m_c)\Gamma(S-m_c+1)}{\Gamma(S+1)} \tag{3}$$

where C_+ is the number of non-empty features (columns), $H_S = \sum_{s=1}^{S} 1/s$ is the S-th harmonic number, and the columns of $\boldsymbol{Z}$ are uniformly ordered. Note that the rows of $\boldsymbol{Z}$ are exchangeable as the right-hand side of Eq. (3) does not depend

on the row indices of $\boldsymbol{Z}$. The name IBP originates from a description of customers entering an Indian buffet restaurant with infinitely many dishes. The first customer (SNV) chooses a Poisson(α) number of dishes (haplotypes). The s-th customer takes dish c with probability m_c/s, with m_c being the number of customers who have tried dish c before, and then tries a Poisson(α/s) number of new dishes. The choices of all customers are recorded in matrix $\boldsymbol{Z}$ where $Z_{sc} = 1$ if the sth customer took the cth dish. Then the matrix $\boldsymbol{Z}$ is said to follow an IBP, and the probability of $\boldsymbol{Z}$ is given by (3) after randomly permuting the columns of $\boldsymbol{Z}$.

Due to the exchangeability of the IBP prior in (3), the conditional distribution $P(Z_{sc} = 1|\boldsymbol{Z}_{-s,c})$ is the same as if $s = S$ were the last customer. That is, $P(Z_{sc} = 1|\boldsymbol{Z}_{-s,c}) = m_{-s,c}/S$ provided $m_{-s,c} > 0$ where $\boldsymbol{Z}_{-s,c}$ is the cth column of $\boldsymbol{Z}$ excluding sth row, $m_{-s,c}$ is the number of 1's in $\boldsymbol{Z}_{-s,c}$, and the distribution of the number of new features (haplotypes) for each row (SNV) is Poisson(α/S).

Conditioning on the number C of features, we assign a Dirichlet distribution to the weights $\boldsymbol{w} = (w_0, w_1, \ldots, w_C) \sim \text{Dir}(a_0, a, \ldots, a)$ with $a_0 < a$ to reflect the prior belief that the proportion of background noise is smaller than the proportion of the haplotypes. Equivalently, each w_c can be written as $w_c = \theta_c / \sum_{c=0}^{C} \theta_c$ where $\theta_0 \sim \text{Gamma}(a_0, 1)$, $\theta_c \sim \text{Gamma}(a, 1)$ for $c = 1, \ldots, C$. The gamma representation of Dirichlet distribution is adopted in our sampling algorithm as described in Sect. 3. Finally, the background SNV frequency ρ is given a beta prior, $\rho \sim \text{Beta}(a_\rho, b_\rho)$. We set $a_\rho \ll b_\rho$ as we expect ρ to be very small.

3 Posterior Inference and Parallel Tempering

The posterior distribution of the model described in Sect. 2 is given by

$$p(\boldsymbol{Z}, \boldsymbol{\theta}, \rho|\boldsymbol{n}) \propto p(\boldsymbol{n}|\boldsymbol{N}, \boldsymbol{p})\, p(\boldsymbol{Z})\, p(\boldsymbol{\theta}|\boldsymbol{Z})\, p(\rho)$$

where the weights $\boldsymbol{w}$ are replaced by the unnormalized weights $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_C)$. Since the posterior is analytically intractable, we use Markov chain Monte Carlo (MCMC) simulation to generate a Monte Carlo sample of $(\boldsymbol{Z}, \boldsymbol{w}, \rho)$ from the posterior. Let $\boldsymbol{p} = (\boldsymbol{Z}, \boldsymbol{\theta}, \rho)$ and let $\boldsymbol{p}^{(i)}$ denote the state of the Markov chain at the ith iteration. Define the following one-step operator (that is, the transition probability of the Markov chain), which takes the current state and data as input, and outputs the state for the next step. In the description below we use "new features" to refer for each customer s to the set of features c that are only selected by customer s.

Operator $\boldsymbol{p}^* = \mathscr{S}(\boldsymbol{p}, \boldsymbol{N}, \boldsymbol{n})$

(1) Update $\boldsymbol{Z}$. We iterate through each row s of $\boldsymbol{Z}$.

(1a) Update existing features. Sample Z_{sc} from its full conditional

$$P(Z_{sc} = 1|\cdot) \propto P(Z_{sc} = 1|\boldsymbol{Z}_{-s,c})p(n_s|N_s, p_s) = \frac{m_{-s,c}}{S}\binom{N_s}{n_s}p_s^{n_s}(1-p_s)^{N_s-n_s},$$

for $c = 1, \ldots, C$ where C is the current number of features.

(1b) Update new features (that is, features that are only selected by customer s). We propose $C^* = \text{Poisson}(\alpha/S)$ new features and for each new feature we propose a new feature-specific parameter θ_c^* from its prior $\theta_c^* \sim \text{Gamma}(a, 1)$ for $c = C+1, \ldots, C+C^*$. We accept the new features and the feature-specific parameters, with probability $\min\{1, r\}$, where the Metropolis-Hasting (M-H) ratio

$$r = \frac{p(\boldsymbol{n}|\boldsymbol{N}, \boldsymbol{p}^*)}{p(\boldsymbol{n}|\boldsymbol{N}, \boldsymbol{p})}$$

reduces to the likelihood ratio because the prior ratio and proposal ratio are canceled out.

(2) Update $\boldsymbol{\theta}$. For $c = 0, \ldots, C$, we propose θ_c^* from a proposal density $q(\theta_c^*|\theta_c)$, and accept it with probability $\min\{1, r\}$ where

$$r = \frac{p(\boldsymbol{n}|\boldsymbol{N}, \boldsymbol{p}^*)p(\theta_c^*|\boldsymbol{Z})q(\theta_c|\theta_c^*)}{p(\boldsymbol{n}|\boldsymbol{N}, \boldsymbol{p})p(\theta_c|\boldsymbol{Z})q(\theta_c^*|\theta_c)}.$$

(3) Update ρ. We propose ρ from a proposal density $q(\rho^*|\rho)$, and accept it with probability $\min\{1, r\}$ where

$$r = \frac{p(\boldsymbol{n}|\boldsymbol{N}, \boldsymbol{p}^*)p(\rho^*)q(\rho|\rho^*)}{p(\boldsymbol{n}|\boldsymbol{N}, \boldsymbol{p})p(\rho)q(\rho^*|\rho)}.$$

Then our Metropolis-within-Gibbs algorithm proceeds by iteratively calling the operator $\boldsymbol{p}^* = \mathscr{S}(\boldsymbol{p}, \boldsymbol{N}, \boldsymbol{n})$.

Algorithm: MCMC

(I) Initialize $\boldsymbol{p}^{(0)} = (\boldsymbol{Z}^{(0)}, \boldsymbol{\theta}^{(0)}, \rho^{(0)})$

(II) Iteratively apply the operator $\boldsymbol{p}^{(i)} = \mathscr{S}(\boldsymbol{p}^{(i-1)}, \boldsymbol{N}, \boldsymbol{n})$ until convergence.

The above MCMC only works for a small to moderate total number N_s of reads. When N_s is large (up to 4437 in our application), the likelihood $p(n_s|N_s, p_s)$ as a function of p_s (hence the posterior distribution) is extremely concentrated at the empirical VAF $\hat{f}_s = n_s/N_s$, which results in effectively zero acceptance rate for the M-H steps (1b), (2) and (3). For this reason, the previous proposed approaches [9, 12] are not suitable for our data. In their applications, N_s is typically around 50.

We instead implement a variation of the *parallel tempering* technique, also known as Metropolis-coupled MCMC [2]. In particular, we run L parallel Markov chains with different target distributions $\pi_1, \ldots, \pi_L$. The chains are "heated" by raising the likelihood to a power $1/T_\ell$ for $T_\ell \geq 1$, creating a target distribution

$$\pi_\ell \propto \prod_{s=1}^{S} p(n_s|N_s, p_s)^{\frac{1}{T_\ell}} p(\boldsymbol{Z})p(\boldsymbol{\theta}|\boldsymbol{Z})p(\rho)$$

for $\ell = 1, \ldots, L$. The first chain is the "cold" chain with the lowest temperature $T_1 = 1$ and its target distribution is our desired posterior distribution $\pi_1 = p(\boldsymbol{Z}, \boldsymbol{\theta}, \rho|\boldsymbol{n})$. The "heated" chain has a less peaked likelihood, which makes it easier to traverse the parameter space. Under the binomial sampling model (1), the power transformation of the likelihood has a direct interpretation as using only a fraction $(n_s/T, N_s/T)$ of the data

$$p(n_s|N_s, p_s)^{\frac{1}{T}} \propto p\left(\frac{n_s}{T}\middle|\frac{N_s}{T}, p_s\right).$$

With a smaller total number of reads (while keeping the empirical VAF unchanged), the posterior distribution becomes flatter, and consequently, the heated chain tends to accept more proposals than a cold chain. In order for the cold chain to mix better, we couple it with the heated chains by swapping the states between two randomly selected chains with a predefined rate P_{swap}. The swapping is then accepted/rejected according to a Metropolis-Hasting ratio.

Algorithm: Parallel-tempered MCMC
(I) Initialize $\boldsymbol{p}_\ell^{(0)} = (\boldsymbol{Z}_\ell^{(0)}, \boldsymbol{\theta}_\ell^{(0)}, \rho_\ell^{(0)})$ for $\ell = 1, \ldots, L$ where the subscript ℓ is the chain indicator.
(II) At each iteration i

(a) With probability P_{swap}, propose a swapping move. Randomly select two chains ℓ and m and swap the states of the two chains with probability $\min\{1, r\}$ where

$$r = \frac{p(\boldsymbol{n}_\ell|\boldsymbol{N}_\ell, \boldsymbol{p}_m^{(i-1)})p(\boldsymbol{n}_m|\boldsymbol{N}_m, \boldsymbol{p}_\ell^{(i-1)})}{p(\boldsymbol{n}_\ell|\boldsymbol{N}_\ell, \boldsymbol{p}_\ell^{(i-1)})p(\boldsymbol{n}_m|\boldsymbol{N}_m, \boldsymbol{p}_m^{(i-1)})},$$

with $\boldsymbol{n}_\ell = \boldsymbol{n}/T_\ell$ and $\boldsymbol{N}_\ell = \boldsymbol{N}/T_\ell$ and similar definition for $\boldsymbol{n}_m$ and $\boldsymbol{N}_m$.

(b) Perform a one-step update for all chains: $\boldsymbol{p}_\ell^{(i)} = \mathscr{S}(\boldsymbol{p}_\ell^{(i-1)}, \boldsymbol{N}_\ell, \boldsymbol{n}_\ell)$ for $\ell = 1, \ldots, L$.
(III) Repeat step (II) until convergence.
Each individual chain is no longer Markov, because of the dependencies across the chains, introduced by swapping. However, the validity of the algorithm can be understood by viewing all the chains as a joint stochastic process which is Markov; technical details can be found in [2].

To summarize the posterior distribution $p(\boldsymbol{Z}, \boldsymbol{w}, \rho|\boldsymbol{n}) = p(\boldsymbol{w}, \rho|\boldsymbol{n}, \boldsymbol{Z})p(\boldsymbol{Z}|\boldsymbol{n}, C)$ $p(C|\boldsymbol{n})$ using MCMC samples, we proceed by first calculating the maximum a posteriori (MAP) estimator $\hat{C}$ from the marginal posterior distribution of C. Conditional on $\hat{C}$, we find the least squares feature allocation estimator [13] $\hat{\boldsymbol{Z}}$ by the following procedure. For any two binary matrices $\boldsymbol{Z}, \boldsymbol{Z}' \in \{0, 1\}^{n\times\hat{C}}$, we define the distance $d(\boldsymbol{Z}, \boldsymbol{Z}') = \min_\pi \mathscr{H}(\boldsymbol{Z}, \pi(\boldsymbol{Z}'))$ where $\pi(\boldsymbol{Z}')$ denotes a permutation of the columns of $\boldsymbol{Z}'$ and $\mathscr{H}(\cdot, \cdot)$ is the Hamming distance of two binary matrices. The point estimator $\hat{\boldsymbol{Z}}$ is then obtained by

$$\hat{\boldsymbol{Z}} = \arg\min_{\boldsymbol{Z}'} \int d(\boldsymbol{Z}, \boldsymbol{Z}')dp(\boldsymbol{Z}|\boldsymbol{n}, \hat{C}),$$

which can be approximated by MCMC samples. The posterior point estimators $\hat{\boldsymbol{w}}$ and $\hat{\rho}$ are then computed by $(\hat{\boldsymbol{w}}, \hat{\rho}) = E(\boldsymbol{w}, \rho|\boldsymbol{n}, \hat{\boldsymbol{Z}})$ conditional on $\hat{\boldsymbol{Z}}$.

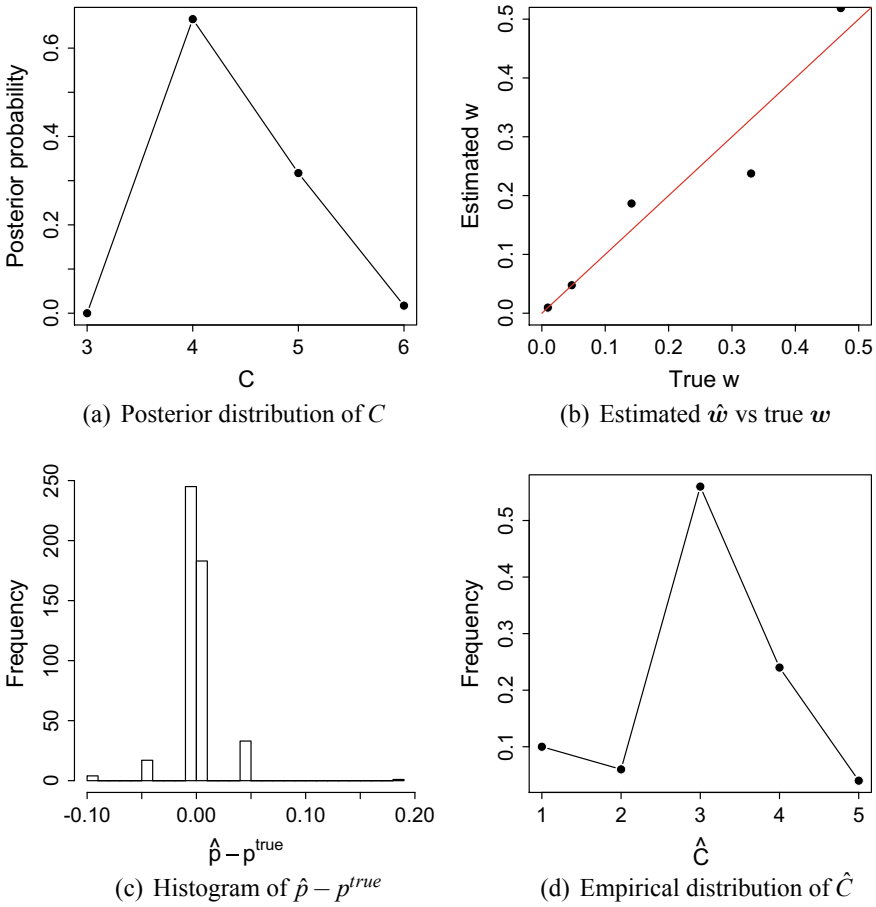
4 Simulations

In this section, we present two simulation studies to assess the performance of the proposed approach. In the first scenario, we compare PTFA with MAD-Bayes using a similar simulation setting as in [12]. We let $\boldsymbol{N}$ be the observed total number of reads from the Glioblastoma multiforme (GBM) data in Sect. 5 which has $S = 483$ SNVs. We assume $C = 4$ haplotypes where each haplotype has the following configuration: haplotype c has variant sequences at the first $100 \cdot c$ SNV positions for $c = 1, \ldots, 4$. The true feature allocation matrix $\boldsymbol{Z}$ is shown in Fig. 2a. We set the true weights $\boldsymbol{w} \propto (0.2, 10, 7, 3, 1)$ and $\rho = 0.01$. Then we generate $n_s \sim \text{Bin}(N_s, p_s^{true})$ where $p_s^{true} = w_0\rho + \sum_{c=1}^{C} w_c Z_{sc}$ for $s = 1, \ldots, S$.

The hyperparameters of our model are specified as: $\alpha = 1, a_0 = 0.1, a = 0.5$, $a_\rho = 1$ and $b_\rho = 100$. We run $L = 10$ parallel chains, each with 50,000 iterations, and set the probability of swapping $P_{swap} = 0.3$. The temperatures are chosen to be $T_\ell = S^{\frac{\ell-1}{L-1}}$ for $\ell = 1, \ldots, L$. Only the cold chain is retained for subsequent analysis. We discard the first 50% of the iterations as burn-in, and thin the chain by taking every 10th sample.

The posterior distribution of the number C of haplotypes is displayed in Fig. 1a, which is peaked at the simulation truth, $\hat{C} = 4$. Conditional on $\hat{C}$, the (posterior) point estimator $\hat{\boldsymbol{Z}}$ is shown in Fig. 2b with mis-allocation rate $\mathscr{H}(\hat{\boldsymbol{Z}}, \boldsymbol{Z})/S/C = 13\%$. Conditional on $\hat{\boldsymbol{Z}}$, the point estimator $\hat{\boldsymbol{w}}$ is plotted against the true $\boldsymbol{w}$ in Fig. 1b. The estimation works well, as can be seen from the very close fit of points to the diagonal line. This can also be seen from the histogram of $\hat{p} - p^{true}$ in Fig. 1c where $\hat{p} = \hat{w}_0\hat{\rho} + \sum_{c=1}^{\hat{C}} \hat{w}_c \hat{Z}_{sc}$.

For comparison, we apply MAD-Bayes to the same simulated data. MAD-Bayes has a tuning parameter λ^2 which penalizes the number of columns in $\boldsymbol{Z}$, with smaller λ^2 implying larger $\hat{C}$. We run the algorithm with 50 different initializations for a range of $\lambda^2 = \{2, 4, 6, 8, 10, 20, 200, 500\}$, recommended by [12]. The best fit is obtained $\lambda^2 = 2$ and the frequency of the estimated $\hat{C}$ across 50 simulations is shown in Fig. 1d. Even with a small penalty, MAD-Bayes tends to select 3 haplotypes (28 out of 50 simulations); in only 12 out of 50 simulations does MAD-Bayes correctly identify the correct number of haplotypes. A typical estimated $\hat{\boldsymbol{Z}}$ with $\hat{C} = 4$ from MAD-Bayes is shown in Fig. 2c with mis-allocation rate 14%. Since MAD-Bayes tends to underestimate C, we alter the simulation truth by removing the first and the third column of $\boldsymbol{Z}$ (therefore $C = 2$) and rerun PTFA and MAD-Bayes. PTFA has a perfect recovery of $\boldsymbol{Z}$ with mis-allocation rate 0%, due to the simplified scenario. For the same reason, the performance of MAD-Bayes is also improved: it correctly estimates C and $\boldsymbol{Z}$ in 32 out of 50 simulations.

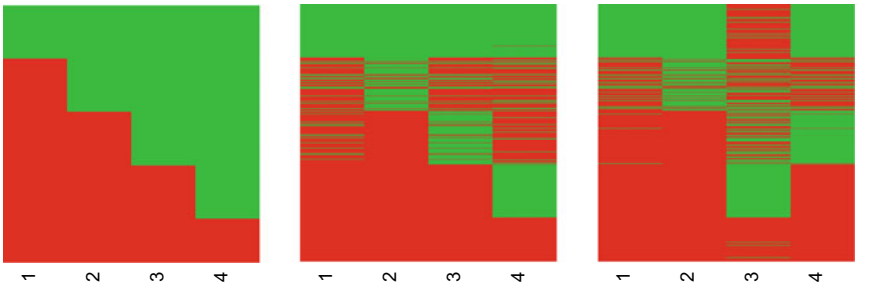


(a) Posterior distribution of C (b) Estimated $\hat{\boldsymbol{w}}$ vs true $\boldsymbol{w}$

(c) Histogram of $\hat{p} - p^{true}$ (d) Empirical distribution of $\hat{C}$

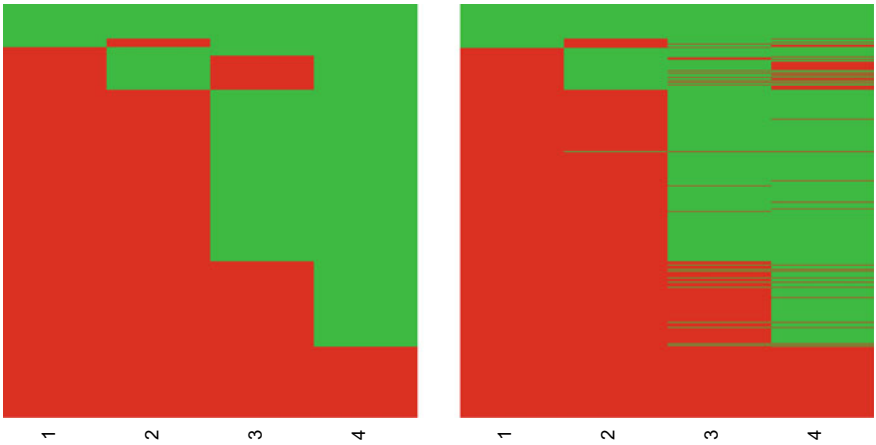
Fig. 1 Simulation results of PTFA and MAD-Bayes in the first scenario. **a** Posterior distribution of C for PTFA. **b** Scatter plot of estimated $\hat{\boldsymbol{w}}$ versus true $\boldsymbol{w}$ for PTFA. **c** Histogram of of $\hat{p} - p^{true}$ for PTFA. **d** Empirical distribution of $\hat{C}$ across 50 runs of MAD-Bayes with $\lambda^2 = 2$

In our second simulation, we consider the scenario where the VAFs are similar to those in the GBM data. Specifically, we obtain the new matrix $\boldsymbol{Z}$ (shown in Fig. 3a) by altering the earlier $\boldsymbol{Z}$: (1) remove the last 50 1's from the second column; (2) remove the first 50 and last 100 1's from the third column; and (3) remove the first 100 1's from the fourth column. We keep other simulation settings unchanged. The proportion of the mutations with VAF > 0.5 is now approximately 10%. We find similar results as in the first scenario. For example, the estimated $\hat{\boldsymbol{Z}}$ is shown in Fig. 3b and the mis-allocation rate is 4%. We also run the same algorithm with different initial values and we don't observe any significant difference. In addition, we investigate how PTFA performs when read depth is 10-fold shallower. We find $\hat{C} = 6$ haplotypes with two haplotypes having only 1 and 3 mutated loci (out of 483 loci).



(a) True Z (b) Estimated $\hat{Z}$ from PTFA (c) Estimated $\hat{Z}$ from MAD-Bayes

Fig. 2 Simulation results of PTFA and MAD-Bayes in the first scenario. Heatmaps of the feature allocation matrix $\boldsymbol{Z}$, where 1 is represented by the color, green and 0 by the color, red. Panel **a** displays the true $\boldsymbol{Z}$. Panel **b** displays the estimated $\hat{\boldsymbol{Z}}$ from PTFA. Panel **c** displays the estimated $\hat{\boldsymbol{Z}}$ from MAD-Bayes



(a) True Z (b) Estimated $\hat{Z}$

Fig. 3 Simulation results of PTFA in the second scenario. Heatmaps of the feature allocation matrix $\boldsymbol{Z}$, where 1 is represented by the color, green and 0 by the color, red. Panel **a** displays the true $\boldsymbol{Z}$. Panel **b** displays the estimated $\hat{\boldsymbol{Z}}$

5 GBM Data Analysis

Glioblastoma multiforme (GBM) is the most common and aggressive form of primary brain cancer in human adults with poor prognosis and a lack of effective therapeutic options. Many recent studies [14] suggest that intra-tumor heterogeneity is crucial to understanding treatment failure, due to the existence of subclones that resist conventional therapies.

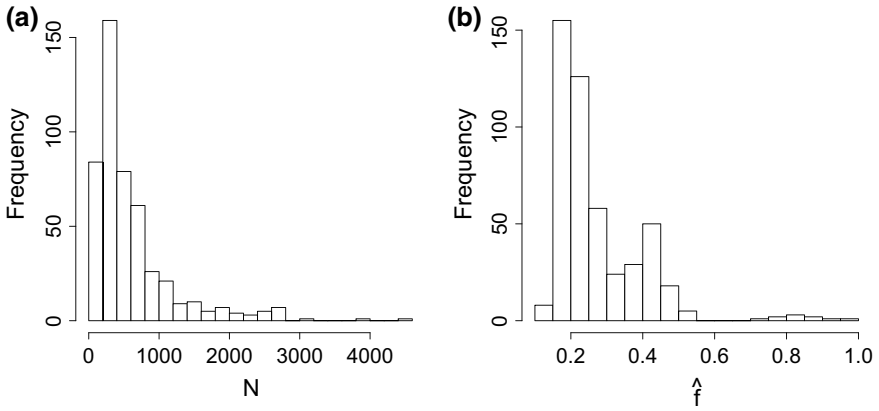


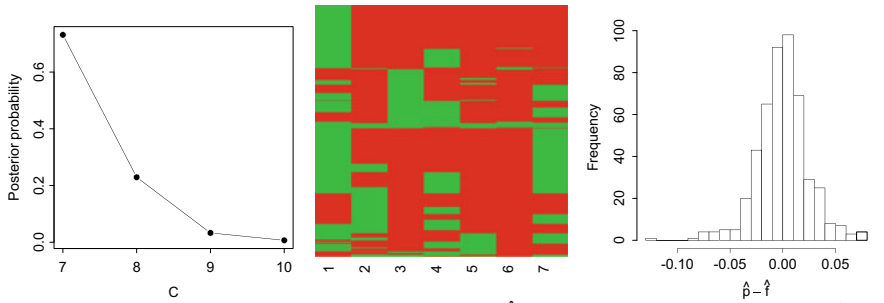
Fig. 4 GBM data summaries. **a** Histogram of the total number of mapped reads N_s. **b** Histogram of the empirical VAF $\hat{f}_s = n_s/Ns$

In this study, frozen GBM tumor tissue and matched blood samples were obtained from the Austin Brain Tumor Repository at St. David's Medical Center. Exome extraction and sequencing were performed at the Genomic Sequencing and Analysis Facility (GSAF) at the University of Texas at Austin. Exome extraction was done using Agilent SureSelect V5+UTR exome-capture kits. DNA sequencing was performed on the Sequenced Illumina HiSeq 25000 NGS platform, with 2×125 bp paired-end reads, and 2×10^7 reads per sample, for an average exome coverage of 500x. Low-quality reads were removed using Samtools, and sequence reads were mapped to the reference human genome hg19, using BWA. Base-quality recalibration, indel calling/realignment, and variant calling/annotation was performed using GATK [15–17]. The resulting dataset consists of a total of $S = 483$ SNVs. We summarize the data in Fig. 4 with a histogram of the total number of mapped reads N_s and a histogram of the empirical VAF $\hat{f}_s = n_s/Ns$ for $s = 1, \ldots, 483$. Deep sequencing technology identifies, not only common clonal mutations, but also rarer variants ($\hat{f}_s \approx 10\%$), which might be unique to particular subclonal lineages. However, the likelihood is essentially a point mass at $\hat{f}_s$, since the total number N_s of reads is very large.

Using a bioinformatics tool such as [18, 19], we estimate the normal contamination to be $\tau = 0.17$, which is incorporated in our model by modifying Eq. (2) as

$$\tilde{p}_s = (1-\tau)p_s = (1-\tau)(w_0\rho + \sum_{c=1}^{C} w_c Z_{sc})$$

and the likelihood (1) as



(a) Posterior distribution of C (b) Estimated $\hat{\boldsymbol{Z}}$ (c) Histogram of $\hat{p} - \hat{f}$

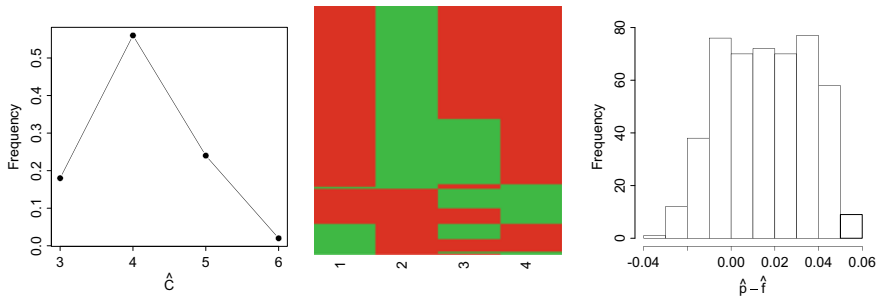
Fig. 5 GBM data analysis using PTFA. **a** Posterior distribution of C. **b** Heatmap of the feature allocation matrix $\boldsymbol{Z}$, where 1 is represented by the color, green and 0 by the color, red. **c** Histogram of of $\hat{p} - \hat{f}$

$$p(\boldsymbol{n}|\boldsymbol{N}, \tilde{\boldsymbol{p}}) = \prod_{s=1}^{S} p(n_s|N_s, \tilde{p}_s) = \prod_{s=1}^{S} \binom{N_s}{n_s} \tilde{p}_s^{n_s} (1 - \tilde{p}_s)^{N_s - n_s}.$$

We apply PTFA using the same specifications as in Sect. 4. The algorithm (implemented in R) takes ∼11 h on a 3.5 GHz Intel Core i7 CPU with 16 GB memory. We show the posterior distribution of C in Fig. 5a and find the posterior mode for $\hat{C} = 7$ haplotypes. Given $\hat{C}$, the posterior point estimator of feature allocation matrix $\hat{\boldsymbol{Z}}$ is provided in Fig. 5b, with green and red indicating mutant and wildtype genotypes, respectively. As before, the rows are SNVs, and the columns are haplotypes.

The extent of intra-tumor heterogeneity is evident from the heatmap. For example, haplotype 1 is characterized by a large number of mutations while haplotype 6 has far fewer mutations. The haplotype proportions are estimated to be $\hat{\boldsymbol{w}} = (0.17, 0.18, 0.19, 0.05, 0.10, 0.11, 0.03)$. The first three haplotypes dominate the subclone distribution, and they differ from one another at multiple loci. To assess the fit of PTFA to the data, we plot the histogram of $\hat{p}_s - \hat{f}_s$ for $s = 1, \ldots, 483$ in Fig. 5c where $\hat{p}_s = \hat{w}_0\hat{\rho} + \sum_{c=1}^{\hat{C}} \hat{w}_c \hat{Z}_{sc}$ and $\hat{f}_s = n_s/N_s$. With the histogram being concentrated around zero the model fit appears to be adequate.

For comparison, we ran MAD-Bayes 50 times (∼2 min) using this data, with $\lambda^2 = 2$, as it yields the best performance in simulation studies, and imposes small penalization on the number of haplotypes. The results are shown in Fig. 6. The empirical distribution of $\hat{C}$ has a peak at 4 (Fig. 6a). Although it is hard to judge which estimate is closer to the true number C of haplotypes, it seems that MAD-Bayes is inefficient in exploring new haplotypes, possibly due to large read depth, which is also seen in simulation studies. We randomly choose one result from 50 MAD-Bayes runs for which $\hat{C} = 4$, and plot the heatmap of $\hat{\boldsymbol{Z}}$ in Fig. 6b and the histogram of $\hat{p} - \hat{f}$ in Fig. 6c. The latter shows some evidence for a worse fit than under PTFA.



(a) Empirical distribution of $\hat{C}$ (b) Estimated $\hat{\boldsymbol{Z}}$ (c) Histogram of $\hat{p} - \hat{f}$

Fig. 6 GBM data analysis, using MAD-Bayes with $\lambda^2 = 2$. **a** Empirical distribution of $\hat{C}$ across 50 runs. **b** Heatmap of the feature allocation matrix $\boldsymbol{Z}$, where 1 is represented by the color, green and 0 by the color, red. **c** Histogram of of $\hat{p} - \hat{f}$

6 Discussion

Characterizing tumor heterogeneity through the identification of subclones is key to understanding cancer genomics, and has the potential to provide vital information in developing personalized treatments. Deep sequencing technology generates more informative data from which to infer subclones, while simultaneously posing a computational challenge, due to peaked likelihood associated with sample means. In this paper, we have developed a parallel-tempered feature allocation algorithm (PTFA) to overcome these difficulties. Simulation studies show PTFA is superior to competing methods when read depth is large. PTFA's applicability was demonstrated by identifying subclones using GBM genomes sequenced at high read depth. Future data analysis will build upon our empirical results by investigating the phylogenetic history of subclonal lineages in GBM and applying tests of neutral evolution versus subclonal selection to the haplotype frequency distribution. We discussed inference for tumor heterogeneity. A similar setup and inference approach could potentially be used with other experiments that use NGS data, including possibly microbiome data if the data include samples from multiple tissue types for each individual. See, for example, [20] for a related discussion.

One drawback of the proposed method is its scalability compared to optimization-based approaches such as MAD-Bayes (which is hundreds times faster). The running speed can be greatly improved by implementing it in C++. However, more clever algorithm is needed if one wants to consider genome-wide sequencing data. Methodologically, in our future work, we will allow misspecification in the feature allocation model. For example, instead of assuming a deterministic relationship in Eq. (2), we could assign a beta prior on p_s which centers on $w_0\rho + \sum_{c=1}^{C} w_c Z_{sc}$. The additional degree of freedom in the beta prior would strengthen the robustness of our method against model misspecification.

Acknowledgements YN, YJ and PM were partially funded by grant NIH R01 CA132891-06A1. MS was supported by the St. David's Foundation impact fund. Specimen collection, processing and analysis were supported by funds from the St. David's Impact Fund and the NeuroTexas Research Foundation.

References

1. Griffiths, T.L., Ghahramani, Z.: Infinite latent feature models and the indian buffet process. NIPS **18**, 475–482 (2005)
2. Geyer, C.J.: Markov chain Monte Carlo maximum likelihood. In: Proceedings of the 23rd Symposium on the Interface, Computing Science and Statistics. Interface Foundation, Fairfax Station, VA (1991)
3. Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al.: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. **2012**(366), 883–892 (2012)
4. Seoane, J., Mattos-Arruda, D., et al.: The challenge of intratumour heterogeneity in precision medicine. J. Intern. Med. **276**(1), 41–51 (2014)
5. De Bono, J., Ashworth, A.: Translating cancer research into targeted therapeutics. Nature **467**(7315), 543–549 (2010)
6. Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J.M., Desrichard, A., Walsh, L.A., Postow, M.A., Wong, P., Ho, T.S., et al.: Genetic basis for clinical response to ctla-4 blockade in melanoma. N. Engl. J. Med. **371**(23), 2189–2199 (2014)
7. Campbell, P.J., Pleasance, E.D., Stephens, P.J., Dicks, E., Rance, R., Goodhead, I., Follows, G.A., et al.: Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc. Natl. Acad. Sci. U. S. A. **105**(35), 13,081–13,086 (2008)
8. Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., et al.: Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. Proc. Natl. Acad. Sci. **112**(47), E6496–E6505 (2015)
9. Lee, J., Müller, P., Gulukota, K., Ji, Y., et al.: A bayesian feature allocation model for tumor heterogeneity. Ann/ Appl. Stat. **9**(2), 621–639 (2015)
10. Green, P.J.: Reversible jump markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**(4), 711–732 (1995)
11. O'Hagan, A.: Fractional Bayes factors for model comparison. J. R. Stat. Soc. Series B **57**(1), 99–138 (1995)
12. Xu, Y., Müller, P., Yuan, Y., Gulukota, K., Ji, Y.: Mad bayes for tumor heterogeneity-feature allocation with exponential family sampling. J. Am. Stat. Assoc. **110**(510), 503–514 (2015)
13. Dahl, D.B.: Model-based clustering for expression data via a Dirichlet process mixture model. In: Bayesian Inference for Gene Expression and Proteomics, pp. 201–218 (2006)
14. Sottoriva, A., Spiteri, I., Piccirillo, S.G., Touloumis, A., Collins, V.P., Marioni, J.C., Curtis, C., Watts, C., Tavaré, S.: Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proc. Natl. Acad. Sci. **110**(10), 4009–4014 (2013)
15. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al.: The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome Res. **20**(9), 1297–1303 (2010)
16. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al.: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. **43**(5), 491–498 (2011)
17. Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al.: From FASTQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. In: Current Protocols in Bioinformatics, pp. 11.10.1–11.10.33 (2013)

18. Qiao, W., Quon, G., Csaszar, E., Yu, M., Morris, Q., Zandstra, P.W.: Pert: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS Comput. Biol. **8**(12), e1002, 838 (2012)
19. Ahn, J., Yuan, Y., Parmigiani, G., Suraokar, M.B., Diao, L., Wistuba, I.I., Wang, W.: Demix: deconvolution for mixed cancer transcriptomes using raw measured data. Bioinformatics (2013)
20. Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C., Trippa, L.: Bayesian nonparametric mixed effects models in microbiome data analysis. arXiv:1711.01241 (2017)

Analysis of T-Cell Immune Responses as Measured by Intracellular Cytokine Staining with Application to Vaccine Clinical Trials

Yunzhi Lin and Cong Han

Abstract Recent advances in single-cell technologies, in particular intracellular cytokine staining (ICS), have enabled multidimensional functional measurements of naturally occurring or vaccine-induced T-cell responses in clinical studies. Analysis of such increasingly multidimensional datasets presents a great challenge to statisticians. Currently, multidimensional functional cell measures are largely analyzed, either by univariate analysis of all combinations of functions individually, or by summarizing a few particular groups of functions separately. Such simple analyses do not reflect comprehensively the polyfunctional profile of the T-cell responses, nor do they allow more sophisticated statistical analysis and inference. In this paper, we introduce a new approach to statistical inference for multidimensional ICS data. We propose to reduce the dimensionality by using a weighted sum, followed by computing the minimum and maximum of the test statistic over all eligible assignments of weights which satisfy the underlying partial ordering of the data. The computation technique is presented. Statistical inference is then based on the minimum and maximum of the test statistic. We illustrate, through an example, that the technique can be useful in reducing the complexity of the multidimensional response data and providing insightful reporting of the results.

Keywords Cell-mediated immunity · Intracellular cytokine staining · Vaccine · Clinical trials · Partial ordering · Min max statistics · Stochastic ordering

1 Introduction

Evaluation of vaccine-induced immunity is essential in vaccine clinical trials in understanding and establishing the immunological basis of the efficacy of test vaccines. Immunological endpoints in vaccine trials are classified by the type of adaptive immune responses invoked—humoral or cellular—and are measured by various

Y. Lin (✉) · C. Han
Takeda Pharmaceutical Company Limited, Cambridge, MA, USA
e-mail: stella.lin@takeda.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_18

immunologic assays at a select few time points. Because of the complexity of the immune system, the majority of the vaccine trials to-date have measured a relatively small number of variables, mainly humoral immune responses measured by antibody titers or antibody concentrations. However, recent advances in technology enable many parameters to be measured quickly and at low cost, allowing much broader immune profiling. In particular, single-cell level intracellular cytokine staining (ICS), an application of flow cytometry, has become an important method for determining T-cell immunogenicity in recent vaccine trials.

Intracellular cytokine staining is a process in which the secretion of multiple functional molecules called cytokines (e.g., interferon (IFN)-γ, tumor necrosis factor (TNF)-α, interleukin (IL)-2, IL-21) is measured simultaneously for individual T cells while the cells pass through the machine in a thin fluid stream. Typically, cell-level measurements are thresholded so that each cell will be classified as positive (+) or negative (−) in expressing each function (cytokine); e.g. IFN-γ+ TNF-α− IL-2+ IL-21+. An ICS experiment measuring K functions therefore defines $2^K - 1$ Boolean combinations of functions (leaving out the category of "all negative" in order to focus on the responding cells), and the set of responses in a sample of blood is characterized by summarizing the frequencies of expression profiles corresponding to the $2^K - 1$ different combinations of functions (Fig. 1).

Recent works have highlighted the importance of such multidimensional functional T-cell assessment in vaccine immunogenicity and efficacy studies. The presence of antigen-specific T cells capable of simultaneously expressing multiple cytokines, known as polyfunctional T cells, has been shown to correlate with favorable clinical outcomes [1–3]. Effective vaccination is expected to generate broad T-cell cytokine expressions [4, 5], making them attractive as potential markers of protection. However, statistical tools for analyzing this complex multidimensional data have been lacking.

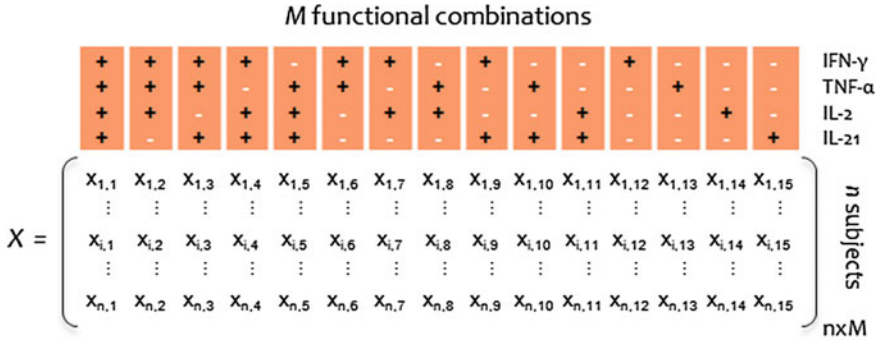


Fig. 1 Illustration of ICS T-cell response data collected from n subjects in an experimental group. Data are number of antigen-specific phenotypic (i.e., CD4+, CD8+) T cells expressing different combinatorial functional cytokines (e.g., the numbers of antigen-specific CD4+ T cells expressing different combinations of cytokines). For K cytokines, this produces an n by M matrix of counts, where $M = 2^K - 1$

Very few tools have been developed specifically for multidimensional T-cell data analysis. To date, most clinical studies have been limited to reporting univariate results, such as summaries and comparisons of individual function expressions (e.g,. IFN-γ, marginalizing over other functions), analyses of "polyfunctionality" (i.e., cells expressing a specified number of functions), or laborious analyses of all functional combinations individually. Simple graphics such as bar plots and pie charts are used to display the summary statistics [6]. The problems with such univariate approaches are that they cannot evaluate T-cell responses as an entirety, ignore the dependency between combinations, and lead to multiple testing problems. These separate analyses require heavy human interpretation into identifying a distinct vaccine-induced T-cell response, e.g., by subjectively applying weights of importance to individual functions or combinations, and this becomes particularly problematic as the number of combinations grows exponentially with the number of cytokines analyzed. A few global statistics have been proposed to address some of these limitations. Nason discussed using Hotelling's T^2 statistic [7, 8], which tests globally whether the treatment groups differ on any of the combinations. However, the test does not differentiate the relative importance of the combinations, nor does it differentiate the direction of the differences. Larsen et al. introduced a polyfunctionality index (PI) that reduces the multidimensional profile into a weighted one-dimensional index value [9]. Although such a reductionist approach inevitably leads to a loss of information, it largely benefits from the numerous analytical tools exclusively compatible with one-dimensional values. The PI however, uses an arbitrary selection of weights which does not discriminate among different cytokines, and thus falls short of describing the true profile of functional T-cell responses, limiting its clinical utility. A few Bayesian frameworks for multidimensional analysis have been proposed, including the mixture models for single-cell analysis (MIMOSA) introduced by Finak et al. [10] and the unbiased combinatorial polyfunctionality analysis of antigen-specific T-cell subsets (COMPASS) proposed by Lin et al. [11].

In vaccine trials the common statistical problem is to assess whether or not there is a difference between the treatment groups in functional T-cell responses. Furthermore, an ideal framework for the analysis should quantify both the magnitude (i.e., amount of cells expressing the cytokines or combinations of cytokines of interest) and the quality (i.e., profiles of the polyfunctional coexpression) of T-cell responses, and thus permits a comprehensive comparison between the treatment groups. To address these needs, we introduce a new approach to statistical inference for multidimensional ICS data. We consider a "generalized cytokine production index" which, similar to the PI, reduces the dimensionality by using a weighted sum. Rather than assigning a fixed set of weights, we allow all possible weights to permit a comprehensive analysis. We further specify that the weights need to follow an underlying partial ordering such that the cells expressing more functions (i.e., more polyfunctional combinations) will be preferred. The resulting generalized index thus quantifies not only the magnitude of response by simply summing up the numbers of cells expressing each combination, but also the quality of response by giving more weight to more polyfunctional responses.

The one-dimensional index enables easy statistical comparison (e.g., t test) of the treatment groups. To allow for the comprehensive analysis, we propose basing statistical inference on the minimum and the maximum of the test statistic over all possible assignments of weights, and present a solution to the problem of finding the minimum (min) and the maximum (max) of the statistics. If the range of the min and max statistics does not include the critical value, then the significance (or non-significance) of the result can immediately be concluded regardless of the choice of weights. On the other hand, if the range includes the critical value, the choice of weights and the corresponding conclusions must be carefully justified.

Our paper is organized as follows. The proposed method is described in Sect. 2 along with the computational approach to find the minimum and maximum of standard two-sample test statistics over all eligible assignments of weights. Section 3 provides an example of T-cell response data collected in a vaccine trial with further details on the multidimensional ICS data. The application of the method is illustrated using the example data. Discussion and conclusions are presented in Sect. 4.

2 Method

In general, suppose K cytokines are assessed. Let M denote the total number of cytokine combinations examined, subtracting out the category of "all negative" in order to focus on the responding cells, i.e., $M = 2^K - 1$. Without loss of generality, in the rest of the paper we will focus on two-sample comparison of CD4+ antigen-specific T-cell responses.

Let n_1 be the sample size in the test group and n_2 be the sample size in the control group. We denote by $x_{i,k}$ and $y_{j,k}$, $i = 1, ..., n_1$, $j = 1, ..., n_2$, $k = 1, ..., M$, the observed frequencies of the M combinations for subjects in the test and control groups, respectively. The M-dimensional observation for each subject is $x_i = \{x_{i,1}, x_{i,2}, ..., x_{i,M}\}$ and $y_j = \{y_{j,1}, y_{j,2}, ..., y_{j,M}\}$. We order the M combinations by decreasing numbers of cytokines expressed as in Fig. 1, i.e., $x_{i,1}$ represents the frequency of responding cells for subject i that are "+" for all K cytokines, $x_{i,2}$ represents the frequency of responding cells for subject i that are "+" for the first $K - 1$ cytokines and "−" for the last (i.e., IFN-γ+ TNF-α+ IL-2+ IL-21−), and so forth.

Let $\mathscr{C} = \{C_1, C_2, ..., C_M\}$ denote the set of the M combinations. Naturally, there is an underlying order to these combinations. As intuitively a polyfunctional expression is considered better than a mono-functional expression, the M combinations are said to be partially ordered such that the kth combination (i.e., C_k) is "better" than (or at least equal to) the lth combination (i.e., C_l) if C_k expresses all the functions expressed in C_l and more. Let $\preceq$ denote the underlying partial ordering. As an example, given combinations as shown in Fig. 1, we would have $C_2, ..., C_{15} \preceq C_1$; $C_6, C_7, C_8, C_{12}, C_{13}, C_{14} \preceq C_2$; but C_2 can't be ordered in respect to the partial ordering with $C_3, C_4, C_5, C_9, C_{10}, C_{11}$, and C_{15}, because the later ones contain cytokines that are not presented in Combination #2; and so forth.

2.1 Testing a Generalized Cytokine Production Index

An intuitive approach for analyzing multidimensional data is to reduce the dimensionality by using a weighted sum. The polyfunctionality index (PI) is one example, which gives linearly increasing weights to combinations corresponding to 1, 2, until K functions while not discriminating the importance of each function. The selection of weights can be much more arbitrary, however, as there is no definitive knowledge yet from biology to determine the relative importance of different functional combinations, apart from the partial ordering described above. There can be an infinite set of possible weights and therefore, an infinite set of possible results to report.

In the context of ordinal data analysis, Agresti (1984) wrote about handling the ambiguities arising from the choice of weights [12]: "*sometimes it is not obvious how to assign scores. In such case it is informative to assign scores a variety of 'reasonable' ways to check whether substantive conclusions depend on the actual choice*." Kimeldorf et al. expanded this idea and proposed a test of two-sample categorical data by assigning all possible weights to the categories, and making inference by examining the maximum and the minimum of the test statistics over all assignments of weights [13, 14]. Along the same line of thought, we propose a test of the two-sample T-cell response data by allowing all possible weight assignments to the M combinations. In this spirit, a "generalized cytokine production index" can be written for subject i (e.g., in the test group) as

$$CPI_i = \sum_{k=1}^{M} \alpha_k x_{i,k} = \alpha^T x_i \tag{1}$$

where $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_M\}$ can be any non-negative, "eligible" weights that are assigned to the M combinations. Apparently, any "eligible" weights need to satisfy the underlying partial ordering $\preceq$. That is, in the above 4-cytokine example, α satisfies $\alpha_1 \geq \alpha_2, ..., \alpha_{15}$; $\alpha_2 \geq \alpha_6, \alpha_7, \alpha_8, \alpha_{12}, \alpha_{13}, \alpha_{14}$, but not constrained with regard to $\alpha_9, \alpha_{10}, \alpha_{11}$, or α_{15}; etc.

Once the multidimensional data are reduced to a one-dimensional continuous endpoint, standard two-sample testing techniques such as t test can be applied to compare the treatment groups. The t statistic comparing the test and control groups computed with weight α is given by

$$t(\alpha) = \frac{\alpha^T(\bar{x} - \bar{y})}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\alpha^T S \alpha}}, \tag{2}$$

where $\bar{x} = (\bar{x}_{\cdot 1}, ..., \bar{x}_{\cdot M})^T$, $\bar{y} = (\bar{y}_{\cdot 1}, ..., \bar{y}_{\cdot M})^T$ are the sample means, and S is the pooled covariance matrix estimate.

A basic calculation, as proposed by Kimeldorf et al., is to find both the minimum and the maximum of the t statistic, t_{MIN} and t_{MAX}, over all eligible assignments of

weights. The statistical inference is then based on the minimum and the maximum of the test statistic:

1. If the resulting minimum t statistic is greater than the critical value, i.e., $t_{\text{MIN}} > t^{\alpha}$, then the result is statistically significant regardless of the choice of weights;
2. If the resulting maximum t statistic is smaller than the critical value, i.e., $t_{\text{MAX}} < t^{\alpha}$, then the result is not significant regardless of the choice of weights;
3. If the optimized t statistics straddle t^{α}, i.e., $t_{\text{MIN}} < t^{\alpha} < t_{\text{MAX}}$, then the result depends on the choice of weights. In this case, care must be taken in the choice of weights and in justifying them. The α values at which the minimum and maximum occur should be examined to evaluate which scenarios lead to significant and non-significant results. The analysis may be inconclusive.

2.2 Computing Min and Max t Statistics

Next we give a procedure to find t_{MIN} and t_{MAX} over all non-negative, non-degenerate weights $\alpha_1, \alpha_2, ..., \alpha_M$ allowed by the partial ordering. In view of the scale invariance of $t(\alpha)$, we assume that the optimized weights satisfy $\max(\alpha_1, \alpha_2, ..., \alpha_M) = 1$, i.e., $\alpha_1 = 1$, which ensures that weights are non-degenerate. Finding t_{MIN} and t_{MAX} thus becomes a nonlinear optimization problem subject to linear equality and inequality constraints as defined in Eqs. (3) and (4):

$$\begin{aligned} \min_{\alpha} \quad & t(\alpha) = \frac{\alpha^T(\bar{x} - \bar{y})}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\alpha^T S \alpha}} \\ \text{s.t.} \quad & C\alpha \geq 0 \\ & \alpha_k \geq 0; \alpha_1 = 1 \end{aligned} \tag{3}$$

$$\begin{aligned} \max_{\alpha} \quad & t(\alpha) = \frac{\alpha^T(\bar{x} - \bar{y})}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\alpha^T S \alpha}} \\ \text{s.t.} \quad & C\alpha \geq 0 \\ & \alpha_k \geq 0; \alpha_1 = 1 \end{aligned} \tag{4}$$

where C is the linear inequality constraint matrix which defines the partial ordering.

We show that the computation of t_{MIN} and t_{MAX} is different under three scenarios determined by the *stochastic ordering* of the study populations. The notion of *stochastic ordering* plays an important role in whether or not the optimized t statistics straddle 0. The following definitions are needed in determining *stochastic ordering* among data sets [14].

Definition 2.1 A subset L of $\mathscr{C}$ is called a lower set with respect to the partial ordering $\preceq$ if $C_i \in \mathscr{C}, C_j \in L$, and $C_i \preceq C_j$ imply $C_i \in L$. A subset U of $\mathscr{C}$ is

called a upper set with respect to the partial ordering $\preceq$ if $C_i \in U$, $C_j \in \mathscr{C}$, and $C_i \preceq C_j$ imply $C_j \in U$.

For example, $U = \{C_1, C_2, C_5, C_8\}$ is an upper set among the combinations shown in Fig. 1. Denote the class of all lower sets by $\mathscr{L}$ and the class of all upper sets by $\mathscr{U}$. For a given partial order, the enumeration and specification of all upper sets (or equivalently lower sets) may require careful algorithms and substantial computational time. As an example, the set of 15 combinations in Fig. 1 would constitute a total of 94 upper sets.

Definition 2.2 The data from Population X are said to be stochastically larger than data from Population Y with respect to the partial ordering $\preceq$, if

$$\sum_{\{j:C_j \in U\}} \bar{x}_{\cdot j} \geq \sum_{\{j:C_j \in U\}} \bar{y}_{\cdot j} \tag{5}$$

for every upper set $U \in \mathscr{U}$.

We have the following equivalence theorem for the stochastic ordering of the study populations with respect to $\preceq$ and the non-negativity of the key statistic $t(\alpha)$. The proof is given in the Appendix.

Theorem 2.1 *If the data from Population X are stochastically greater (smaller) than data from Population Y with respect to the partial order $\preceq$, then $t(\alpha) \geq 0$ ($t(\alpha) \leq 0$) for all eligible weights consistent with the partial order $\preceq$.*

The applications of Theorem 2.1 is immediate. If the data from population X are stochastically larger than the data from population Y with respect to the partial order $\preceq$, then $t_{\text{MIN}} \geq 0$. On the other hand, if the population X data are stochastically smaller than the population Y data, then $t_{\text{MAX}} \leq 0$. If they are incomparable, then $t_{\text{MIN}} \leq 0 \leq t_{\text{MAX}}$. The computation of t_{MIN} and t_{MAX} therefore can be considered separately for these three cases:

Scenario 1. *If data from Population X (test) are stochastically greater than data from Population Y (control) with respect to the partial order $\preceq$*, then

(1) t_{MIN} *is attained at one of the extreme points of the feasible region;*
(2) t_{MAX} *can be found by convex programming.*

This result follows from the observation that $t(\alpha)$ is a quasiconcave function under Scenario 1, and that the feasible region defined by the partial ordering constraints is a convex set.

To see that function $t(\alpha)$ is quasiconcave under Scenario 1, let us first notice that $t(\alpha)$ is a ratio of a linear function and the square root of a quadratic term. Under Scenario 1, the linear term is non-negative for all α that satisfies the partial ordering constraints, and therefore a non-negative concave function. Next, we observe that the pooled covariance matrix estimate S is positive definite, which suggests that

the square root of the quadratic term is strictly positive convex [15]. From Avriel et al. [16], this implies that $t(\alpha)$, as a ratio of a non-negative concave function and a strictly positive convex function, is quasiconcave.

It follows from Theorem 3.5.3 of Bazaraa and Shetty (1979) that $t(\alpha)$ attains its minimum among the *extreme points* of the feasible region defined by the linear inequality constraints [17]. Let $S = \{(\alpha_1, \alpha_2, ..., \alpha_M), C\alpha \geq 0, \alpha_k \geq 0, \text{and } \alpha_1 = 1\}$ denote the feasible region, and $\mathscr{U}$ denote the class of all upper sets of the partial order $\preceq$.

Definition 2.3 The extreme points of the feasible set S are given by $p_r, r = 1, ..., s$, where $p_r = (\alpha_1, \alpha_2, ..., \alpha_M)$, and $\alpha_k, k = 1, ..., M$, is defined by

$$\alpha_k = \begin{cases} 1, & C_k \in U_r, \\ 0, & \text{otherwise}, \end{cases}$$

for each upper set $U_r \in \mathscr{U}, r = 1, ..., s$.

Thus, to find t_{MIN} under Scenario 1, we need to compute t for all the extreme points of S, and an extreme point where t takes the minimum gives t_{MIN}. For a given partial order, the specification of all extreme points (or equivalently upper sets) requires careful enumeration. As mentioned above, for example, the set of 15 combinations in Fig. 1 would constitute a total of 94 upper sets and thus 94 extreme points.

Next we consider computing the maximum of $t(\alpha)$ under Scenario 1. Because $t(\alpha)$ is quasiconcave, and the feasible region S is a convex set, it follows from Proposition 1 and Theorem 3.37 of Avriel et al. that any local maximum is also a global maximum of Eq. (4) under Scenario 1 [16]. Therefore, Eq. (4) can be solved by any convex programming algorithms.

In like manner we obtain the results for the analogous case when population X data are stochastically smaller than population Y data:

Scenario 2. *If data from Population X (test) are stochastically smaller than data from Population Y (control) with respect to the partial order* $\preceq$, then

(1) t_{MIN} *can be found by convex programming;*
(2) t_{MAX} *is attained at one of the extreme points of the feasible region.*

Scenario 3. *If data from Population X (test) and Y (control) are stochastically incomparable with respect to the partial order* $\preceq$, then

(1) t_{MIN} *can be found by convex programming by solving Eq.* (3) *with an additional constraint* $\alpha^T(\bar{x} - \bar{y}) \leq 0$;
(2) t_{MAX} *can be found by convex programming by solving Eq.* (4) *with an additional constraint* $\alpha^T(\bar{x} - \bar{y}) \geq 0$.

Scenario 3 implies that $t_{\text{MIN}} \leq 0$. Hence solving Eq. (3) is equivalent to solving it with the additional constraint of $\alpha^T(\bar{x} - \bar{y}) \leq 0$, which requires $t(\alpha) \leq 0$ within the feasible region. Thus $t(\alpha)$ is quasiconvex in the feasible region and t_{MIN} can be

found by convex programming. Similarly $t(\alpha)$ is quasiconcave under the additional constraint of $\alpha^T(\bar{x} - \bar{y}) \geq 0$, and t_{MAX} can be found by convex programming.

3 Example

We illustrate the application of the proposed method on a publicly available dataset of ICS T-cell responses [11]. We consider T-cell production data generated as part of the RV144 HIV vaccine study among health adults in Thailand [18, 19]. Expression of a set of six functions (TNF-α, IFN-γ, IL-4, IL-2, CD40L, and IL-17α) was measured in CD4+ T cells by ICS to determine if there are differences in the induced T-cell responses between vaccine ($n = 226$) and control ($n = 36$) groups.

Using the ICS assay, each individual cell is classified as either positive (+) or negative (−) upon antigen stimulation for each cytokine based on fixed thresholds, and the numbers of CD4+ antigen-specific cells expressing each functional combination are counted. These measurements are available for each person on CD4+ T-cells, against the 92TH023-Env peptide pool. Although a total of 63 ($2^6 - 1$) functional combinations can be defined by the six functions, only 15 of these had non-negligible cell counts (over five cells in more than two subjects). Hence for the purpose of illustration in this paper we will focus on these 15 combinations. Figure 2 gives an illustration of the data for CD4+ antigen-specific T cells in a particular treatment group (e.g., vaccine group); the columns represent the functional combinations and the data are numbers of cells producing each combination per million cells, after subtracting background (unstimulated) values.

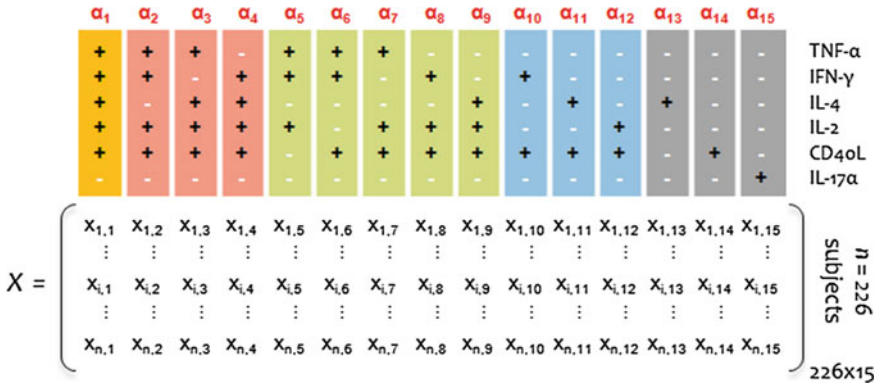


Fig. 2 Illustration of RV144 ICS T-cell response data in the vaccine group. Data are background-subtracted numbers of antigen-specific CD4+ T cells, per million cells, expressing the 15 functional combinations. Combinations are grouped by color = "degree of functionality". The α's are the weights assigned to each combination in a weighted analysis

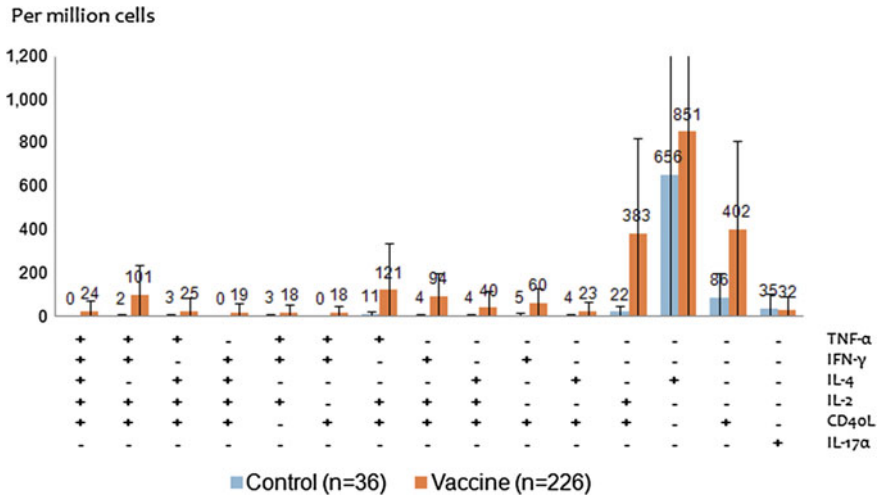


Fig. 3 Polyfunctional profiles of antigen-specific CD4+ T cells. The bar represents the mean number of CD4+ T cells expressing each of the 15 combinations per million cells in each group and the whisker represents the standard deviation

As shown in Fig. 3, the vaccine induced considerably larger numbers of CD4+ cells exhibiting polyfunctional responses compared to the control. The magnitude of response was significantly higher in the vaccine group compared to the control group, for all combinations except for the 13th (TNF-α- IFN-γ- IL-4+ IL-2- CD40L- IL-17α-) and 15th (TNF-α- IFN-γ- IL-4- IL-2- CD40L- IL-17α+) combinations, at the Bonferroni-adjusted level of $\alpha = 0.0033$. Applying the simple polyfunctionality index, the vaccine group reported a mean PI of 631, significantly higher than the control group, which reported mean PI of 155 ($p = 0.0002$).

We would like to see if the significant result could remain for any choice of weights that satisfy the partial ordering, which could help us confidently claim a significant effect of the vaccination. Using Definition 2.2, we first observe that the empirical distribution for the vaccine group is stochastically larger than the empirical distribution for the control group by checking the class of all upper sets. Note that this particular set of 15 combinations with its partial ordering constitutes a total of 167 upper sets. Thus under Scenario 1, the weights α_{MIN} which minimize $t(\alpha)$ over all weights consistent with the partial ordering can be found by calculating $t(\alpha)$ for all extreme points. The weights α_{MAX} which maximize $t(\alpha)$ can be found by convex programming. The resulting optimizing weights α_{MIN} and α_{MAX} are tabulated in Table 1A with $t_{\text{MIN}} = 0.6147$ ($p = 0.5382$) and $t_{\text{MAX}} = 5.3688$ ($p < 0.0001$). We know therefore, in this straddling case, there are some weights that produce significance and some that do not.

Upon examining the weights leading to t_{MIN}, we notice that α_{MIN} gives 0 weights to one tetra-functional and several tri-functional combinations. Although this is theoretically allowed by the partial ordering, we might ask whether this is sensible

Table 1 Minimum and Maximum t statistics for testing a treatment difference with the RV144 T cell response data, and corresponding optimizing weights and p-values: **A** under the general partial ordering constraints; **B** under informative constraints incorporating medical judgment

		$\bar{x} - \bar{y}$	(A) Genrel constraints		(B) Genrel constraints	
			α_{MIN}	α_{MAX}	α_{MIN}	α_{MAX}
α_1	$++++-+-$	23.85	1	1	1	1
α_2	$++-++-$	98.84	0	1	1	1
α_3	$+-+++-$	21.74	1	0.26	1	0.26
α_4	$-+++-$	19.48	1	1	1	1
α_5	$++-+--$	15.21	0	0	1	0
α_6	$++--+-$	17.11	0	1	1	1
α_7	$+--++-$	110.45	0	0.26	1	0.26
α_8	$-+-++-$	89.61	1	0.26	1	0.26
α_9	$--+++-$	36.19	1	0.26	1	0.26
α_{10}	$-+--+-$	54.82	0	1	1	1
α_{11}	$--+-+-$	18.61	1	0.26	1	0.26
α_{12}	$---++-$	361.78	0	0.26	1	0.26
α_{13}	$--+---$	195.58	1	0.01	1	0.01
α_{14}	$----+-$	316.59	0	0.26	0	0.26
α_{15}	$-----+$	−3.80	0	0	1	0
	t p-value		$t_{MIN} =$ 0.6147 0.5382	$t_{MAX} =$ 5.3688 <0.0001	$t_{MIN} =$ 2.0018 0.0463	$t_{MAX} =$ 5.3688 <0.0001

medically and biologically. In light of this consideration, we note that our method can be easily adapted to incorporate medical insight and common sense when appropriate. Specifically, basic medical knowledge and/or judgment can be incorporated into the statistical calculation by being quantified in the form of simple constrains to the weights. In this example, for instance, a clinical researcher might require weights no less than 0.5 for the tetra-functional combinations, no less than 0.3 for the triple-functional combinations, and no less than 0.1 for the dual functional combinations; i.e., $\alpha_2, ..., \alpha_4 \geq 0.5$, $\alpha_5, ..., \alpha_9 \geq 0.3$, and $\alpha_{10}, ..., \alpha_{12} \geq 0.1$. We solve t_{MIN} and t_{MAX} similarly under these added constraints—we can call them "informative constraints"—and the results are given in Table 1B. Here t_{MAX} remains the same and $t_{MIN} = 2.0018$ ($p = 0.0463$). If there is sufficient medical evidence to support the informative constraints, we can conclude the null hypothesis that vaccine and control recipients exhibit the same T-cell responses can always be rejected.

4 Discussion

Analysis of increasingly multidimensional T-cell response datasets presents a great challenge to statisticians. An intuitive approach for analyzing multidimensional data is to reduce the dimensionality by using a weighted sum. The selection of weights can be arbitrary, as there are few biological reason for considering one functional combination more important than the others. There can be, therefore, a limitless set of possible results to report. To handle the ambiguities arising from the choice of weights, we propose a solution by computing the maximum and the minimum of the test statistic over all possible assignments of weights. Statistical inference is based on the maximum and the minimum of the test statistic. We illustrate that under non-straddling cases our approach allows for a strong scientific statement concerning the significance or insignificance of the difference between the two populations or treatments. In the straddling case, we become aware that care and interpretability are important in choosing the numeric weights by which we analyze the data.

Our method can still provide helpful insight into the data under the straddling cases. We suggest the scientific meaning of the weights that produce both significance and insignificance be examined in the context of the study for their relevance. For example, one can become more confident with the significant results, if the minimizing weights (or more generally the sets of weights that produce smaller t statistics) are much less clinically meaningful as the maximizing weights. One can also use the degree of overlap of t_{MIN} and t_{MAX} relative to their appropriate critical value as an indication of the strength of experimental evidence. A larger sample size might be helpful if a trend towards significance is observed. This could be a direction of future work in this topic. Nonetheless, the straddling cases, we expect, will always be somewhat difficult in its interpretability due to the inherent ambiguity in the data.

It is important to note that there are still many areas of cell-mediated immunity analyses that could benefit from increased statistical input. For example, the identification a vaccine-induced immune response that predicts vaccine protection, i.e. correlate of protection, has long been a central goal of vaccine research. Multiple authors have proposed methods and frameworks to assess and establish such immune correlates, with one or a few immune response variables involved. We expect the task of establish immune correlates and a prediction model becomes exponentially more challenging, with multidimensional ICS data. In addition, the focus of the current paper gears more towards testing rather than estimation, which could be an important area for future work. More exploratory work will need to be done in this area to fully utilize these multidimensional data to really benefit clinical investigation.

Appendix

Proof of Theorem 2.1. Let population X data be stochastically greater with respect to the partial ordering $\preceq$ than population Y data. We need to show $\text{t}(\alpha) \geq 0$ for all eligible weights consistent with $\preceq$.

Given any α consistent with the partial ordering $\preceq$, let $\alpha_{(1)} \geq \alpha_{(2)} \geq ... \geq \alpha_{(M)}$ denote them placed in descending order. The M combinations corresponding to this order can be written as $\{C_{(1)}, C_{(2)}, ..., C_{(M)}\}$ and the corresponding data as $\bar{x}^* = (x_{\cdot(1)}, ..., x_{\cdot(M)})^T$ and $\bar{y}^* = (y_{\cdot(1)}, ..., y_{\cdot(M)})^T$.

Denote a sequence of sets $\{U_{(k)}\}$ such that $U_{(k)} = \{C_{(1)}, C_{(2)}, ..., C_{(k)}\}$, $k = 1, 2, ..., M$. It is easy to see that each $U_{(k)}$ is an upper set. Then by Definition 2.2 we have $\sum_{j=1}^{k} \bar{x}_{\cdot(j)} \geq \sum_{j=1}^{k} \bar{y}_{\cdot(j)}$, $k = 1, 2, ..., M$. Let $\Delta_{(j)} = \bar{x}_{\cdot(j)} - \bar{y}_{\cdot(j)}$, we have $\sum_{j=1}^{k} \Delta_{(j)} \geq 0$, $k = 1, 2, ..., M$.

Given $\alpha_{(1)} \geq \alpha_{(2)} \geq ... \geq \alpha_{(M)}$, let us write $\alpha_{(1)} = \alpha_{(2)} + \delta_{(1)}$, $\alpha_{(2)} = \alpha_{(3)} + \delta_{(2)}$, ..., $\alpha_{(M-1)} = \alpha_{(M)} + \delta_{(M-1)}$, with $\delta_{(1)}, ..., \delta_{(M-1)} \geq 0$. That is, $\alpha_{(j)} = \alpha_{(M)} + \sum_{r=j}^{M-1} \delta_{(r)}$, $j = 1, 2, ..., M$. It follows that

$$\begin{aligned} \alpha^T(\bar{x} - \bar{y}) &= \sum_{j=1}^{M} \alpha_{(j)}(\bar{x}_{\cdot(j)} - \bar{y}_{\cdot(j)}) \\ &= \sum_{j=1}^{M} \{\alpha_{(M)} + \sum_{r=j}^{M-1} \delta_{(r)}\} \Delta_{(j)} \\ &= \alpha_{(M)} \sum_{j=1}^{M} \Delta_{(j)} + \sum_{j=1}^{M-1} \delta_{(j)} \sum_{r=1}^{j} \Delta_{(r)} \geq 0, \end{aligned} \tag{6}$$

where the inequality follows from the fact that $\alpha_{(k)} \geq 0, \delta_{(k)} \geq 0$, and $\sum_{j=1}^{k} \Delta_{(j)} \geq 0$, for $k = 1, 2, ..., M$. Hence we have proven Theorem 2.1.

References

1. De Rosa, S.C., Lu, F.X., Yu, J., Perfetto, S.P., Falloon, J., Moser, S., Evans, T.G., Koup, R., Miller, C.J., Roederer, M.: Vaccination in humans generates broad T cell cytokine responses. J. Immunol. **173**(9), 5372–5380 (2004)
2. Rodrigue-Gervais, I.G., Rigsby, H., Jouan, L., Sauv, D., Saly, R.P., Willems, B., Lamarre, D.: Dendritic cell inhibition is connected to exhaustion of CD8+ T cell polyfunctionality during chronic hepatitis C virus infection. J. Immunol. **184**(6), 3134–3144 (2010)
3. Ciuffreda, D., Comte, D., Cavassini, M., Giostra, E., Bhler, L., Perruchoud, M., Heim, M.H., Battegay, M., Genn, D., Mulhaupt, B., Malinverni, R.: Polyfunctional HCVspecific Tcell responses are associated with effective control of HCV replication. Eur. J. Immunol. **38**(10), 2665–2677 (2008)
4. Precopio, M.L., Betts, M.R., Parrino, J., Price, D.A., Gostick, E., Ambrozak, D.R., Asher, T.E., Douek, D.C., Harari, A., Pantaleo, G., Bailer, R.: Immunization with vaccinia virus induces polyfunctional and phenotypically distinctive CD8+ T cell responses. J. Exp. Med. **204**(6), 1405–1416 (2007)
5. Darrah, P.A., Patel, D.T., De Luca, P.M., Lindsay, R.W., Davey, D.F., Flynn, B.J., Hoff, S.T., Andersen, P., Reed, S.G., Morris, S.L., Roederer, M.: Multifunctional T H 1 cells define a correlate of vaccine-mediated protection against Leishmania major. Nat. Med. **13**(7), 843 (2007)
6. Roederer, M., Nozzi, J.L., Nason, M.C.: SPICE: exploration and analysis of postcytometric complex multivariate datasets. Cytometry Part A **79**(2), 167–174 (2011)

7. Morrison, D.F.: Multivariate Statistical Methods. McGraw-Hill, New York (1976)
8. Nason, M.: Patterns of immune response to a vaccine or virus as measured by intracellular cytokine staining in flow cytometry: hypothesis generation and comparison of groups. J. Biopharm. Stat. **16**, 483–498 (2006)
9. Larsen, M., Sauce, D., Arnaud, L., Fastenackels, S., Appay, V., Gorochov, G.: Evaluating cellular polyfunctionality with a novel polyfunctionality index. PLoS One **7**(7), e42403 (2012)
10. Finak, G., McDavid, A., Chattopadhyay, P., Dominguez, M., De Rosa, S., Roederer, M., Gottardo, R.: Mixture models for single-cell assays with applications to vaccine studies. Biostatistics **15**(1), 87–101 (2013)
11. Lin, L., Finak, G., Ushey, K., Seshadri, C., Hawn, T.R., Frahm, N., Scriba, T.J., Mahomed, H., Hanekom, W., Bart, P.A., Pantaleo, G.: COMPASS identifies T-cell subsets correlated with clinical outcomes. Nat. Biotechnol. **33**(6), 610 (2015)
12. Agresti, A.: Analysis of Ordinal Cathegorical Data. Willey, New York (1984)
13. Kimeldorf, G., Sampson, A.R., Whitaker, L.R.: Min and max scorings for two-sample ordinal data. J. Am. Stat. Assoc. **87**(417), 241–247 (1992)
14. Sampson, A.R., Singh, H.: Min and max scorings for two sample partially ordered categorical data. J. Stat. Plan. Inference **107**(1–2), 219–236 (2002)
15. Landsman, Z.: Minimization of the root of a quadratic functional under an affine equality constraint. J. Comput. Appl. Math. **216**(2), 319–327 (2008)
16. Avriel, M., Diewert, W.E., Schaible, S., Zang, I.: Generalized Concavity. Plenum Press, New York (1988)
17. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms. Wiley, New York (2013)
18. Haynes, B.F., Gilbert, P.B., McElrath, M.J., Zolla-Pazner, S., Tomaras, G.D., Alam, S.M., Evans, D.T., Montefiori, D.C., Karnasuta, C., Sutthent, R., Liao, H.X.: Immune-correlates analysis of an HIV-1 vaccine efficacy trial. N. Engl. J. Med. **366**(14), 1275–1286 (2012)
19. Rerks-Ngarm, S., Pitisuttithum, P., Nitayaphan, S., Kaewkungwal, J., Chiu, J., Paris, R., Premsri, N., Namwat, C., de Souza, M., Adams, E., Benenson, M.: Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N. Engl. J. Med. **361**(23), 2209–2220 (2009)

Project Data Sphere and the Applications of Historical Patient Level Clinical Trial Data in Oncology Drug Development

Greg Hather and Ray Liu

Abstract As scientific data sharing initiatives become more popular, an increasing amount of oncology clinical trial data is becoming available to the public. This historical data has the potential to help improve the design and analysis of future studies of new oncology compounds. Project Data Sphere is one such public database of oncology studies, with patient level data from over 76,000 patients. Here, we review the contents of this database and describe several examples of how the data has been used or could potentially be used in drug development. Applications include population selection, historical comparisons, and identification of stratification factors.

Keywords Oncology · Data sharing · Project Data Sphere · Population selection · Stratification · Historical comparison

1 Introduction

Since the first human clinical trials of oncology compounds began, over 35,000 oncology clinical trials have been completed [1]. Traditionally, the patient level data from these trials has been kept private due to commercial, legal, and patient privacy concerns. This private data can only be reused by the data owner, thus limiting its utility. The resulting inefficiency and duplication of effort slows the progress of oncology drug development.

In recent years, some data owners have agreed to share their oncology data through various initiatives. These initiatives include Project Data Sphere [2], which is an oncology focused data sharing platform. Project Data Sphere was started by the CEO Roundtable on Cancer Life Sciences Consortium as a platform to voluntarily share deidentified historical oncology clinical trial datasets for the sake of advancing future cancer research. The goal was for industry, academia, and governmental orga-

G. Hather (✉) · R. Liu
Takeda Pharmaceuticals Inc., Cambridge, MA, USA
e-mail: Greg.hather@takeda.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_19

nizations to share their data with the public for scientific reuse. Project Data Sphere began sharing patient data in 2014.

Project Data Sphere can be compared to several other data sharing initiatives that include oncology clinical trial data. For example, ClinicalStudyDataRequest.com [3], Yale University Open Data Access [4], and Pfizer's data transparency initiative [5], all include patient level oncology data. However, these platforms require a research proposal and an approval process which can take many months [6]. These access restrictions discourage the use of the data in drug development. Other potential sources of oncology clinical trial data include the NCI Genomic Data Commons [7] and dbGaP [8]. However, most studies in these databases lack detailed clinical data, since the focus of these initiatives is to share genomic data. Shared data is also available for specialized oncology populations, such as the MMRF CoMMpass study, which has clinical data from a large cohort of multiple myeloma patients [9].

In terms of the number of patients included, Project Data Sphere is currently the largest public source of patient level oncology clinical trial data, to the knowledge of the authors. In addition, the data is detailed and relatively easy to access. While the number of studies in Project Data Sphere is only a small fraction of all completed oncology trials, we believe that for many cancer types, there is now enough publically available patient level data to routinely consider using it as an aid in the design and analysis of new trials. In this paper, we describe the contents of Project Data Sphere and potential applications of this historical patient level data for oncology drug development.

2 Project Data Sphere

Project Data Sphere currently contains data from over 76,000 patients [10]. These patients are in 108 studies volunteered by 25 different providers. Users may browse and download available study data through the platform's website (https://www.projectdatasphere.org/). Table 1 shows the number of studies provided for each type of cancer. Most of the studies began on or after 2005 (see Fig. 1). Older data is present in the database, but the older data may be less valuable if current treatment patterns and outcomes have changed. The large majority of the studies (105 out of 108) are Phase III trials. All trials include the protocol, CRF, data dictionary, and SAS datasets.

Project Data Sphere comes with very few access restrictions. To obtain access, the user submits a brief application online. Although users are encouraged to submit a brief description (up to 1,000 characters) of their research, no proposal is required. In the authors' experience, we were granted access within one business day of applying. Once access is granted, the user has access to the majority (85 out of 108) of trials in Project Data Sphere. The remaining trials are those contributed by the NCI, which require additional steps to access. To access an NCI trial, the user must fill out a brief online form and provide a short research plan specific to the requested data (up to 1,000 characters). In addition, the user must have a member of their

Table 1 The number of studies available for each type of cancer. Note that some studies are counted more than once in this table because they include multiple types of cancer

Cancer type	Number of studies
Breast	27
Prostate	26
Colorectal	15
Lung (Non-small cell)	13
Ovarian	7
Central nervous system	6
Leukemia	6
Pancreatic	6
Head-neck	5
Gastric	4
Liver	4
Lung (Small cell)	4
Myelodysplastic syndrome (MDS)	4
Neuroblastoma	4
Bone sarcoma	3
Brain	3
Germ	3
Kidney	3
Lymphoma (Non-hodgkins)	3
Multiple myeloma	3
Bladder	2
Melanoma	2
Myeloproliferative neoplasm	2
Soft tissue sarcoma	2
Testicular	2
Esophageal	1
Lymphoma (Hodgkins)	1
Mucositis	1
Myelofibrosis	1
Neuroendocrine tumors	1
Uterine cervix	1

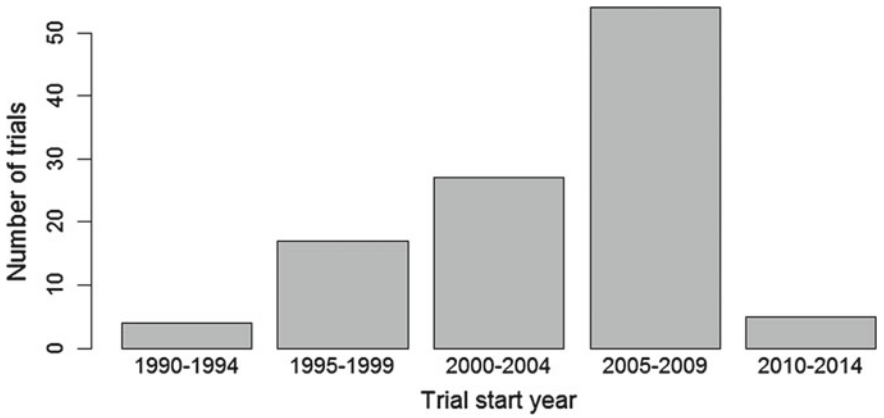


Fig. 1 Distribution of the trial start year for studies in Project Data Sphere

organization's legal department sign the NCI data use agreement. The agreement requires that patient level data must not be shared, and that Project Data Sphere must be acknowledged in publications.

To give an example of the richness of the data in Project Data Sphere, we highlight the data available from a single study, NCT00415194, conducted by Eli Lilly. In this study, have data for 397 subjects with metastatic cancer in the standard of care arm, which was Placebo + Cisplatin. For each patient, we have age, sex, tobacco use history, and cancer primary site. In addition, we have lab values and collection dates, vitals at each visit, including weight and ECOG status, and list of concomitant medicines. The data also provides AE records, with start and end dates. Finally, we have the average Cisplatin dose (planned and actual), Overall Survival Time, and Progression-free Survival Time.

One limitation of Project Data Sphere is that for the majority of studies (73 out of 108), only control data is available. Another limitation of Project Data Sphere is that the data is not standardized across formats. Data is provided in raw data format or analysis data format. The level of data standardization for a particular study depends on the age of the trial and the level of CDISC standards adoption. Therefore, pooling datasets across studies may require substantial effort to integrate the data. Furthermore, users should note that the maintainers of Project Data Sphere are not responsible for data quality; instead, this is the responsibility of the organizations that volunteer the data. Therefore, the data quality can vary between studies. An additional limitation of Project Data Sphere is that for most studies, only a subset of the recorded variables have been released to the database. For example, for trial NCT00415194, lesion measurements, radiotherapy information, surgery information, and protocol deviations were recorded, but these measurements were not provided to Project Data Sphere.

3 Application of Project Data Sphere in Drug Development

Although Project Data Sphere is a relatively recent initiative, several papers published have made use of Project Data Sphere data. Wendling et al. [11] identified prognostic biomarkers for survival in pancreatic cancer. Gill et al. [12] and Geifman et al. [13] identified concomitant medications and surgical treatments that are associated with survival in prostate cancer patients. As another example, Abdallah et al. [14] built a predictive model for prostate cancer survival as part of a crowd-sourcing challenge. Finally, Green et al. [15] performed an indirect comparison of two different drug regiments for prostate cancer.

Although the published applications of Project Data Sphere we described above were useful in understanding disease progression and selecting treatments, supporting drug development was not the focus of these papers. Thus, we describe below several possible use cases for Project Data Sphere for the design and analysis of new oncology trials. These use cases are drawn from papers using patient level historical data. Here, we present 3 major use cases of how the patient level data in Project Data Sphere can support drug development and provide value beyond what summary data can provide. These applications are population selection, historical comparisons, and identification of stratification factors.

4 Population Selection

Choosing which population to run a trial in is one of the major decisions in clinical research. Generally, this decision is made by clinicians based on the drug's mechanism of action, preclinical data, and early clinical data. The project team's assessment of unmet medical need and the competitive landscape also contribute to this decision. However, historical patient level data can provide additional information to help the team identify the right patient population [16].

When setting the inclusion/exclusion criteria for a trial, one approach is to identify a population with a poor expected outcome under the standard of care. Not only does this approach focus on a population with unmet medical need, but it also increases the statistical power to detect a drug effect. To understand why, consider a trial where the population selected is expected to have a long survival time under standard of care. Under this scenario, few events would be observed before the end of the study, and the power to detect a drug effect would be low [17]. While it is possible to extend the length of the trial or increase the number of patients studied to increase the power, both options would increase the cost of the clinical trial. A relevant study from Project Data Sphere could help compare different patient subgroups under standard of care. Patient subgroups that are unlikely to experience an event before the end of the trial could be identified and potentially excluded from the study. In this way, much of the prognostic modeling that has already been done with Project Data Sphere data [11–14] could help to set inclusion criteria.

Another requirement for a successful trial is that the population must be large enough for a clinical trial to be run. If the population selected is too small, the trial may suffer from low enrollment. Project teams can use Project Data Sphere data to determine the relative size of populations considered by the team. Such an analysis would give the team quantitative evidence to ensure that the selected population is large enough.

Statisticians can contribute to population selection. First, the statistician can search for a relevant study in Project Data Sphere, with input from the project team. If a relevant dataset can be identified, the statistician can then summarize the distribution of relevant covariates and outcomes, and show how each covariate is related to the outcome. Based on these results, the team can propose several candidate sets of inclusion/exclusion criteria.

To compare different proposed populations based on candidate sets of inclusion/exclusion criteria, the statistician can compute the number of patients in the dataset that fall within each population definition. In addition, the statistician can present the distribution of outcomes for each population. When one population is a subset of the other population, the statistician can also present the distribution of outcomes of the excluded patients. If necessary, the statistician may also perform a sample size calculation or simulation comparing the proposed inclusion/exclusion criteria to estimate trial performance characteristics [18]. Based on these results, the team may arrive at a final set of criteria or new proposals to test.

5 Historical Comparisons

Phase II oncology trials traditionally do not use a control group. Instead, the results are usually compared to historical summary level data. Historical comparisons with summary level control data have a long history in the statistics literature [19]. One shortcoming of historical comparisons is the possibility that the populations are different, thus confounding the comparison. This problem can be partly corrected by adjusting for baseline covariates in the current and historical populations. While several methods for adjustment exist with summary level historical data [20, 21], more accurate adjustment methods can be used when patient level historical data is available [15, 22–27]. Therefore, by using patient level data from Project Data Sphere as a historical control, the comparisons are likely to be more accurate. Common methods of analysis when full patient level data is available include matched pair analysis [15, 22–24], inverse probability weighting [25–27], and regression modeling [22, 23]. All these methods require patient level data from both studies, and they can account for differences in baseline covariates between the treatment group and the historical control group. These methods can reduce both the bias and the variance in the estimated treatment effect [27].

Regression modeling is one approach to historical comparisons. With regression modeling, the outcome is modeled as a function of the covariates, and the treatment arm is included as a covariate. The estimated "treatment arm effect" from the regression model is the estimated drug effect. Unlike matched pair analysis, regression modeling allows all data to be used. One disadvantage of this approach is that if the regression model is misspecified, the resulting estimate for the drug effect will be biased.

Another approach to historical comparison is inverse probability weighting. With this method, one fits a regression model where the outcome is the treatment group and the covariates are the baseline measurements. This type of model is called a propensity score model. Patients in the historical group can then be reweighted so that, assuming the model is correct, the distribution of baseline covariates is identical to that of the new trial. The drug effect can then be estimated using standard approaches (e.g. Cox regression) applied to the reweighted data. While a propensity model may be less intuitive than the regression model, the propensity score may depend on fewer covariates and may be easier to model. However both regression modeling and inverse probability weighting require the model to be correctly specified, otherwise the estimated drug effect will be biased.

An alternative approach to historical comparison is matched pair analysis, which matches similar patients in the current and historical data. Here, similarity is determined by the propensity score. One advantage of this method is that once the pairs are selected, the analysis is the same as that of a real randomized trial. A disadvantage of this method is that due to the difference in the sample sizes of the two studies, some of the data points will not be part of a matched pair, and these data will not be used in the analysis. This reduces the efficiency of the analysis, and it may result in higher uncertainty in the estimates.

When using Project Data Sphere data as a historical control, the Project Data Sphere study and analysis method should be pre-specified to be statistically valid. In addition, the statistician should work with clinical experts to confirm that the Project Data Sphere study is similar to the current study in terms of the eligibility criteria, control treatment, treatment evaluation, and distribution of baseline characteristics [19]. Ideally, the Project Data Sphere study should be recent, to reduce the amount that patient care methods may have changed. Finally, the Project Data Sphere study should have a sufficiently large sample size.

6 Stratification

In many Phase III oncology trials, the analysis is stratified by pre-specified baseline covariates [28, 29]. One reason for using stratification is the belief that certain patient populations may have different treatment efficacy. Another reason to use stratification is that even if the treatment is the same across the subgroups, stratification can potentially increase the power of the analysis because the heterogeneity within each subgroup is expected to be lower than the heterogeneity of the entire population.

The choice of stratification factor is generally made by clinicians based on the clinician's knowledge of the population, the drug's mechanism of action, preclinical data, and early clinical data. However, historical patient level data can provide additional information to help the team identify the right stratification factors and the right thresholds for the strata [29].

One way that a statistician can help the team is to fit a model that predicts the outcome of interest using all the available baseline covariates. Next, the statistician could use the model to calculate the fraction of variability in the outcome explained by each of the baseline covariates. The list of covariates ranked by the magnitude of their association with the outcome could then be supplied to the project team. The project team can then consider these covariates, along with other prior knowledge about the disease, when deciding which covariates to stratify by.

Another way that the statistician can support the stratification plan is to plot the distribution of each covariate that is being considered as a stratification factor. If cut points are set too high or low, some of the strata may have very few patients. Analyzing the covariate distribution can help the team set better cut point(s) for the strata.

Finally, one could perform a simulation analysis of different stratification approaches by drawing on data from Project Data Sphere. For example, one could simulate a clinical trial by random sampling of patients with replacement from a Project Data Sphere study [18]. These patients could then be randomly assigned to the treatment or control group of a trial. This would allow us to simulate a drug that had no effect. The simulated data could be analyzed to calculate the treatment effect, and the standard error of the estimated effect size could be computed. Next, one could simulate different stratification approaches and check for the effect on the standard error of the estimate. A stratification analysis that produced the lowest standard error of the estimate would be preferred.

7 Conclusion

Project Data Sphere is a large public database of patient level oncology clinical trial data. The database continues to grow and include coverage of less common cancers. Project Data Sphere contains rich clinical data, and the data is easy to access. Based on these feature, Project Data Sphere shows promise to be routinely applied to support oncology drug development. In this paper, we summarized the contents of Project Data Sphere and identified 3 major applications of Project Data Sphere data in drug development: clinical trial population selection, historical comparisons, and identification of stratification factors.

References

1. Trialtrove. Informa PLC, London. https://pharmaintelligence.informa.com/products-and-services/data-and-analysis/trialtrove (2017). Last accessed 14 Aug 2017
2. Green, A.K., Reeder-Hayes, K.E., Corty, R.W., Basch, E., Milowsky, M.I., Dusetzina, S.B., Bennett, A.V., Wood, W.A.: The project data sphere initiative: accelerating cancer research by sharing data. Oncologist **20**, 464-e20 (2015)
3. Strom, B.L., Buyse, M., Hughes, J., Knoppers, B.M.: Data sharing, year 1—access to data from industry-sponsored clinical trials. New. Engl. J. Med. **371**, 2052–2054 (2014)
4. The YODA project summary of data inquiries and requests. Yale University, New Haven. http://yoda.yale.edu/summary-data-inquiries-and-requests (2017). Last accessed 14 Aug 2017
5. Pfizer Trial Data & Results. Pfizer Inc, New York. http://www.pfizer.com/science/clinical-trials/trial-data-and-results (2017). Last accessed 14 Aug 2017
6. Geifman, N., Bollyky, J., Bhattacharya, S., Butte, A.J.: Opening clinical trial data: are the voluntary data-sharing portals enough? BMC Med. **13**, 280 (2015)
7. NCI Genomic Data Commons. National Cancer Institute, Bethesda. https://gdc.cancer.gov/ (2017). Last accessed 8 Oct 2017
8. NCI Genomic Data Commons. National Cancer Institute, Bethesda. https://gdc.cancer.gov/ (2017). Last accessed 8 Oct 2017
9. MMRF Researcher Gateway. Multiple Myeloma Research Foundation, Norwalk. https://research.themmrf.org/ (2017). Last accessed 8 Oct 2017
10. Project Data Sphere. Project Data Sphere, LLC. https://www.projectdatasphere.org/ (2017). Last accessed 14 Aug 2017
11. Wendling, T., Mistry, H., Ogungbenro, K., Aarons, L.: Predicting survival of pancreatic cancer patients treated with gemcitabine using longitudinal tumour size data. Canc. Chemo. and Pharmacol. **77**, 927–938 (2016)
12. Gill, B., Khoja, L., Hamilton, R.J., Abdallah, K., Pintilie, M., Joshua, A.M.: Project data sphere (PDS) in prostate cancer: a first look including concomitant medication use. Bone **1144**, 19–78 (2015)
13. Geifman, N., Butte, A.J.: A patient-level data meta-analysis of standard-of-care treatments from eight prostate cancer clinical trials. Sci. Data **3**, 160027 (2016)
14. Abdallah, K., Hugh-Jones, C., Norman, T., Friend, S., Stolovitzky, G.: The Prostate Cancer DREAM Challenge: a community-wide effort to use open clinical trial data for the quantitative prediction of outcomes in metastatic prostate cancer. Oncologist **20**, 459–460 (2015)
15. Green, A.K., Corty, R.W., Wood, W.A., Meeneghan, M., Reeder-Hayes, K.E., Basch, E., Milowsky, M.I., Dusetzina, S.B.: Comparative effectiveness of mitoxantrone plus prednisone versus prednisone alone in metastatic castrate-resistant prostate cancer after docetaxel failure. Oncologist **20**, 516–522 (2015)
16. Romero, K., Ito, K., Rogers, J.A., Polhamus, D., Qiu, R., Stephenson, D., Mohs, R., Lalonde, R., Sinha, V., Wang, Y., Brown, D.: The future is now: Model-based clinical trial design for Alzheimer's disease. Clin. Pharmacol. Ther. **97**, 210–214 (2015)
17. Fijal, B.A., Hall, J.M., Witte, J.S.: Clinical trials in the genomic era: effects of protective genotypes on sample size and duration of trial. Contemp. Clin. Trials. **21**, 7–20 (2000)
18. Williamson, F: Using External Patient Data in Clinical Trial Simulation. Paper presented at the Joint Statistical Meetings, session 530, McCormick Place, Chicago 30 July–4 August (2016)
19. Pocock, S.J.: The combination of randomized and historical controls in clinical trials. J. Chronic Dis. **29**, 175–188 (1976)
20. Signorovitch, J.E., Wu, E.Q., Andrew, P.Y., Gerrits, C.M., Kantor, E., Bao, Y., Gupta, S.R., Mulani, P.M.: Comparative effectiveness without head-to-head trials. Pharmacoeconomics **28**, 935–945 (2010)
21. Caro, J.J., Ishak, K.J.: No head-to-head trial? Simulate the missing arms. Pharmacoeconomics **28**, 957–967 (2010)

22. Dimopoulos, M.A., Orlowski, R.Z., Facon, T., Sonneveld, P., Anderson, K.C., Beksac, M., Benboubker, L., Roddie, H., Potamianou, A., Couturier, C. and Feng, H.: Retrospective matched-pairs analysis of bortezomib plus dexamethasone versus bortezomib monotherapy in relapsed multiple myeloma. Haematologica, 112037 (2014)
23. Selaru, P., Tang, Y., Huang, B., Polli, A., Wilner, K.D., Donnelly, E., Cohen, D.P.: Sufficiency of single-arm studies to support registration of targeted agents in molecularly selected patients with cancer: lessons from the clinical development of Crizotinib. Clin. Trans. Sci. **9**, 63–73 (2016)
24. Rubin, D.B.: Matching to remove bias in observational studies. Biometrics **29**, 159–183 (1973)
25. Kawamura, K., Ichikado, K., Suga, M., Yoshioka, M.: Efficacy of azithromycin for treatment of acute exacerbation of chronic fibrosing interstitial pneumonia: a prospective, open-label study with historical controls. Respiration **87**, 478–484 (2014)
26. Gökbuget, N., Kelsh, M., Chia, V., Advani, A., Bassan, R., Dombret, H., Doubek, M., Fielding, A.K., Giebel, S., Haddad, V., Hoelzer, D.: Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. Blood Cancer J. **6**, 473 (2016)
27. Robins, J.M., Rotnitzky, A., Zhao, L.P.: Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc. **89**, 846–866 (1994)
28. Zelen, M.: The randomization and stratification of patients to clinical trials. J. Chronic Dis. **27**, 365–375 (1974)
29. Yusuf, S., Wittes, J., Probstfield, J., Tyroler, H.A.: Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA **266**, 93–98 (1991)

Novel Test for the Equality of Continuous Curves with Homoscedastic or Heteroscedastic Measurement Errors

Zhongfa Zhang, Yarong Yang and Jiayang Sun

Abstract Testing equality of two curves occurs often in functional data analysis. In this paper, we develop procedures for testing if two curves measured with either homoscedastic or heteroscedastic errors are equal. The method is applicable to a general class of curves. Compared with existing tests, ours does not require repeated measurements to obtain the variances at each of the explanatory values. Instead, our test calculates the overall variances by pooling all of the data points. The null distribution of the test statistic is derived and an approximation formula to calculate the p value is developed when the heteroscedastic variances are either known or unknown. Simulations are conducted to show that this procedure works well in the finite sample situation. Comparisons with other test procedures are made based on simulated data sets. Applications to our motivating example from an environmental study will be illustrated. An R package was created for ease of general applications.

Keywords Functional data analysis · Hypothesis test · Local regression · Tube formula

2011 MSC Primary · 62G08 · 62J02. Secondary · 93E14 · 62G10 · 62H15

1 Curve Test and its Motivating Example: An Environmental Study—The Lead Project

In this paper, we present a new test procedure to formally test if two continuous curves measured with errors are statistically equal over the defining interval.

Before we introduce the model, we will first introduce a real motivating study for developing the test procedures.

Z. Zhang (✉) · J. Sun
Department of Statistics, Case Western Reserve University, Cleveland, OH 44106, USA
e-mail: jacob.zhang@takeda.com

Y. Yang
Department of Statistics, North Dakota State University, Fargo, ND 58104, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_20

Studies have demonstrated a close adverse relationship between lead exposure and children's health see, for example, [27]. Among major sources of lead exposure identified [1] are leaded paint, leaded gasoline, food, drinking water, industry wastes and industry products etc. However, it is hard to identify the major source(s) of the lead concentrations over time, in part due to the difficulties in obtaining long-term high quality lead concentrations in children as well in their grown-up environment retrospectively. Fortunately, there is a high correlation between lead concentration measurements in children and that in their teeth. Once a tooth has finished its development, the lead concentration in it will remain unchanged throughout one's life. By measuring the lead concentration in teeth, we can trace the lead concentration in a child's blood back to the year when the corresponding tooth was formed. This lead concentration in blood will then be used as surrogate for lead exposure of child in that year to be correlated with child's health and lead concentrations from different sources.

Each patient who participated in the study had either his/her first (molar 1) or second molar (molar 2) extracted, and then the lead concentration as well as lead isotopic ratios from that tooth were measured. The times for half-maximal enamel formation for molar 1 and 2 are different, at about age 2 and 6 respectively. From this point forward, we will refer to a patient as a member in group $M1$ if his/her first molar was used and in group $M2$ otherwise.

The scatter plot of lead concentrations for the two groups is displayed in Fig. 1a. From the plot, we see that points from the two groups are well mixed with each other. Lead concentrations for both groups generally increased during the phase-in period (1936–1960), then decreased afterward during the phase out period (1960–1990). The phase-in and phase-out periods were mainly referring to the time periods when the leaded gasoline was introduced and then phased out gradually till its totally ban by EPA (1996).

In Fig. 1b, we imposed smoothing curves fitted by a local regression method.

To answer the question as of whether the two curves are statistically equal, we need develop test procedures to do this.

The outline for the rest of this paper is as follows. In Sect. 2, we will formally introduce the statistical model for the test procedure, following a brief literature review in the field in Sect. 3. In Sect. 4, we introduce a few lemmas that are needed to prove our theorems. In Sect. 5, we describe our test procedures, which are based on the Tube formula and local regressions. Simulation studies are made in Sect. 7 and performance comparison with other test procedures will be presented in Sect. 8. Test procedures are then applied to the motivating (teeth) data set mentioned above in Sect. 9.

2 Model Setup

In this section, we will formally introduce some statistical notations and model setups to be used for the development of the proposed test procedures that will be discussed in the next few sections.

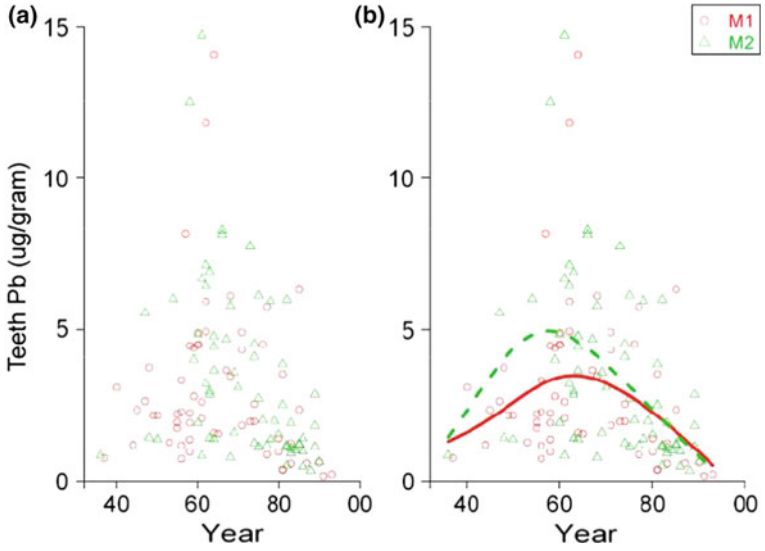


Fig. 1 Scatter plot of teeth enamel lead concentrations (in μg/gram) against year (1936–2000) with (**a**) or without (**b**) local smoothing curves superimposed for each group

We formalize our model as following. Suppose

$$Y_1(t) = f_1(t) + \varepsilon_1(t), \tag{2.1}$$

and

$$Y_2(t) = f_2(t) + \varepsilon_2(t), \tag{2.2}$$

where $\varepsilon_1(t)$ and $\varepsilon_2(t)$ are two independent homogeneous Gaussian random errors indexed by t with means $E(\varepsilon_1(t)) = E(\varepsilon_2(t)) = 0$ and variances $Var(\varepsilon_1(t)) = \sigma_1^2$ and $Var(\varepsilon_2(t)) = \sigma_2^2$ for all t. The term also implies that the errors at any t_1, t_2 with $t_1 \neq t_2$, $\varepsilon_i(t_1)$ and $\varepsilon_i(t_2)$ are independent for both $i = 1, 2$.

Of interest is to test $H_0 : f_1(t) = f_2(t)$ for all $t \in \mathcal{T}$ versus $H_1 : f_1(t) \neq f_2(t)$ for at least one $t \in \mathcal{T}$, for some domain $\mathcal{T}$ (e.g., an interval).

We assume that both $f_1(t)$ and $f_2(t)$ are smooth functions with continuous derivatives up to order 2. Such restrictions on choices of f are reasonable and sometimes necessary.

One special case occurs when $f_2(t) \equiv 0$ (or any other nonzero constant c). In this case, we are testing if curve $Y_1(t)$ is statistically different from (homogeneous) Gaussian random errors.

The observed data are $\{(t_{1,i}, Y_{1,i}),\ i = 1, \ldots, n_1\}$ for model (2.1) and $\{(t_{2,j}, Y_{2,j})$ $j = 1, \ldots, n_2\}$ for model (2.2), where sample sizes n_1 and n_2 may or may not equal. The errors $\epsilon_{i,j}$ are assumed to be independent for $i = 1, 2, \quad j = 1, \ldots, n_i$.

Notice here that the two sequences of $t_{1,i}$ and $t_{2,i}$ are not necessarily the same(but they do need to be on the same support interval). At any given t value, our model

does allow only one response value, so no repeated measurements are required. This adds more flexibility in applications than other procedures, which we will introduce in the next section.

3 Related Works of Functional Data Analysis

In this section, we will give a brief literature review and related works in this field.

Studies that extend numerical data analysis to functional data analysis can be traced back to Parzen [18]. Ramsay and Daizell [19] coined the term *Functional Data Analysis* (FDA) to distinguish it from ordinary data analysis. Since then, several people have tried doing research on various aspects in the field. For example, [14] considered canonical correlation analysis when the data are curves. James and Hastie [9] discussed functional linear discriminant analysis for irregularly sampled curves. A related work was done by Kitska [12], in which he extended the ANOVA test procedure from discrete time to continuous time, termed his method Functional ANOVA (FANOVA). But this method applies only when the underlying regression is linear. See a recent summary in [26]. More closely related to our present research is work done by [7], who gave a procedure to test significance of differences between two curves by executing a Fourier or Wavelet transformation (on the curves), and then using the partial coefficients to test the hypothesis. We call this an *indirect* method since the test is based on transformed functions (curves) rather than on the original functions (curves) themselves. This test procedure requires the two curves to be on the same supporting sets (i.e, t values), but for each t value, there should have multiple response values (repeated measurements) from each group being tested. The algorithm will fail if there is only one response value at some point, because the procedure will fail to calculate the variance of response at that point. In this paper, we present a direct method to test the hypothesis by estimating the involved tail probability *directly* under the null.

4 A Few Lemmas

In this section, we will introduce a few lemmas which will be needed to prove our main theorem described in the next section.

Lemma 4.1 *Assume that* $X_1 \sim \chi^2_{n_1}$ *and* $X_2 \sim \chi^2_{n_2}$ *are independent random variables having* χ^2 *distribution with degree of freedom* n_1 *and* n_2 *respectively, where* n_1 *and* n_2 *are relatively large. Let* $G_i = X_i(X_1 + X_2)^{-1}/(n_i(n_1 + n_2)^{-1})$. *Then* $G_i \to 1$ *in distribution. In particular, for* $i = 1, 2$,

$$E\{G_i\} \to 1, \quad E\left\{G_i^2\right\} \to 1. \tag{4.1}$$

Proof See Appendix A.

Now suppose that X_1, X_2, S_1, S_2 are four independent random variables such that $X_i \sim \chi^2_{ni}$ and $\nu_i S_i \sim \chi^2_{\nu_i}$ for $i = 1, 2$. Let $n = n_1 + n_2$ and

$$Y = \frac{X_1 + X_2}{X_1/S_1 + X_2/S_2}. \tag{4.2}$$

Lemma 4.2 *Assume $n_i > \nu_i$ for $i = 1, 2$. Under the conditions of Lemma 4.1, $Y \sim \chi^2_\nu/\nu$ approximately with degree of freedom ν estimated by formula*

$$\nu = \frac{n^2\nu_1\nu_2}{n_2^2\nu_1 + n_1^2\nu_2}. \tag{4.3}$$

Proof See Appendix B.

To see how close the approximated distribution of Y stated in this lemma is to the true distribution of Y, we did some numeric study based on generated data. First, we choose ν_i, n_i for $i = 1, 2$. Then we generate random samples of X_i and S_i following the distribution described in this lemma, for $i = 1, 2$. The corresponding Y calculated by formula (4.2) will serve our true samples of Y. A random sample drawn from a χ^2_ν/ν distribution, where ν is calculated by formula (4.3) will be our approximated samples of Y. For $\nu_1 = 120$, $\nu_2 = 300$, $n_1 = 800$ and $n_2 = 1000$, we have generated Fig. 2. In this figure, the density curve of χ^2_ν/ν, where ν is calculated by (4.3) is presented as dashed curve. The true density curve of Y is presented as solid black curve. To compare, the density curves of X_1 and X_2 are also displayed in the plot.

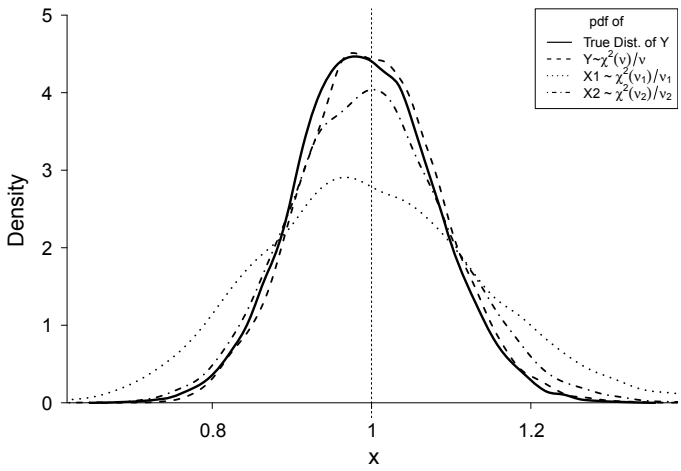


Fig. 2 The true density of Y(solid black) by numerical method on generated data and the density of χ^2_ν/ν (dashed), with ν calculated by formula (4.3). The density curves of $\chi^2_{\nu_i}/\nu_i$ are also added on the plot. $n_1 = 800$, $n_2 = 1000$, $\nu_1 = 120$, $\nu_2 = 300$

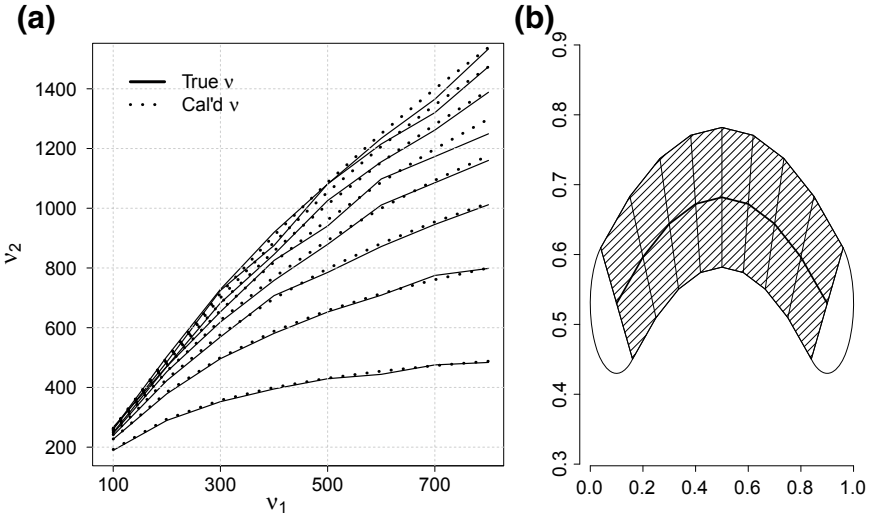


Fig. 3 **a** Compare the degrees of freedom ν calculated by formula (4.3) (dotted lines) and the degrees of freedom ν by generated data with $\nu = 4\pi m^2$ (as true values, solid line) with different combination of values $\nu_1 = 100, 200, \ldots, 800$ (x-axis) and $\nu_2 = 200, \ldots, 1500$ (from bottom curve up). Here $n_1 = 1000, n_2 = 1500$. **b** Tube with 2 endpoints around a 1-dimensional manifold (dark line in the middle) embedded in R^2. This tube has its main part (shaded area) and 2 boundaries (white areas)

By allowing the 4 parameters to take a sequence of values of its own, we can compare how close the two distributions are under different cases, by just comparing the degrees of freedom of the two χ_2 distributions. The estimated degrees of freedom were obtained by the following steps. First, draw samples of $X_i \sim \chi^2_{n_i}, S_i \sim \chi^2_{\nu_i}, \quad i = 1, 2.$ independently, then compute Y by formula (4.2) to get a sample of Y. Repeat this step K times to get a set of $i.i.d.$ samples of Y of size K. Then its pdf $p(y)$ is estimated and its peak value $m = \max_y p(y)$ is calculated and ν is calculated based on formula $\nu = 4\pi m^2$. This formula is obtained theoretically through finding the relationship between the degree of freedom ν and the peak value of density function of a χ^2_ν/ν distribution. Our simulation plot (Fig. 3a) shows that they agree with each other. See also Table 1 of numerical version of Fig. 3a.

To prove our Theorems, we also need the following definition and lemma.

Definition 4.3 (Tube) A tube T, with radius r of a manifold $\mathcal{M}(t) = \{m^n(t) := (m_1(t), m_2(t), \ldots, m_n(t)), t \in \mathcal{T}\}$ embedded in n-dimensional space X (either Euclidean space R^n or Spherical surface $S^{n-1} \subset R^n$) is defined to be the set of all points $x \in X$ such that $d(x, \mathcal{M}(t)) \leq r$, or $T = T(r) = \{x \in X : \ d(x, y) \leq r for \ at least \ one \ y \in \mathcal{M}(t)\}$, where d is the usual Euclidean distance. It is one dimensional if the domain $\mathcal{T}$ of t is 1-dimensional.

Table 1 Comparison of true degrees of freedom (upper element) with calculated degrees of freedom ν (lower element, as estimated d.f. via formula 4.3) for different combinations of degrees of freedom of ν_1 and ν_2

ν_1	ν_2							
	100	200	300	400	500	600	700	800
100	192.7	232.6	242.4	253.6	252.8	258.6	264	268.5
	192.3	227.3	241.9	250	255.1	258.6	261.2	263.2
200	289.5	379	422.9	449.6	471.8	468.7	497.2	500
	294.1	384.6	428.6	454.5	471.7	483.9	493	500
300	354.8	496.2	580.1	625.4	643.1	677.4	705.3	732
	357.1	500	576.9	625	657.9	681.8	700	714.3
400	407.3	570.9	697.8	763.9	808.8	844.7	889.8	896.7
	400	588.2	697.7	769.2	819.7	857.1	886.1	909.1
500	422.2	672.3	789.9	879.5	941	1013.1	1038.9	1098.7
	431	657.9	797.9	892.9	961.5	1013.5	1054.2	1087
600	456.1	721.3	868.2	1011.5	1062.6	1165.2	1180.2	1256.9
	454.5	714.3	882.4	1000	1087	1153.8	1206.9	1250
700	471.6	768.9	943.5	1095.5	1204.6	1256.3	1355.2	1409.4
	473	760.9	954.5	1093.8	1198.6	1280.5	1346.2	1400
800	486.7	793.8	1040	1152.3	1267.9	1433.2	1451.3	1534.3
	487.8	800	1016.9	1176.5	1298.7	1395.3	1473.7	1538.5

Since the distance d in R^n can be equivalently defined by the inner product:

$$d(x, y)^2 = \langle x - y, x - y \rangle, \tag{4.4}$$

where $x, y \in R^n$, the tube can also be defined by the inner product. When $x, y \in S^{n-1}$, the relation between distance and inner product reduces to:

$$d(x, y)^2 = 2 - 2\langle x, y \rangle.$$

The volume $vol(T)$ of a tube T can be roughly divided into two parts. The first part is the (main) tubular area in the middle (shaded area), while the second part is the one resulting from boundary correction if the manifold has boundaries (unshaded white areas). See the illustration picture in Fig. 3b. This problem of calculating the volume of a tube was first proposed and solved by Hotelling [8] in a 1-dimensional manifold situation. Weyl [25] extended the results to higher dimensional manifolds; that is, when $\mathcal{M}(t)$ is a surface. Other publications on the subject include [11, 13, 16, 17, 23].

When $X = S^{n-1}$ is the surface of the unit sphere in R^n and the manifold is a curve, we have:

Lemma 4.4 *The volume of tube can be obtained by the following formula:*

$$Vol(T) = \frac{\kappa_0 A_n}{2\pi} P(\beta_{1,\ (n-2)/2} \geq w^2) + \frac{EA_n}{4} P(\beta_{1/2,\ (n-1)/2} \geq w^2), \tag{4.5}$$

where $\beta_{a,b}$ denotes a random variable following a β distribution with parameters a and b, $A_n = 2\pi^{n/2}/\Gamma(n/2)$ is the surface area of the unit sphere in R^n, $w = 1 - r^2/2$, κ_0 is the length of the manifold $\mathcal{M}(t)$ as before and E is the Euler characteristic number (or the number of end points of $\mathcal{M}(t)$ in this case).

This theorem was provided by Knowles and Siegmund [13], extended by Sun and Loader [22] [2.5, p1331] to cases when the dimension of the manifold $d \geq 2$:

$$\begin{aligned} &Pr\left\{\sup_{x\in\mathcal{X}} < T(x), U >\geq w\right\} \\ &= \kappa_0 J_0(w) + \frac{\xi_0}{2} J_1(w) + \frac{\kappa_2 + \xi_1 + m_0}{2\pi} J_2(w) + O((1-w^2)^{(n-d+2)/2}), \end{aligned} \tag{4.6}$$

where $J_e(w) = (A_{n-d+e-1}/A_n) \int_w^1 (1-u^2)^{(n-d+e-3)/2} u^{d-e} du$, $e = 0, 1, 2$.

It is easy to see the corresponding relationship between formula (4.5) and (4.6): apart from a constant A_n, which is needed to be divided from the volume of tube to get the probability, $J_0(w) = 1/(2\pi) P(\beta_{1,\ (n-2)/2} \geq w^2)$, $J_1(w) = P(\beta_{1/2,\ (n-1)/2} \geq w^2)/2$. Formula (4.5) thus takes only the first two terms in formula (4.6).

5 Theorems and Test Procedures

In this section, we will outline a few theorems and describe the test procedures when different model assumptions arise. The lengthy proofs of these theorems will be given in Appendix.

Remember we are testing: $H_0 : f_1(t) = f_2(t)$ for all $t \in \mathcal{T}$ versus $H_1 : f_1(t) \neq f_2(t)$ for at least one $t \in \mathcal{T}$, for some domain $\mathcal{T}$ based on observed data sets $\{(t_{i,j}, Y_{i,j}),\ i = 1, 2,\ j = 1, \ldots, n_i\}$, where $Y_{i,j} = f_i(t_{i,j}) + \epsilon_{i,j}$ with $\epsilon_{i,j} \sim N(0, \sigma_i^2)$ assumed to be independent for $i = 1, 2,\ \ j = 1, \ldots, n_i$. σ_i and σ_2 may or may not equal. In case $\sigma_1 = \sigma_2$, we have the homoscedastic errors, while otherwise, we have the heteroscedastic errors.

5.1 Homoscedastic Case

Here we assume the variances of the two homogeneous Gaussian random errors ε_i, $i = 1, 2$ in models (2.1) and (2.2) are equal, i.e., we assume that $\sigma_1^2 = \sigma_2^2 := \sigma^2$.

Consider the quadratic local regression estimation. Then f_1 can be obtained by solving the following optimal problem

$$\min\left\{(a_0,a_1):\sum_{i=1}^{n_1}W(\frac{t-t_{1.i}}{h})(Y_{1,i}-a_0-a_1(t_{1,i}-t))^2\right\},$$

where $W(t)$ is a kernel function and h is the window width with $h>0$ and $h\to 0$.

The estimated $f_1(t)$ from model (2.1) is

$$\hat{f}_1(t)=\hat{a}_0=\sum_{i=1}^{n_1}l_{1,i}(t)Y_{1,i}=\langle \mathbf{l}_1(t),\mathbf{Y}_1\rangle, \tag{5.1}$$

where $\mathbf{Y}_1=(Y_{1,1},\ldots,Y_{1,n_1})'$, $\mathbf{l}_1(t)=(l_{1,i}(t),\ i=1,\ldots,n_1)'$,

$$l_{1,i}(t)=\frac{w_{1,i}\sum_j w_{1,j}(t-t_{1,j})(t_{1,i}-t_{1,j})}{\sum_i\left\{w_{1,i}\sum_j w_{1,j}(t-t_{1,j})(t_{1,i}-t_{1,j})\right\}}, \tag{5.2}$$

and $w_{1,i}=W((t-t_{1.i})h^{-1})$ for $i=1,\ldots,n_1$.

Similarly, the estimated $f_2(t)$ by quadratic local regression for model (2.2) can be expressed

$$\hat{f}_2(t)=\sum_{i=1}^{n_2}l_{2,i}(t)Y_{2,i}=\langle \mathbf{l}_2(t),\mathbf{Y}_2\rangle, \tag{5.3}$$

where $\mathbf{Y}_2,\mathbf{l}_2(t)$ are similarly defined.

If both estimators $\hat{f}_i(t)$ are unbiased, i.e., $E\hat{f}_i(t)=f_i(t)=\sum_j l_{i,j}(t)\mu_{i,j}$ for $i=1,2$, where $\mu_{i,j}=EY_{i,j}$ for all i,j. Then for $i=1,2$,

$$\begin{aligned}E\hat{f}_i(t) &= E(\sum_j l_{i,j}(t)Y_{i,j})=\sum_j l_{i,j}(t)\mu_{i,j}=\langle \mathbf{l}_i(t),\mu_i\rangle,\\ Var(\hat{f}_i(t)) &= \sigma^2\sum_j l_{i,j}^2(t)=\sigma^2\|\mathbf{l}_i(t)\|_2^2,\end{aligned}$$

where $\|\cdot\|_2$ denotes the L^2 norm.

Because $\varepsilon_1(t)$ and $\varepsilon_2(t)$ are assumed to be independent, $Var(\hat{f}_1(t)-\hat{f}_2(t))=\sigma^2(\sum l_{1,i}^2(t)+\sum l_{2,i}^2(t))=\sigma^2(\|\mathbf{l}_1(t)\ \|^2+\|\mathbf{l}_2(t)\|^2)$ and the standard deviation of $\hat{f}_1(t)-\hat{f}_2(t)$ is $sd(\hat{f}_1(t)-\hat{f}_2(t))=\sigma\sqrt{\|\mathbf{l}_1(t)\|^2+\|\mathbf{l}_2(t)\|^2}$.

Estimate of the standard deviation σ. Let $\hat{\varepsilon}_\mathbf{i}:=(\hat{\varepsilon}_{i,1},\hat{\varepsilon}_{i,2},\ \ldots,\hat{\varepsilon}_{i,n_i})'=(Y_{i,1}-\hat{Y}_{i,1},Y_{i,2}-\hat{Y}_{i,2},\ldots,Y_{i,n_i}-\hat{Y}_{i,n_i})'$, where $\hat{Y}_{i,j}=\hat{f}_i(t_{i,j})$, for $i=1,2$. Let

$$L_i = \begin{pmatrix} l_{i,1}(t_{i,1}) & l_{i,2}(t_{i,1}) & \dots & l_{i,n_i}(t_{i,1}) \\ l_{i,1}(t_{i,2}) & l_{i,2}(t_{i,2}) & \dots & l_{i,n_i}(t_{i,2}) \\ \dots & \dots & \dots & \dots \\ l_{i,1}(t_{i,n_i}) & l_{i,2}(t_{i,n_i}) & \dots & l_{i,n_i}(t_{i,n_i}) \end{pmatrix}$$

be the matrix such that $\hat{\mathbf{Y}}_i = L_i \mathbf{Y}_i$. Therefore, $\hat{\varepsilon}_i = (I_i - L_i)Y_i$, where I_i is the identity matrix of order n_i. Let $A_i = (I_i - L_i)(I_i - L_i)'$. Cleveland and Devin [5] showed that $E(\hat{\varepsilon}_i'\hat{\varepsilon}_i) = \sigma^2 tr(A_i)$, $Var(\hat{\varepsilon}_i'\hat{\varepsilon}_i) = 2\sigma^4 tr(A_i^2)$.

Also for $i = 1, 2$, let $\delta_{i,1} := tr(A_i), \quad \delta_{i,2} := tr(A_i^2)$, and let

$$\nu_i := \delta_{i,1}^2/\delta_{i,2}, \quad \nu = \nu_1 + \nu_2. \tag{5.4}$$

Then by equating the first two moments of random variables from both sides, it can be shown that $((\hat{\varepsilon}_i'\hat{\varepsilon}_i)\delta_{i,1})(\sigma^2\delta_{i,2})^{-1} \sim_{approx} \chi^2_{\nu_i}$, for $i = 1, 2$.

Since $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$ are independent, we have $\sum_{j=1}^{2}((\hat{\varepsilon}_j'\hat{\varepsilon}_j)\delta_{j,1})(\sigma^2\delta_{j,2})^{-1} \sim_{approx} \chi^2_\nu = \chi^2_{\nu_1+\nu_2}$. If we estimate σ^2 by

$$\hat{\sigma}^2 = \frac{(\hat{\varepsilon}_1'\hat{\varepsilon}_1)\delta_{1,1}}{\nu\delta_{1,2}} + \frac{(\hat{\varepsilon}_2{}'\hat{\varepsilon}_2)\delta_{2,1}}{\nu\delta_{2,2}}, \tag{5.5}$$

then $\nu\hat{\sigma}^2/\sigma^2 \sim \chi^2_\nu$.

This expression of estimated σ^2 can be viewed in another way. Let $\hat{\sigma}_i^2 = (\hat{\varepsilon}_i'\hat{\varepsilon}_i)/\delta_{i,1}$, then $(\nu_i\hat{\sigma}_i^2)/\sigma_i^2 \sim \chi^2_{\nu_i}$ approximately for $i = 1, 2$. The right hand side of Eq. (5.5) can thus be written as:

$$\hat{\sigma}^2 = \frac{\nu_1\hat{\sigma}_1^2 + \nu_2\hat{\sigma}_2^2}{\nu_1 + \nu_2}.$$

Therefore, our estimate of $\hat{\sigma}^2$ is a *weighted* average of estimates of the two individual variances. This mimics the pooled estimation of variance in the two sample t-test situation.

Let

$$Z(t) := \frac{(\hat{f}_1(t) - \hat{f}_2(t)) - (f_1(t) - f_2(t))}{sd(\hat{f}_1(t) - \hat{f}_2(t))}. \tag{5.6}$$

Under $H_0 : f_1(t) = f_2(t)$, $EZ(t) = 0$, $Var(Z(t)) = 1$ for all $t \in \mathcal{T}$.

$Z(t)$ is approximately a Gaussian random field. Let

$$u_i(t) := \frac{\mathbf{l}_i(t)}{\sqrt{\|\mathbf{l}_1(t)\|_2^2 + \|\mathbf{l}_2(t)\|_2^2}}, \qquad i = 1, 2.$$

Then $Z(t)$ can be expressed in terms of $u_1(t)$ and $u_2(t)$:

$$
\begin{aligned}
Z(t) =& \frac{(\hat{f}_1(t) - f_1(t)) - (\hat{f}_2(t) - f_2(t))}{sd(\hat{f}_1(t) - \hat{f}_2(t))} = \frac{\hat{f}_1(t) - f_1(t)}{sd(\hat{f}_1(t) - \hat{f}_2(t))} - \frac{\hat{f}_2(t) - f_2(t)}{sd(\hat{f}_1(t) - \hat{f}_2(t))} \\
=& \langle \frac{\mathbf{l}_1(t)}{\sqrt{\|\mathbf{l}_1(t)\|_2^2 + \|\mathbf{l}_2(t)\|_2^2}}, \frac{\mathbf{Y}_1 - E\mathbf{Y_1}}{\sigma} \rangle - \langle \frac{\mathbf{l}_2(t)}{\sqrt{\|\mathbf{l}_1(t)\|_2^2 + \|\mathbf{l}_2(t)\|_2^2}}, \frac{\mathbf{Y}_2 - E\mathbf{Y}_2}{\sigma} \rangle \\
=& \langle u_1(t), \frac{\varepsilon_\mathbf{1}}{\sigma} \rangle - \langle u_2(t), \frac{\varepsilon_\mathbf{2}}{\sigma} \rangle = \langle u_1(t), \xi_1 \rangle - \langle u_2(t), \xi_2 \rangle ,
\end{aligned}
$$

where $\xi_i = \varepsilon_\mathbf{i}/\sigma,\ i = 1, 2$ are multivariate standard normal (following a multivariate $N(0, 1)$ distribution), and are independent to each other.

The correlation function $\rho(t, t')$ of the random field $Z(t)$ is computed by

$$\rho(t, t') := corr(Z(t), Z(t')) = \langle u_1(t), u_1(t') \rangle + \langle u_2(t), u_2(t') \rangle \tag{5.7}$$

by the fact that ξ_i's are independent standard normal and that $u_i(t) \in S^{n-1}$ for $i = 1, 2$.

Let us return to our primary test problem. Recall that we want to test $H_0 : f_1(t) = f_2(t)$ for all $t \Leftrightarrow H_1 : f_1(t) \neq f_2(t)$ for at least one t at a pre-specified level α. Consider the test statistic

$$T = max_{t \in \mathcal{T}} \|Z(t)\|. \tag{5.8}$$

If the realized $T = t_0$ is too large, we reject the null hypothesis. More specifically, we need to find the (tail) probability $Pr_{H_0}(T > t_0)$. If this probability is larger than α, we accept the null and declare that there is no difference between curves $f_1(t)$ and $f_2(t)$. Otherwise, we will reject the null hypothesis. The key issue is how to estimate the tail probability $Pr_{H_0}(T > t_0)$. We offer the following theorem to estimate this probability, which is generalization of the Theorem in [24].

Theorem 5.1 (Tail Probability Estimation for Homoscedastic Case) *Suppose* $\mathcal{T} = [a, b]$ *and* $\hat{f}_1(t)$ *and* $\hat{f}_2(t)$ *are unbiased estimates of* $f_1(t)$ *and* $f_2(t)$, *and* $\mathbf{l}_1(t)$ *and* $\mathbf{l}_2(t)$ *are defined in (5.2) and (5.3) respectively. If* σ^2 *is known, then*

$$Pr_{H_0}(T > t_0) \approx \frac{\kappa_0}{\pi} \exp(-\frac{t_0^2}{2}) + E(1 - \Phi(t_0)) \qquad as\, t_0 \to \infty. \tag{5.9}$$

If σ^2 *is unknown and is estimated by* $\hat{\sigma}^2$ *in (5.5) so that* $\nu\hat{\sigma}^2/\sigma^2 \sim \chi_\nu^2$, *then*

$$Pr_{H_0}(T > t_0) \approx \frac{\kappa_0}{\pi}(1 + \frac{t_0^2}{\nu})^{-\nu/2} + \frac{E}{2} P(|t_\nu| > t_0) \qquad as\, t_0 \to \infty, \tag{5.10}$$

where $t_\nu, \nu \neq 0$ *follows a standard t distribution with degree of freedom* ν,

$$\kappa_0 = \int_{\mathcal{T}} |C(t)|^{1/2} dt, \tag{5.11}$$

$C = \partial\rho(t,t')/\partial t\partial t'|_{t'=t}$, and E is the Euler-Poincare characteristic of manifold $\mathcal{M}(t) = (u_1(t), -u_2(t))$ from $\mathcal{T}$ to $\mathcal{S}^{n-1} = \{x \in \mathcal{R}^n : \|x\| = 1\}$, the $n = n_1 + n_2$ dimensional unit surface. $E = 0$ if $\mathcal{M}(a) = \mathcal{M}(b)$, $E = 2$ if $\mathcal{M}(a) \neq \mathcal{M}(b)$ and $\mathcal{M}$ has no self-overlap.

Proof See Appendix C.

When σ^2 is known, $Z(t) =< u_1(t), \xi_1 > - < u_2(t), \xi_2 >$ in (5.6) is actually a finite Karhunen-Loeve expansion of the random field $Z(t)$. The covariance function $\rho(s,t)$ of $Z(t)$ in formula (5.7) has a finite expansion (of up to n terms of form $Z_{i,j}(s)Z_{i,j}(t)$, for $i = 1, 2$, $j = 1, \ldots, n_i$). The constant κ_0 can be approximated by the following formula.

Computation of κ_0: If $\mathcal{T} = [a, b]$, and is partitioned into k small intervals $a = t_0 < t_1 < t_2 < \cdots < t_k = b$ such that $max_i|t_i - t_{i-1}| \to 0$ as $k \to \infty$. Then κ_0 can be approximated by

$$\begin{aligned}\kappa_0 &= \int_{\mathcal{T}} |C(t)|^{1/2} dt \\ &\approx \sum_{i=1}^{k} [\|u_1(t_i) - u_1(t_{i-1})\|_2^2 + \|u_2(t_i) - u_2(t_{i-1})\|_2^2]^{1/2}, \end{aligned} \tag{5.12}$$

where $\|.\|_2$ as before denotes the L^2 norm. Its computation is often straight forward.

5.2 Special Case when $f_2(t) \equiv 0$

This is a much simpler version of the previous problem and is equivalent to a normality test (over a continuous domain). Studies of this problem can be found in James and Stein [10], Shapiro and Wilk [21], Chakravarti et al. [3], but with a different setup.

Throughout this section, we will suppress the group subscripts to simplify our notation.

Our test procedure will first use local regression to estimate $f(t)$ for a selected window width h and kernel function $W(t)$ as before. The estimated curve $f(t)$ can be expressed

$$\hat{f}(t) = \sum_{i=1}^{n} l_i(t) Y_i = \langle \mathbf{l}(t), \mathbf{Y} \rangle, \tag{5.13}$$

where $\mathbf{Y}$ and $\mathbf{l}(t)$ are defined as before in (5.2), with index i being omitted.

Let $T(t) = \mathbf{l}(t)/\|\mathbf{l}(t)\|$. Theorem (5.1) is still valid and κ_0 can be calculated by $\kappa_0 = \int_{\mathcal{T}} \|T'(t)\| dt$. Similar to the estimation formula of κ_0 in (5.12), we can estimate κ_0 by $\kappa_0 \approx \sum_{i=1}^{k} \|T(t_i) - T(t_{i-1})\|$, if $\mathcal{T} = [a, b]$ and it is partitioned into k intervals $a = t_0 < t_1 < t_2 < \cdots < t_k = b$ with $max_i|t_i - t_{i-1}| \to 0$ as $k \to \infty$. E is similarly defined as it was in Theorem 5.10. QED

5.3 Heteroscedastic Case

In this case, the assumption $\sigma_1^2 = \sigma_2^2$ is no longer valid. Instead, we assume the two fitted models have different standard deviations for its error terms (see Sect. 2 for detail).

Let $\hat{f}_i(t) = \sum_{j=1}^{n_i} l_{i,j}(t) Y_{i,j}$, for $i = 1, 2$, be estimated as before in the homoscedastic case (note the estimate of $f_i(t)$ does not depend on σ_i^2's at all) and let

$$Z(t) := \frac{[\hat{f}_1(t) - f_1(t)] - [\hat{f}_2(t) - f_2(t)]}{sd(\hat{f}_1(t) - \hat{f}_2(t))}$$

$$u_i(t) := \frac{\sigma_i \mathbf{l}_i(t)}{\sqrt{\sigma_1^2 \|\mathbf{l}_1(t)\|^2 + \sigma_2^2 \|\mathbf{l}_2(t)\|^2}}, \qquad \xi_i := \frac{\varepsilon_{\mathbf{i}}}{\sigma_i} = \frac{Y_i - EY_i}{\sigma_i},$$

for $i = 1, 2$, where $\xi_i \in \mathcal{R}^{n_i}$ is (multivariate) standard normal for $i = 1, 2$, independent with each other, $\mathbf{l}_i(t)$, $i = 1, 2$ are defined similarly in (5.2).

$Z(t)$ can be expressed in terms of $u_1(t)$ and $u_2(t)$ (again, with assumption that $\hat{f}_i(t), i = 1, 2$ are unbiased, as we did in the homoscedastic case):

$$Z(t) = \langle u_1(t), \xi_1 \rangle - \langle u_2(t), \xi_2 \rangle.$$

The correlation function $\rho(t, t')$ of this random field $Z(t)$ can be calculated as

$$\rho(t, t') := corr(Z(t), Z(t')) = \langle u_1(t), u_1(t') \rangle + \langle u_2(t), u_2(t') \rangle \tag{5.14}$$

Each individual σ_i^2 for $i = 1, 2$ can be estimated separately as:

$$\hat{\sigma}_i^2 = \frac{\hat{\varepsilon}_i' \hat{\varepsilon}_i}{tr((I - L_i)(I - L_i)')} \tag{5.15}$$

where L_i is the matrix in the estimation equation $\hat{Y}_i = L_i Y_i$, $\hat{\varepsilon}_i = \mathbf{Y}_i - \hat{\mathbf{Y}}_i = (I - L_i)\mathbf{Y}_i$. Such estimates satisfy the property that $\hat{\sigma}_i^2/\sigma_i^2 \approx \chi_{\nu_i}^2/\nu_i$ for $i = 1, 2$, as we have discussed before.

To test $H_0 : f_1(t) = f_2(t)$ for all $t \Leftrightarrow H_1 : f_1(t) \neq f_2(t)$ for at least one $t \in \mathcal{T}$ at pre-specified level α, consider the test statistic

$$T = max_{t \in \mathcal{T}} \|Z(t)\|. \tag{5.16}$$

Depending on the realized $T = t_0$, we will reject or accept the null hypothesis if t_0 is large or small. More specifically, we need to find the probability $Pr(T > t_0)$. If this is larger than α, we will accept the null and declare that there is no difference between curves $f_1(t)$ and $f_2(t)$.

The probability $Pr(T > t_0)$ will be estimated by a generalization and modification of the theorem in Sun and Loader [22], p1330. Modifications are made here to estimate both the common degree of freedom ν and the variances when they are unknown.

Theorem 5.2 (Tail Probability Estimation for Heteroscedastic Case) *When the variances are not equal, the conclusions in Theorem 5.1 are still valid with estimated ν in (4.3) being replaced by*

$$\nu = \frac{n^2 \nu_1 \nu_2}{n_2^2 \nu_1 + n_1^2 \nu_2}.$$

That is:

$$Pr(T > t_0) \approx \frac{\kappa_0}{\pi} \exp(-\frac{t_0^2}{2}) + E(1 - \Phi(t_0)) \quad as\, t_0 \to \infty, \tag{5.17}$$

for known variances, and

$$Pr(T > t_0) \approx \frac{\kappa_0}{\pi} (1 + \frac{t_0^2}{\nu})^{-\nu/2} + \frac{E}{2} P(|t_\nu| > t_0), \tag{5.18}$$

for unknown variances, with σ_i^2 being estimated by formula (5.15).

Proof See Appendix D.

Remarks: when $n_1 = n_2$ and $\nu_1 = \nu_2$, it can be readily shown that the estimate of ν in Theorem 5.2 is reduced to the estimate of ν in (5.4) for the homoscedastic case.

Computation of κ_0: κ_0 in Theorem 5.2 will be estimated using formula (5.12) in Theorem 5.1. The assumption of whether we have equal or unequal variances will not affect the calculation of this quantity; see the definition of $u_i(t)$ of $Z(t)$, and [13, 22].

6 Boundary Correction and Automatic Bandwidth Selection

In this section, we will propose a few options to improve the curve fittings, namely with methods such as boundary corrections and automatic bandwidth selections proposed in the literature.

For kernel regression, the bias in the boundary area is automatically corrected in most cases [4] when the degree of the polynomial is properly selected. However,since our method fixed the degree of polynomial to be 1 (so it is locally linear fitting), our algorithm may need to address the boundary bias issue. We implemented a simple idea to correct the boundary biases as was discussed in [6].

Another issue often mentioned in the literature is the automatic bandwidth selection problem. As numerous publications have pointed out (see [4] for an overview),

bandwidth selection can be important in deciding the goodness of fit. In the implementation of our algorithm, we have offered three different methods to specify the bandwidth.

The first method is to specify a fixed bandwidth directly for all data points in the domain. The second method is to specify the percent of neighboring data points to be use for each point in the domain. We call this percent of neighboring data points in the later method as smoothing α. This method is more adapted than the fixed bandwidth method. For example, if $\alpha = 0.5$, we are using a bandwidth spanning a window that have 50% of all data points falling in the window. Thus for different points, the window sizes may change. Smoothing $\alpha = 0.5$ has been set as the default value for the parameter. Our simulation shows that this selection achieves the best fitting for most situations. When both bandwidth and smoothing alpha are specified, the bandwidth will be used, and the smoothing alpha will be discarded.

The third method is termed *optimal* one. When choosing this option, the algorithm will make a series curve fittings using a sequence of α values. For each α, the algorithm will calculate the generalized cross validation (GCV) statistic. The best α value will be selected to be the 'optimal' one if it generates the minimal GCV value. This method has been used in the locfit package, see [15].

The performances of different combinations of these parameters will be illustrated below in Fig. 4. Panels a and b were generated using the same parameters except that

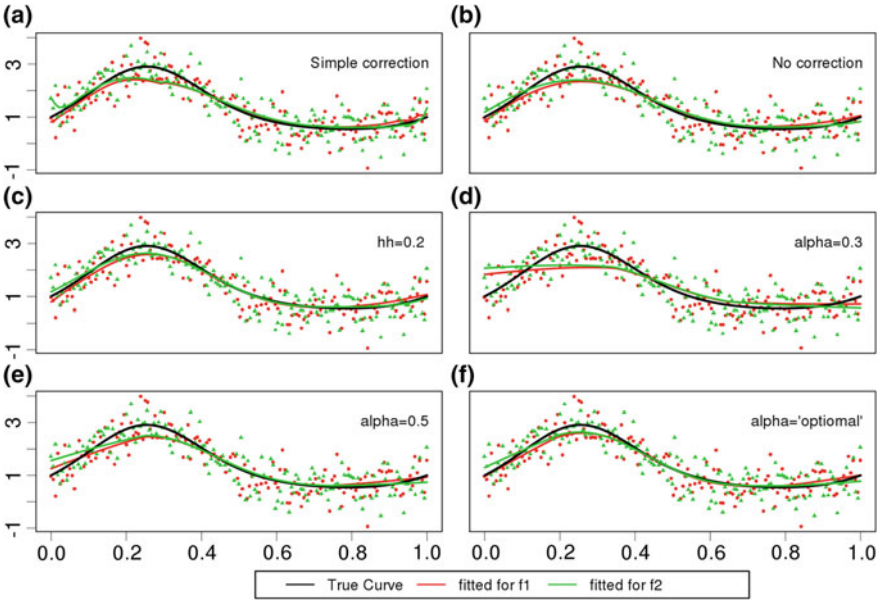


Fig. 4 Performance comparisons using different parameters in the model. True curves are assumed to be $f_1(x) = f_2(x) = x * (1 - x) + sin(2\pi x)$. **a** Simple correction was used. **b** No correction was used. **c** fixed bandwidth= 0.2. **d–f** Different smoothing alphas were used to fit the curves

A used a simple correction method to the boundary biases, while B used none. Panels c–f compare the effect using different smoothing parameters, which include a fixed bandwidth ($bw = 0.2$, c) and smoothing $\alpha = 0.3$(d), 0.5(e) and optimal(f). The simple correction specification seems to generate an over-corrected curve in the boundary area in our example (See the upward pattern in a), while it performs well with the non boundary correction specification. When the smoothing $\alpha = 0.3$, it is apparently under-fitted the data (see panel d), while the default $\alpha(= 0.5)$ greatly improve the fitting. The optimal smoothing α method improved the fitting again. However, optimal bandwidth selection based on GCV is computationally intensive, often generates a curve or curves that is over-fitted.

7 Simulations

In this section, we will show the overall performance of our proposed procedure under different assumptions based on simulated datasets.

First, we assume that the random errors are homoscedastic. We then assign values to n_1 and n_2, the sample sizes for the two groups of data sets that define the two curves. The $t_{i,j}$'s for $i = 1, 2$, $j = 1, \ldots n_i$ are chosen to be equally spaced between 0 and 1, so our $\mathcal{T} = [0, 1]$. This should not cause any loss of generality, since otherwise,

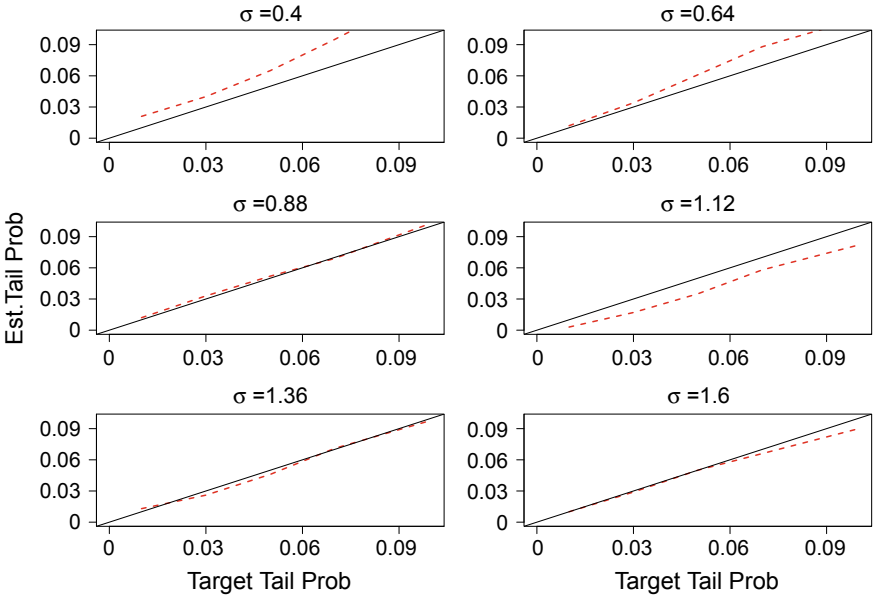


Fig. 5 Simulation results for test $f_1(t) = f_2(t), t \in \mathcal{T} = [0, 1]$: homoscedastic variances were assumed. 10000 repetitions were used

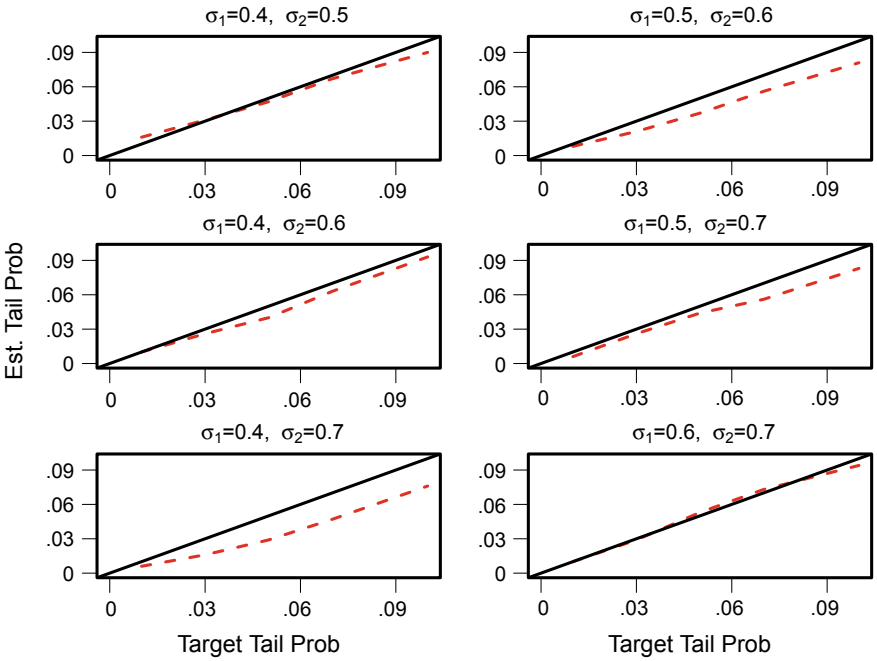


Fig. 6 Simulation results for test: $f_1(t) = f_2(t)$, for $t \in \mathcal{T} = [0, 1]$: homoscedastic variance assumption was dropped. 20000 repetitions were used. Heteroscedastic variances were used with different σ_1^2 and σ_2^2

we can scale the domain to force it to be between 0 and 1. Set $f_1(t) = f_2(t) = x * (1 - x) + sin(2\pi x)$ so that H_0 is true. Generate n_1 and n_2 $i.i.d.$ Gaussian $N(0, \sigma^2)$ random samples $\varepsilon_{i,j}$ for $i = 1, 2,\ j = 1, \ldots, n_i$, where σ is the common standard deviations of the two curves. The Y_i is obtained by adding $f_i(t)$ and ε_i together. The actual used σ are 0.4, ..., 1.6 in our simulation to produce Fig. 5.

Now partition $\mathcal{T} = [0, 1]$ into n equally spaced subintervals with the $n + 1$ end points $t_i = i/n$ for $i = 0, 1, \ldots, n$, where n is chosen by user, and is independent of both n_1 and n_2. These $t's$ are used to represent the continuous curves of f in computer. The default smoothing $\alpha = 0.5$ was used to fit the two curves to obtain the estimated function values $\hat{f}_1(t_i)$ and $\hat{f}_2(t_i)$, for $i = 1, \ldots n$. κ_0 is computed based on formula (5.12) and the realized test statistic t_0 defined in (5.16) is computed, together with the p-value based on the right-hand-side of formula (5.10). This p-value is compared with a sequence of pre-selected levels (e.g., 0.005, 0.01, 0.02, 0.05, 0.1). For each of the choice of σ^2, the proportions that H_0 is falsely rejected at each level among 10,000 iterations are calculated. The result is plotted in Fig. 5. Similar simulation results were generated in Fig. 6 for the heteroscedastic case.

Simulation results in Figs. 5 and 6 show that our approximation formulas (5.1) and (5.2) are correct and accurate, with deviations from the true target tail probabilities (the solid black line) in our study being usually less than 1%.

8 Comparison with Other Test Procedures

In this section, we will compare the performance of our curvetest procedure with that of other test procedures from the literature.

We selected two test procedures from literature for making comparisons. One is the test procedure called *hanova*, which is based on Adaptive Neyman Statistic from [7] using Fourier transformation. Another is the ordinary $ANOVA$ test for testing group differences in linear regression model (i.e., test g in $y \sim x * g$).

To generate the data, we first set a few parameters, including the standard deviations $\sigma_i, i = 1, 2$ for the errors of curves, the true curve functions $f_1(t)$ and $f_2(t)$,

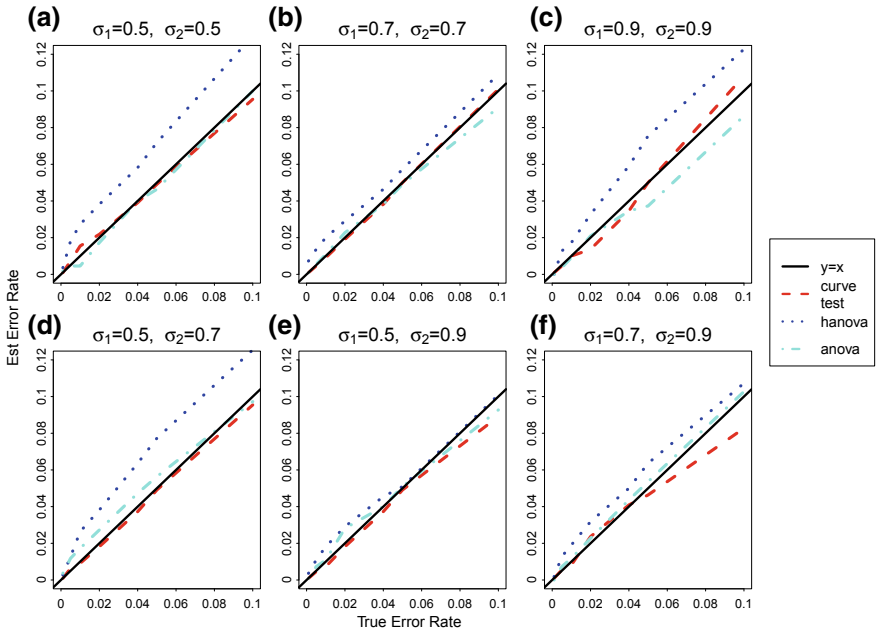


Fig. 7 Comparison of different tail probability calculations for *curvetest* (in red), *anova* (in blue) and or *hanova* (in cyan) for test: $f_1(t) = f_2(t)$. Homoscedastic (**a**–**c**) or heteroscedastic (**d**–**f**) variances were assumed here. 1000 repetitions were used to find the tail probably at different levels. σ_1 and σ_2 are the standard deviations used to generate the random errors for curves in group 1 and 2 respectively with the true curve function being set to be $f_1(t) = f_2(t) = t^{1-t} * (t - 1)$ for $t \in [0, 1]$. There are $N_1 = 45$ and $N_2 = 40$ repeated measurements defining the curves 1 and 2 respectively for each generated data set

a sequence of time points from 0 to 1 of length N_t (the common number of t-values for both curves, as it is required by *hanova* and $ANOVA$ procedures), and N_1 and N_2 for the number of repeated measurements at each time points t. At any time t, there will have N_i repeated measurements for $i = 1, 2$. As a result, there will be a total number of $N_t * N_i$ data points as our observed data for the corresponding curve $i, i = 1, 2$. The data points representing the curve i was $f_i(t)$ plus $N(0, \sigma_i^2)$ for $i = 1, 2$.

This data generating schema with repeated measurements at each time points is necessary, as *hanova* and $ANOVA$ procedures will fail if we have only one points at any individual time point. *hanova* also requires to have a relatively large number of data points at each time to have a reasonable approximation result.

Figure 7 displays the simulation results for comparison with the other two test procedures, for both homoscedastic (a–c) and heteroscedastic (d–f) cases. For either case, *hanova* gives slightly larger tail probability estimation than the targeted values. Both $ANOVA$ and curvetest in most cases give the tail estimation close to the targeted values. For homoscedastic error cases, the curvetest procedure gives tail probabilities that are most close to the targeted tail probabilities when the underlining homoscedastic errors are large. For heteroscedastic case, $ANOVA$ gives slightly better estimates than curvetest, especially when the differences between the two standard deviations are large.

9 Test Results on Teeth Lead Data Set

In this section, we will display the test results on the motivating data set from running our test procedure for illustration purpose.

To do this, the proposed test procedure was applied to the Teeth data set based on our R routine $curvetest$ for illustration (see Appendix E), with kernel function

$$W(t) = (1 - |t|^3)^3 I_{[-1,1]}(t),$$

where I denotes the indicator function and $bw = 14$ for both groups. $bw = 14$ was chosen for the common window width was on the scatter plot with smoothing curves (Fig. 1b). The p-values for the test of null hypothesis are 0.213 and 0.210 respectively if equal or unequal variances are assumed. The scatter plot with smoothing curves is displayed in Fig. 1b.

10 Discussions

We have described and developed a general test procedure to test if two curves measured with either homoscedastic or heteroscedastic errors are statistical equal, with mild requirements on the fitted curves. This test procedure, as we discussed

earlier, will have a wide range of applications in diversified areas. Ideally if we want to compare the longitudinal trends of some quantity for two or more groups, we can use longitudinal data analysis or the curve testing procedure developed here. Furthermore, the curves can be functions defined over time (time related) or over locations. For example, researchers often ask whether the growth curves (weights or tumor volumes over time) between two groups of animals are significantly different. The test developed here tests if the underlying continuous (growth) curves are actually equal or not for each value within the study interval, though the curves are measured at only finite discrete points with errors being either homoscedastic or heteroscedastic among the two curves. This property makes our test procedure unique. Our procedure directly tests whether the two curves are equal or not on *all* values in the definition interval by estimating the tail probability when null is true, while ANOVA or its variations have a different hypothesis. Our test does not require repeated values (measurements) at each times/locations points for each of the groups, as other test procedures do. Therefore, our curve test procedure will have a wider applications than for example, *hanova*.

Testing equality of curves is closely related to building a simultaneous confidence band (SCB) around a curve. To test if two curves $f_1(t)$ and $f_2(t)$ measured with errors are statistically equal, we could alternatively and equivalently build a proper SCB around the difference curve $f_1(t) - f_2(t)$. If part of line 0 in the test region is outside of this SCB, we will claim that the two curves are statistically different and vice versa. However, subtle technique differences between SCB and our curve test exist. For example, curve test allows the two curves to have different measurement standard errors, that is not true when using SCB approach. For more detail, see papers of Sun and Loader [22], Sun [24], Naiman [16].

The choice of window width can have a relatively large effect on the test from the algorithm, as it does to the smoothing algorithms. Therefore, we have offered several methods for users to choose from for selection of the bandwidth, including an automatic method. However, the choice of kernel function seems to affect little about the test result. This agrees with our previous experience in the smoothing study, see, for example, Cleveland and Devin [5].

In summary, we have developed a formal test procedure to test if two continuous curves measured with either homoscedastic or heteroscedastic errors are statistically equal or not, and developed an easy to use R-package for applications.

Appendix

Appendix A. Proof of Lemma 4.1

Let $X_1 = \sum_{i=1}^{n_1} W_i^2$, where $W_i \sim_{iid} N(0,1)$. Since $EX_1 = n_1, Var(X_1) = 2n_1$, we can approximate X_1 by $X_1 \stackrel{d}{=} n_1 + \sqrt{2n_1}Z_1 + o(\sqrt{n_1})$, X_2 by $X_2 \stackrel{d}{=} n_2 + \sqrt{2n_2}Z_2 + o(\sqrt{n_2})$, where Z_1 and Z_2 are independent standard normal random

variables and notation $\stackrel{d}{=}$ denotes equality in distribution. Therefore,

$$\begin{aligned} G_i =&(\frac{X_i}{X_1+X_2})/(\frac{n_i}{n_1+n_2}) \\ \stackrel{d}{=}&\frac{n_i+\sqrt{2n_i}Z_i+o(\sqrt{n_i})}{n_1+n_2+\sqrt{2n_1}Z_1+\sqrt{2n_2}Z_2+o(\sqrt{n_2})+o(\sqrt{n_1})}/(\frac{n_i}{n_1+n_2}) \\ =&\frac{1+\frac{\sqrt{2n_i}Z_i+o(\sqrt{n_i})}{n_i}}{1+\frac{\sqrt{2n_1}Z_1+\sqrt{2n_2}Z_2+o(\sqrt{n_2})+o(\sqrt{n_1})}{n_1+n_2}} \stackrel{d}{\sim}_{approx} 1, \end{aligned}$$

as $n_i \to \infty$ for $i = 1, 2$. QED

Appendix B. Proof of Lemma 4.2

Clearly by $S_i/\nu_i \sim \chi^2_{\nu_i}$, we have $ES_i = 1, \quad Var(S_i) = 2/\nu_i$, and there exists a sequence of *i.i.d.* standard normal random variables $W_{i,k}$, for $k = 1, 2, \ldots, \nu_i, i = 1, 2$ such that $S_i = (1/\nu_i)\sum_{k=1}^{\nu_i} W_{i,k}$ by definition. A large sample theory gives that $S_i \stackrel{d}{=} 1 + \sqrt{\frac{2}{\nu_i}}Z_i + o(\frac{1}{\sqrt{\nu_i}})$, where $Z_i \sim N(0, 1)$.

Expanding $Y = f(S_1, S_2) = (X_1 + X_2)/(X_1/S_1 + X_2/S_2)$ around $(1, 1)$, we have:

$$Y = f(S_1, S_2) = 1 + \sqrt{\frac{2}{\nu_1}}Z_1\frac{X_1}{X_1+X_2} + \sqrt{\frac{2}{\nu_2}}Z_2\frac{X_2}{X_1+X_2} + o(\frac{1}{\sqrt{\nu_1}}+\frac{1}{\sqrt{\nu_2}}).$$

It is easy to find that $EY \to 1$. Conditioning on X_1, X_2, we have

$$\begin{aligned} Var(Y) =&Var(E(Y|X_1, X_2)) + E(Var(Y|X_1, X_2)) \\ =&E(Var(Y|X_1, X_2)) + o(\frac{1}{\sqrt{\nu_1}}+\frac{1}{\sqrt{\nu_2}}) \\ \approx&E\left\{Var(\sqrt{\frac{2}{\nu_1}}Z_1\frac{X_1}{X_1+X_2} + \sqrt{\frac{2}{\nu_2}}Z_2\frac{X_2}{X_1+X_2}|X_1, X_2)\right\} \\ =&E\left\{\frac{2}{\nu_1}(\frac{X_1}{X_1+X_2})^2 + \frac{2}{\nu_2}(\frac{X_2}{X_1+X_2})^2\right\} \\ =&\frac{2}{\nu_1}E(\frac{X_1}{X_1+X_2})^2 + \frac{2}{\nu_2}E(\frac{X_2}{X_1+X_2})^2 \\ \approx&\frac{2}{\nu_1}(\frac{n_1}{n_1+n_2})^2 + \frac{2}{\nu_2}(\frac{n_2}{n_1+n_2})^2 \qquad \text{(by Lemma (4.1))} \\ =&\frac{2n_1^2\nu_2 + 2n_2^2\nu_1}{(n_1+n_2)^2\nu_1\nu_2}. \end{aligned}$$

Thus by comparing $Var(\chi^2_\nu/\nu) = 2/\nu$ with $Var(Y) = (2n_1^2\nu_2 + 2n_2^2\nu_1)/((n_1 + n_2)^2\nu_1\nu_2)$, we have estimate ν in formula (4.3). QED

Appendix C. Proof of Theorem 5.1

$Z(t)$ in formula (5.6) can be written as $Z(t) = \langle u_1(t),\ \xi_1\rangle - \langle u_2(t),\ \xi_2\rangle = \langle \mathcal{M}(t),\ \xi\rangle$, where $\mathcal{M}(t) = (u_1(t), -u_2(t))' \in \mathcal{S}^{n-1}$ and $\xi = (\xi_1, \xi_2)' = (\varepsilon_1/\sigma, \varepsilon_2/\sigma)' \in \mathcal{R}^n$ is standard multivariate normal. Conditioning on $\|\xi\|$, the probability can be written as

$$\begin{aligned} Pr(T \geq t_0) =& Pr(\sup_{t\in\mathcal{T}} |\ \langle \mathcal{M}(t),\ \xi\rangle\ | \geq t_0) \\ =& \int_{t_0}^{\infty} Pr(\sup_{t\in\mathcal{T}} \left|\left\langle \mathcal{M}(t),\ \frac{\xi}{\|\xi\|}\right\rangle\right| \geq \frac{t_0}{y}\ \mid\ \|\xi\| = y) g(y,n)dy \end{aligned} \quad (10.1)$$

where $g(y, n)$ is the probability density function (pdf) of the square root of a χ^2 random variable with n degrees of freedom. Since $U = \xi/\|\xi\| \sim uniform(\mathcal{S}^{n-1})$ is independent of $\|\xi\|$, we can drop the condition in the probability. Let $T = \{x \in S^{n-1} : \sup_{t\in\mathcal{T}} |< \mathcal{M}(t),\ x >| \geq (t_0/y)\}$, we then have tubes around curve $\mathcal{M}(t)$ and curve $-\mathcal{M}(t)$ embedded in S^{n-1}, with radius $r = \sqrt{2 - 2t_0/y}$ (See relation (4.4)). The probability inside the integral of (10.1) can be calculated by $Vol(T)/Vol(S^{n-1})$. We then plug-in the tube formula (4.5) to get result (5.9). See also [22] [Proposition 1, p. 1330]. Result (5.10) is obtained by replacing $g(y, n)$ of the pdf of $\|\xi\| = \sqrt{\varepsilon'\varepsilon/\sigma^2}$ by the pdf of $\|\hat{\xi}\| = \sqrt{\varepsilon'\varepsilon/\hat{\sigma}^2}$, where $\|\hat{\xi}\|^2/n \sim F_{n,\nu}$. For details, see the above citation. QED

Appendix D. Proof of Theorem 5.2

Suppose σ_1^2 and σ_2^2 are known. Let $\mathcal{M}(t) = (u_1(t),\ \ u_2(t))' \in \mathcal{S}^{n-1} \subseteq \mathcal{R}^n$ for $n = n_1 + n_2,\ t \in \mathcal{T}$, as before. Let $\xi = (\xi_1,\ \ \xi_2)'$ and $U = \xi/\|\xi\|$. Then U is uniformly distributed (over $\mathcal{S}^{n-1}$), independent of $\|\xi\|$. Since σ_1 and σ_2 are known, $\|\xi\|^2$ will follow a χ^2_n distribution.

Conditioning on value $\|\xi\|$, making use of the fact that ε_1 and ε_2 are independent, we have

$$\begin{aligned} Pr(T > t_0) =& Pr(\sup_{t\in\mathcal{T}} \|Z(t)\| \geq t_0) = Pr(\sup_{t\in\mathcal{T}} \langle \mathcal{M}(t), \xi\rangle \quad \geq t_0) \\ =& Pr(\sup_{t\in\mathcal{T}} \langle \mathcal{M}(t), \frac{\xi}{\|\xi\|}\rangle \geq \frac{t_0}{\|\xi\|}) \\ =& \int_{y\geq t_0} Pr(\sup_{t\in\mathcal{T}} \langle \mathcal{M}(t), U\rangle \geq \frac{t_0}{y} \mid \|\xi\| = y) f_{\|\xi\|}(y) dy \\ =& \int_{y\geq t_0} Pr(\sup_{t\in\mathcal{T}} \langle \mathcal{M}(t), U\rangle \geq \frac{t_0}{y}) f_{\|\xi\|}(y) dy, \end{aligned}$$

where $f_{\|\xi\|}(y)$ denotes the pdf of $\|\xi\|$, whose square follows a χ^2_n distribution. The last equation holds because of the fact that $\xi/\|\xi\|$ and $\|\xi\|$ are independent.

When t_0 is large, the following tube formula [cf: Lemma 4.5] can be plugged into the last equation

$$Pr(\sup_{t\in\mathcal{T}} \|\langle \mathcal{M}(t), U\rangle\| \geq c) \approx 2 * \left\{ \frac{\kappa_0}{2\pi}(1-c^2)^{n/2-1} + \frac{E}{4} Pr(\beta_{\{1/2,(n-1)/2\}} \geq c^2 \right\}$$

where $\beta_{\{1/2,(n-1)/2\}}$ denotes a Beta random variable with parameters 1/2 and (n–1)/2. The factor 2 corresponds to the two curves satisfying the probability condition: one is $\mathcal{M}(t)$, another $-\mathcal{M}(t)$. This gives formula (5.17).

Suppose we don't know both $\sigma_i^2, i = 1, 2,$ but they are estimated as in formula (5.5). Let

$$X_i = \frac{\varepsilon_i'\varepsilon_i}{\sigma_i^2}, \quad S_i = \frac{\hat{\sigma}_i^2}{\sigma_i^2}, \qquad i = 1, 2, \tag{10.2}$$

$$Y = \frac{X_1}{S_1} + \frac{X_2}{S_2},\ X = \frac{X_1 + X_2}{X_1/S_1 + X_2/S_2} = \frac{X_1 + X_2}{Y}. \tag{10.3}$$

Then the requirements of Lemma 4.2 are satisfied, hence we have $X \sim_{approx} \chi^2_\nu/\nu$, with degrees of freedom ν estimated in formula (4.3). Therefore, $Y = (X_1 + X_2)/X \to (\chi^2_n)/(\chi^2_\nu/\nu) \sim nF_{n,\nu}$.

Now let

$$\hat{\mathcal{M}}(t) = \frac{(\hat{\sigma}_1 \mathbf{l}_1(t), -\hat{\sigma}_2 \mathbf{l}_2(t))}{\sqrt{\hat{\sigma}_1^2 \|\mathbf{l}_1(t)\|^2 + \hat{\sigma}_2^2 \|\mathbf{l}_2(t)\|^2}} \in \mathcal{S}^{n-1}, \qquad U = \frac{\hat{\xi}}{\|\hat{\xi}\|},$$

$$\hat{\xi} = (\frac{\varepsilon_1}{\hat{\sigma}_1}, \frac{\varepsilon_2}{\hat{\sigma}_2}) \in \mathcal{R}^n, \qquad \hat{Z}(t) =< \hat{\mathcal{M}}(t), \hat{\xi} > .$$

Then

$$
\begin{aligned}
Pr(T > t_0) =& Pr(\sup_{t\in\mathcal{T}} \|\hat{Z}(t)\| \geq t_0) = Pr(\sup_{t\in\mathcal{T}} \langle \hat{\mathcal{M}}(t), \hat{\xi}\rangle \quad \geq t_0) \\
=& Pr(\sup_{t\in\mathcal{T}} \langle \hat{\mathcal{M}}(t), \frac{\hat{\xi}}{\|\hat{\xi}\|}\rangle \geq \frac{t_0}{\|\hat{\xi}\|}) \\
=& \int_{y\geq t_0} Pr(\sup_{t\in\mathcal{T}} \langle \hat{\mathcal{M}}(t), U\rangle \geq \frac{t_0}{y} \mid \|\hat{\xi}\| = y) f_{\|\hat{\xi}\|}(y)dy \\
=& \int_{y\geq t_0} Pr(\sup_{t\in\mathcal{T}} \langle \hat{\mathcal{M}}(t), U\rangle \geq \frac{t_0}{y}) f_{\|\hat{\xi}\|}(y)dy,
\end{aligned}
$$

where $Y =: \|\hat{\xi}\|^2 \sim_{approx} \chi^2_\nu/\nu$, as Y was defined in (4.2).

Let the σ_i^2 be known values in $\hat{\mathcal{M}}(t)$ in order to estimate κ_0, so that we can use the Tube formula (4.5) for the estimation of the probability inside the integral. After some calculation, we get result (5.18). QED

Appendix E. Software

Our R package *curvetest* is freely available at http://www.r-project.org/. This package tests the equality of curves as described in this paper. The main function curvetest has the following parameters:

formula: specified the regression formula.
data1: data.frame representing the first (discretized) curve.
data2: a data frame representing the second curve. If it is NULL, then the test is test $f(t) == 0$. data1 and data2 must have two columns with same column names, that can be retrieved by calls on the formula.
equal.var: logic value, specifies if equal.variances are assumed. Default=TRUE.
plotit: logic, asks if *curvetest* should generate the scatter plots and smoothing curves. It is useful to plot it to select the window width bw below. Default=F.
bw: Window bandwidth for both curves.
alpha: Smoothing parameter. Default=0.5.
nn: number of points used to smooth the curves. The points are equally spaced between the domains that appeared in the two data set. Default=100.
myx: x (or t) values to estimate the curves. Default= NULL. This will put n points specified by nn in the data range. If myx is non-null, papameter nn will be suppressed.
bcorrect: method to use for boundary correction. Default='simple'. Other options are: 'none'=no corrections.
Conf.level: the α value for the type I error level. Default=.05.
kernel: kernel function to choose for smoothing. Users can choose one of 'Trio', 'Gaussian', 'Uniform', 'Triweight', 'Triangle', 'Epanechnikov', 'Quartic'. See the definitions of them in Table 2.

Table 2 List of kernel functions

Kernel	K(u)
Epanechnikov	$(3/4)(1-u^2)I(\|u\| \leq 1)$
Gaussian	$1/(\sqrt{2\pi})exp(-u^2/2)$
Quartic	$(15/16)(1-u^2)^2 I(\|u\| \leq 1)$
Triangle	$(1- \| u \|) I(\|u\| \leq 1)$
Trio	$(1-\|u\|^3)^3 I(\|u\| \leq 1)$
Triweight	$(35/32)(1-u^2)^3 I(\|u\| \leq 1)$
Uniform	$I(\|u\| \leq 1)/2$

Usage:

```
          n1=150; n2=155 ##numbers of data points for the two curves.
          f1<-f2<-function(x){x*(1-x)+sin(2*pi*x)}; ##True functions.
          x1=seq(0,1, length=n1);
          x2=seq(0, 1, length=n2);
          y1=f1(x1)+rnorm(n1, 0, 0.2)
          y2=f2(x2)+rnorm(n2, 0, 0.2) ###Measured data for the
                                      ###two curves with noises.
          curvetest(y~x,data.frame(x=x1,y=y1),
            data.frame(x=x2,y=y2), alpha =  0.7,
            equal.var=TRUE,plotit=TRUE)
Output:
                ======================
                Curve Test  Procedures
                ======================
 The p-value to test H0:f1(x)=f2(x) is 1.

 With test statistics equals            1.2,
 Estimated degree of freedom is         294.

 Equal variances assumed.
 Estimated common sigma^2 is            0.0707.
 ====================
```

References

1. ATSDR: The nature and extent of lead poisoning in children in the united states: a report to congress. Technical report, Agency for Toxic Substances and Disease Registry, Atlanta: US Department of Health and Human Services, Public Health Service(1988)
2. Besse, P., Ramsay, J.O.: Principle components analysis of sampled functions. Psychometrika **51**(2), 285–311 (1986). https://doi.org/10.1007/bf02293986
3. Chakravarti, I.M., Laha, R.G., Roy, J.: Handbook of Methods of Applied Statistics, Vol. I. John Wiley and Sons (1967)
4. Clive, R.: Loader: Bandwidth selection: classical or plug-in? Ann. Statist. **27**(2), 415–438 (1999)

5. Cleveland, W., Devin, S.: Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. **83**(403), 596–610 (1988). https://doi.org/10.1080/01621459.1988.10478639
6. Dai, J., Sperlich, S.: Simple and effective boundary correction for kernel densities and regression with an application to the world income and Engel curve estimation. Comput. Stat. Data Anal. Elsevier **54**(11), 2487–2497 (2010)
7. Fan, J., Lin, S.: Test of significance when data are curves. J. Am. Stat. Assoc. **93**, 1007–1021 (1998). https://doi.org/10.1080/01621459.1998.10473763
8. Hotelling, H.: Tubes and spheres in n-spaces, and a class of statistical problems. Am. J. Math. **61**, 440–460 (1939)
9. James, G., Hastie, T.: Functional linear discriminant analysis for irregularly sampled curves. J. R. Stat. Soc. Ser. B **63**(3), 533–550 (2001). https://doi.org/10.1111/1467-9868.00297
10. James, W., Stein, C.: Estimation with quadratic loss. In: Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability Theory University of California Press, pp 361–380 (1961)
11. Johansen, S., Johnstone, I.: Hotelling's theorem on the volume of tubes: some illustrations in simultaneous inference and data analysis. Ann. Stat. **18**, 652–684 (1990)
12. Kitska, D.J.: Simultaneous inference for functional linear models. Ph.D. thesis, Case Western Reserve University (2005)
13. Knowles, M., Siegmund, D.: On hotelling's approach to testing for a nonlinear parameter in regression. Int. Stat. Rev. **57**(3), 205–220 (1989). https://doi.org/10.2307/1403794
14. Leurgans, S.E., Moyeed, R.A., Silverman, B.W.: Canonical correlation analysis when the data are curves. J. R. Stat. Soc. Ser. B **55**(3), 725–740 (1993)
15. Loader, C.: Local Regression and Likelihood. Springer, New York (1999)
16. Naiman, D.Q.: Simultaneous confidence bounds in multiple regression using predictor variable constraints. J. Am. Stat. Assoc. **82**, 214–219 (1987). https://doi.org/10.2307/2289156
17. Naiman, D.Q.: On volumes of tubular neighborhoods of spherical polyhedra and statistical inference. Ann. Stat. **18**(2), 685–716 (1990)
18. Parzen, E.: An approach to time series analysis. Ann. Math. Stat. **32**(4), 951–989 (1961)
19. Ramsay, J., Daizell, C.: Some tools for functional data analysis. J. R. Stat. Soc. Ser. B **53**(3), 539–572 (1991). https://doi.org/10.2307/2345586
20. Robbins, N., Zhang, Z., Sun, J., Ketterer, M., Lalumandier, J., Shulze, R.: Childhood lead exposure and uptake in teeth in the Cleveland area during the era of leaded gasoline. Sci. Total Environ. **408**(19), 4118–27 (2010). https://doi.org/10.1016/j.scitotenv.2010.04.060
21. Shapiro, S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**(3 & 4), 591–611 (1965)
22. Sun, J., Loader, C.: Simultaneous confidence bands for linear regression and smoothing. Ann. Stat. **22**, 1328–1345 (1994). https://doi.org/10.1214/aos/1176325631
23. Sun, J.: Tail probabilities of the maxima of gaussian random fields. Ann. Probab. **21**(1), 34–71 (1993). https://doi.org/10.1214/aop/1176989393
24. Sun, J.: Multiple comparisons for a large number of parameters. Biom. J. **43**, 627–643 (2001). https://doi.org/10.1002/1521-4036(200109)43:53.3.CO;2-6
25. Weyl, H.: On the volume of tubes. Am. J. Math. **61**(2), 461–472 (1939)
26. Wang, J.-L., Chiou, J.-M., Mller1, H.-G.: Functional data analysis. Ann. Rev. Stat. Appl. **3**, 257–295 (2016). https://doi.org/10.1146/annurev-statistics-041715-033624
27. Xintaras, C.: Impact of lead-contaminated soil on public health (1992). http://www.cdc.gov/search.do. (Technical report, Agency for Toxic Substances and Disease Registry)

Quality Control Metrics for Extraction-Free Targeted RNA-Seq Under a Compositional Framework

Dominic LaRoche, Dean Billheimer, Kurt Michels and Bonnie LaFleur

Abstract The rapid rise in the use of RNA sequencing technology (RNA-seq) for scientific discovery has led to its consideration as a clinical diagnostic tool. However, as a new technology the analytical accuracy and reproducibility of RNA-seq must be established before it can realize its full clinical utility (SEQC/MAQC-III Consortium, 2014; VanKeuren-Jensen et al. 2014). We respond to the need for reliable diagnostics, quality control metrics and improved reproducibility of RNA-seq data by recognizing and capitalizing on the relative frequency nature of RNA-Seq data. Problems with sample quality, library preparation, or sequencing may result in a low number of reads allocated to a given sample within a sequencing run. We propose a method, based on outlier detection of Centered Log-Ratio (CLR) transformed counts, for objectively identifying problematic samples based on the total number of reads allocated to the sample. Normalization and standardization methods for RNA-Seq generally assume that the total number of reads assigned to a sample does not affect the observed relative frequencies of probes within an assay. This assumpion, known as Compositional Invariance, is an important property for RNA-Seq data which enables the comparison of samples with differing read depths. Violations of the invariance property can lead to spurious differential expression results, even after normalization. We develop a diagnostic method to identify violations of the Compositional Invariance property. Batch effects arising from differing laboratory conditions or operator differences have been identified as a problem in high-throughput measurement systems (Leek et al. in Genome Biol 15, R29 [14]; Chen et al. in PLoS One 6 [10]). Batch effects are typically identified with a hierarchical clustering (HC) method or principal components analysis (PCA). For both methods, the multivariate distance between the samples is visualized, either in a biplot for PCA or a dendrogram for HC, to check for the existence of clusters of samples related to batch. We show that CLR transformed RNA-Seq data is appropriate for evaluation in a PCA biplot

D. LaRoche (✉) · K. Michels · B. LaFleur
HTG Molecular Diagnostics, Inc., Tucson, AZ, USA
e-mail: dlaroche@htgmolecular.com

D. Billheimer
Department of Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ, USA

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_21

and improves batch effect detection over current methods. As RNA-Seq makes the transition from the research laboratory to the clinic there is a need for robust quality control metrics. The realization that RNA-Seq data are compositional opens the door to the existing body of theory and methods developed by Aitchison (The statistical analysis of compositional data, Chapman & Hall Ltd., 1986) and others. We show that the properties of compositional data can be leveraged to develop new metrics and improve existing methods.

Keywords RNA-Seq · Next generation sequencing · Composition · Quality control · Relative abundance · Normalization

1 Introduction

We develop quality control diagnostics for targeted RNA-Seq using the theory of compositional data. Targeted sequencing allows researchers to efficiently measure transcripts of interest for a particular disease by focusing sequencing efforts on a select subset of transcript targets. Targeted sequencing offers several benefits over traditional whole-transcriptome RNA-Seq for clinical use including the elimination of amplification bias, reduced sequencing cost, and a simplified bioinformatics workflow. Moreover, extraction-free targeted sequencing technologies, such as HTG EdgeSeq, permit the use of very small sample volumes. However, extraction free technologies create the need for post-sequencing quality control metrics since poor quality samples, which would likely be removed after unsuccessful RNA extraction in extraction-based technologies, can still be sequenced. The post-sequencing methods described here should be easily extensible to traditional extraction-based RNA-Seq because targeted and traditional RNA-Seq data share many of the same properties.

Relative frequency measures are characterized as a vector of proportions of some whole. These proportions are necessarily positive and sum to a constant which is determined by the measurement system and not the measurand. Targeted and whole transcriptome RNA-Seq measurements from NGS-based instruments provide only relative frequencies of the measured transcripts. The measurement technology, along with sample preparation, preclude the measurement of absolute abundance. The total number of reads in a sequencing run for high-throughput RNA-Seq instruments is determined by the maximum number of available reads and not the absolute number of reads in a sample. For example, the Illumina Mi-Seq is limited to 25 million reads in a sequencing run while the Roche 454 GS Junior $^{(TM)}$, with longer read lengths, claims approximately 100,000 reads per run for shotgun sequencing. These reads are distributed across all of the samples included in a sequencing run and, therefore, impose a total sum constraint on the data. This constraint cascades down to each probe or tag within a sample which is, in turn, constrained by the total number of reads allocated to the sample thereby creating a natural hierarchical structure to RNA-Seq data.

Previous authors have identified the relative abundance nature of RNA-Seq data [6, 13, 16, 22, 23]. For example, [23] consider counts of RNA tags as relative abundances in their development of a model for estimating differential gene expression implemented in the Bioconductor package edgeR. Similarly, Robinson and Oshlack [22] explicitly acknowledge the mapped-read constraint when developing their widely used Trimmed-Mean of M-values (TMM) normalization method for RNA-Seq data. Finally, the commonly used $\log_2$ Counts per Million (CPM) re-scaling transformation proposed by Law et al. [13] divides each sequence count by the total number of reads allocated to the sample thereby transforming the data for each sample into a vector of proportions.

The positivity and summation constraint complicate the analysis of relative frequency data. As early as 1896 Pearson [21] identified the spurious correlation problem associated with compositions. John Aitchison observed that relative frequency data is compositional and developed a methodology based on the geometric constraints of compositions [1]. Recent authors have argued that ignoring the sum constraint can lead to unexpected results and erroneous inference [15]. Despite the evidence that RNA-Seq data are compositional in nature, few researchers have extended the broad set of compositional data analysis theory and operations for use in RNA-Seq analysis problems.

We provide a brief background on compositional methods. We then extend existing compositional data methodology to develop two quality control metrics and improve batch effect detection for RNA-Seq data.

2 Methods

2.1 Compositional Data

Compositional data is defined as any data in which all elements are non-negative and sum to a fixed constant [1]. For RNA-seq data, the total sum constraint is imposed by the limited number of available reads in each sequencing run. Since this total differs between sequencing platforms we will refer to the total number of available reads as $\mathbb{T}$. These reads are distributed among the D samples in a sequencing run such that:

$$\sum_{i=1}^{D} t_i = \mathbb{T} \tag{1}$$

where t_i represents the total reads for sample i. Because of the total sum constraint, the vector $\mathbf{t}$ is completely determined by $D - 1$ elements since the D^{th} element of $\mathbf{t}$ can be determined from the other $d = D - 1$ elements and the total $\mathbb{T}$:

$$t_D = \mathbb{T} - \sum_{i=1}^{d} \mathbf{t_i} \tag{2}$$

In Eq. 2, any of the elements can be chosen for t_D with the remaining elements labeled 1, ..., d in any order [1]. Similarly, the total reads for each sample (t_i) are distributed among the P transcript targets in the assay such that $\sum_{j=1}^{P} p_{ij} = t_i$, where p_{ij} is the number of reads allocated to target j in sample i. We highlight the hierarchical structure of RNA-Seq data as it leads to useful properties when developing quality control metrics.

From Eqs. 1 and 2 it is clear that the total reads allocated to each of the D samples represent a $D - 1 = d$ dimensional simplex ($\mathcal{S}^d$). This leads to problems when using methods developed for standard Euclidean sample spaces such as interpreting the traditional $D \times D$ covariance structure or measuring the distance between vectors. In particular, it is clear that for a D-part composition $\mathbf{x}$, $\text{cov}(x_1, x_1 + \cdots + x_D) = 0$ since $x_1 + \cdots + x_D$ is a constant. Moreover, the sum constraint induces negativity in the covariance matrix,

$$\text{cov}(x_1, x_2) + \cdots + \text{cov}(x_1, x_D) = -\text{var}(x_1). \tag{3}$$

Equation 3 shows that at least one element of each row of the covariance matrix must be negative. Aitchison refers to this as the "negative bias difficulty" (although 'bias' is not used in the traditional sense; [1], p. 53). The structurally induced negative values create problems for the interpretation of the covariance matrix. Similarly, the use of naive distance metrics in the simplex may not be interpretable as in Euclidean space. Because of these difficulties, standard statistical methodology is not always appropriate [1] and can produce misleading results [16].

To overcome these obstacles, Aitchison [5] proposed working in ratios of components. We focus on the Centered Log-Ratio (CLR) which treats the parts of the composition symmetrically and provides an informative covariance structure. The CLR transformation is defined for a D-part composition $\mathbf{x}$ as:

$$y_i = \text{CLR}(x_i) = log\left(\frac{x_i}{g(\mathbf{x})}\right), \tag{4}$$

where $g(\mathbf{x})$ is the geometric mean of $\mathbf{x}$. The $D \times D$ covariance matrix is then defined as:

$$\Gamma = \left[\text{cov}\left(y_i, y_j\right) : i,\ j = 1, ..., D\right] \tag{5}$$

The CLR transformation is similar to the familiar Counts per Million (CPM) transformation [13] defined as, $log_2\left(\frac{r_{gi}+0.5}{t_i+1} \times 10^6\right)$, where r_{gi} is the number of sequence reads for each probe (g) and sample (i), (scaled to avoid zero counts), adjusted for the number of mapped reads (library count) for each sample t_i (scaled by a constant 1 to ensure the proportional read to library size ratio is greater than zero). The primary difference between the CLR and log(CPM) transformations is in the use of the geometric mean in the denominator of the CLR transformation. The use of the geometric mean results in subtracting the mean of the log transformed values from

each log-transformed element thereby centering the vector of log-ratio transformed read counts. The difference appears minor but has important implications for the application of several common statistical methods.

Although the CLR transformation preserves the original dimension of the data, and gives equal treatment to every element of **x**, the resulting covariance matrix, Γ, is singular. Therefore, care should be taken when using general multivariate methods on CLR transformed data. Aitchison [1] proposed an alternative transformation, the additive log-ratio (ALR), which does not treat the components symmetrically but results in a non-singular covariance matrix. The ALR transformation is defined as,

$$y_i = \mathrm{ALR}(x_i) = log\left(\frac{x_i}{x_D}\right), \tag{6}$$

where x_D, the D^{th} component of x, can be any component.

As noted above, the compositional geometry must be accounted for when measuring the distance between two compositional vectors or finding the center of a group of compositions [4]. Aitchison [2] outlined several properties for any compositional difference metric which must be met: scale invariance, permutation invariance, perturbation invariance (similar to translation invariance for Euclidean distance), and subcompositional dominance (similar to subspace dominance of Euclidean distance). The scale invariance requirement is ignorable if the difference metric is applied to data on the same scale (which is generally not satisfied in raw RNA-seq data due to differences in read depth). The permutation invariance is generally satisfied by existing methods such as Euclidean distance [19]. However, the perturbation invariance and subcompositional dominance are not generally satisfied [19].

Aitchison [1, 2] suggests using the sum of squares of all log-ratio differences. Billheimer et al. [8] use the geometry of compositions to define a norm which, along with the perturbation operator defined by Aitchison [1], allow the interpretation of differences in compositions. Martin-Fernandez et al. [19] showed that applying either Euclidean distance or Mahalanobis distance metric to CLR transformed data satisfies all the requirements of a compositional distance metric. Euclidean distance on CLR transformed compositions is referred to as Aitchison distance:

$$d_A(x_i, x_j) = \left[\sum_{k=1}^{D}\left(log\left(\frac{x_{ik}}{g(x_i)}\right) - log\left(\frac{x_{jk}}{g(x_j)}\right)\right)^2\right]^{\frac{1}{2}} \tag{7}$$

or

$$d_A(x_i, x_j) = \left[\sum_{k=1}^{D}\left(clr(x_{ik}) - clr(x_{jk})\right)^2\right]^{\frac{1}{2}}. \tag{8}$$

To avoid numerical difficulties arising from sequence targets with 0 reads, Martin-Fernandez et al. [18] suggest an additive-multiplicative hybrid transformation. If zeros are present in the data We recommend using the Martin-Fernandez

transformation with a threshold value of $\delta = \frac{0.55}{\text{Total Reads}}$ to account for differences in sequencing depth. The CLR transformation is then applied to the Martin-Fernandez transformed data which contains no zeros.

Up to this point we have referred to the total reads available per sequencing run, $\mathbb{T}$. However, it is more typical to work with the aligned reads in practice. The total aligned reads, T, is always a fraction of the total reads available for a sequencing run, $\mathbb{T}$. The fraction of the total reads aligned can be affected by multiple factors, including the choice of alignment algorithm, which we do not address here. We assume that T imposes the same constraints on the data as outlined above for $\mathbb{T}$ and will refer exclusively to T hereafter.

3 Fractional Allocation of Aligned Reads to Samples

Problems with sample quality, library preparation, or sequencing may result in a low number of reads allocated to a given sample within a sequencing run. The Percent Pass Filter (% PF) metric provided on Illumina sequencers provides a measure that can identify problems with sequencing that result in a low number of reads allocated to a sample. However, % PF will not necessarily catch problems associated with poor sample quality or problems with sample pre-processing since these processes may affect cluster generation, and not just cluster quality. This is particularly important for extraction-free RNA-Seq technologies, such as the HTG EdgeSeq$^{(tm)}$, which allow for the use of smaller input amounts but lack the intermediate steps for checking sample quality. There is currently no objective way to evaluate sample quality based on the total number of reads attributed to a sample. We propose a method for objectively identifying problematic samples based on the total number of reads allocated to the sample.

For most experimental designs we expect the number of reads allocated to each sample in a sequencing run to arise from the same general data generating mechanism, namely the chemistry of the NGS-based measurement system, regardless of experimental condition. The objective is then to determine which samples arise from a different mechanism. Outlier detection is well suited for discovering observations that deviate so much from other observations that they are likely to have arisen from a different mechanism [12]. We base our method off Tukey's box-plots [27], which is a commonly used and robust method for detecting outliers [7].

We expect the total number of reads allocated to each sample, t_i, to be equivalent notwithstanding random variation. For a given sequencing run with D samples we define the vector of total reads allocated to each sample as $\mathbf{t}$. Since the D dimensional vector $\mathbf{t}$ is a composition we have $\mathbf{t} \in \mathcal{S}^{D-1}$, the $D-1$-dimensional simplex. As noted above, traditional statistical methods may not be appropriate for data in the simplex. Therefore, we map $\mathbf{t} \in \mathcal{S}^{D-1} \rightarrow \mathbf{x} = CLR(\mathbf{t}) \in \mathcal{R}^D$ using the Centered Log Ratio transformation Eq. 4. We then apply Tukey's method for detecting outliers to $\mathbf{x}$, which simply identifies those observations which lie outside 1.5 times the interquartile range.

Definition 1 x_i is a quality control sample failure if $x_i <$ lower-quartile$-1.5\times$ IQR *or* $x_i >$ upper-quartile$+1.5\times$ IQR, where IQR is the interquartile range of **x**.

We demonstrate the utility of our sample quality control measure using two sets of targeted RNA-Seq data: (1) 120 mRNA technical replicate universal-RNA samples prepared with the HTG EdgeSeq Immuno-Oncology assay and sequenced in 5 different equally sized runs, and (2) 105 miRNA technical replicate samples of human plasma, FFPE tissue, and Brain RNA prepared with the HTG EdgeSeq Whole Transcriptome miRNA assay. These two data sets differ in the both the type of RNA (mRNA versus miRNA) and the number of sequence targets in each assay (558 versus 2,280 targets, for the mRNA and miRNA assays respectively). All samples were prepared for sequencing using the HTG EdgeSeq Processor and sequenced with an Illumina Mi-Seq sequencer.

We compare the utility of our method to evaluation of the un-transformed total counts. Figure 1 shows a boxplot and heat-map of the total number of reads allocated to each sample for each of 5 sequencing runs. Figure 2 shows the same data after CLR transformation. After transformation the poor samples become much more visually

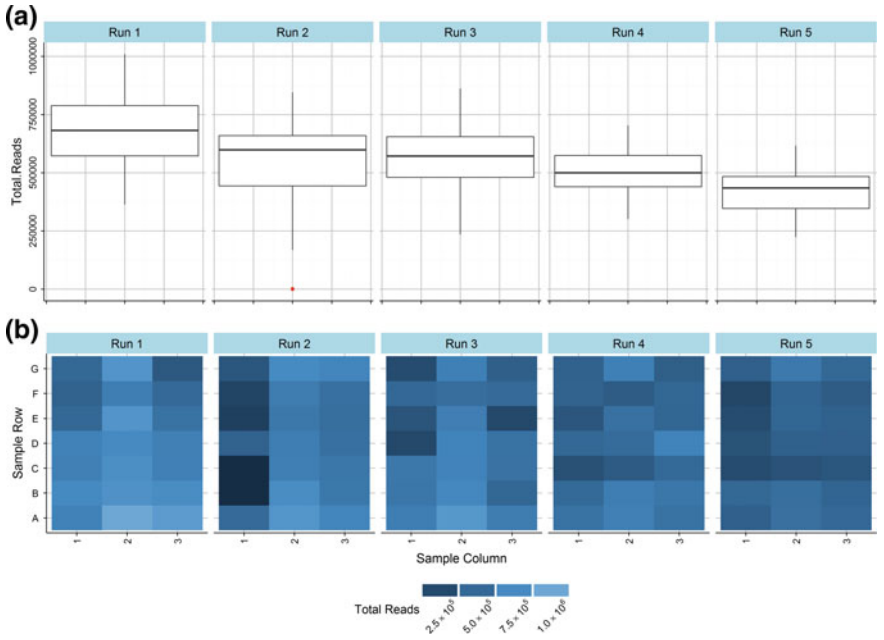


Fig. 1 **a** Distributions of total reads allocated to each sample in 5 runs on an Illumina Mi-Seq sequencer. Only 1 sample is identified as a problematic sample. **b** Heat-maps showing the relative totals for each sample within each run. The darker heat-maps for runs 4 and 5 reflect the generally lower number of total reads in those sequencing runs as compared to runs 1 and 2. This is caused by normal variation in the number of reads available in a sequencing run

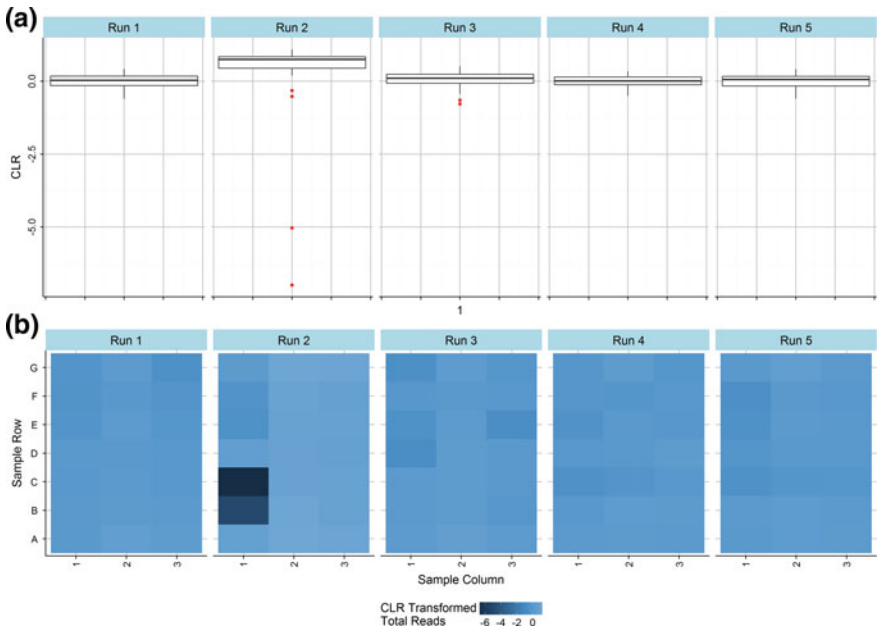


Fig. 2 **a** Distributions of CLR transformed total reads allocated to each sample in 5 runs on an Illumina Mi-Seq sequencer. After CLR transformation, 6 samples are identified as a problematic. **b** Heat-maps showing the relative CLR transformed totals for each sample within each run

evident in the heat maps. Additionally, the ability to detect outlying values increases and the number of poor samples detected increases from 1 to 6.

4 Testing for Compositional Invariance

Normalization and standardization methods for RNA-Seq generally assume that the total number of reads assigned to a sample does not affect the observed relative frequencies of probes within an assay. For example, implicit in the CPM transformation is the idea that if you re-scale the counts (by dividing by the total for each sample) then the resulting counts are comparable and any differences are due to underlying differences in expression. Other methods which apply a scaling factor to each sample, such as Trimmed-mean of M values (TMM) or Quantile normalization, also rely on this assumption. In the parlance of compositional data these methods assume *Compositional Invariance*, i.e. the underlying composition is statistically independent of the total size of the composition (the total counts for a sample, t).

Compositional invariance (CI) is an important property for RNA-Seq data which enables the comparison of samples with differing read depths. However, it is well documented that the quality of RNA-Seq depends on the read depth of the sequencing run with higher read-depths associated with higher quality data [25, 26]. Read depth may affect the measurement of relative abundances for the target RNA sequences as some targets may receive proportionally more reads as the read depth increases. This would be a direct violation of CI and could lead seemingly differential expression between samples with different read depths, even after normalization. Another form of CI violation, that is perhaps more likely in RNA-Seq experiments, is the dependence between the variance of read counts and the read depth.

Aitchison [1] outlined a simple model for testing compositional covariance using the ALR transformation,

$$[y_1 \dots y_d] = \begin{bmatrix} 1 & t \end{bmatrix} \begin{bmatrix} \alpha_1 & \cdots & \alpha_d \\ \beta_1 & \cdots & \beta_d \end{bmatrix} + [e_1 \dots e_d], \tag{9}$$

where $y_1 \dots y_d$ are the d ALR transformed components, t is the vector of sample total aligned reads, $\alpha_1 \dots \alpha_d$ are the probe specific log-ratio intercepts, and $\beta_1 \dots \beta_d$ are the coefficients relating the the total aligned reads to the relative expression of the probe. A test for compositional invariance for the experiment then becomes a test of the null hypothesis, $H_o : \beta_1 = \cdots = \beta_d = 0$. This test can be re-parameterized to test for dependence between the variance and total aligned reads as well.

Unfortunately, the small sample sizes and large number of probes typically associated with RNA-Seq experiments complicates the application of Aitchison's model. We propose an alternative visualization for simultaneously detecting both violations of compositional invariance described above. We use the multivariate Aitchison distance (8) between all pairs of samples in a heat-map with the samples ordered by total aligned reads. If CI is violated we expect pairs samples with similar total aligned reads will have smaller scalar distances than those with large differences in total aligned reads. This will result in visual clustering around the 45° axis. If the variance depends on the total aligned reads, we expect the scalar distance between sample pairs to decrease with increasing read depth resulting in a visual gradient in the distance heat map. To reduce the visual noise associated with outlier samples in the heat-map we also provide a dot-plot of the distance between each CLR transformed sample and the compositional center of the samples in the top quartile of total reads.

We demonstrate this visualization with two sets of miRNA samples (Fig. 3) and two sets of mRNA samples (Fig. 4). The miRNA samples are composed of 40 technical replicates each of (1) plasma samples and (2) brain samples. In the miRNA data there is a clear gradient along the 45° axis for the plasma samples (Fig. 3a). This indicates a dependence between the total aligned reads and the variance of the samples (as indicated by the increasing multivariate distance between replicates as the total aligned reads decreases). In contrast, there is no clear gradient in the brain

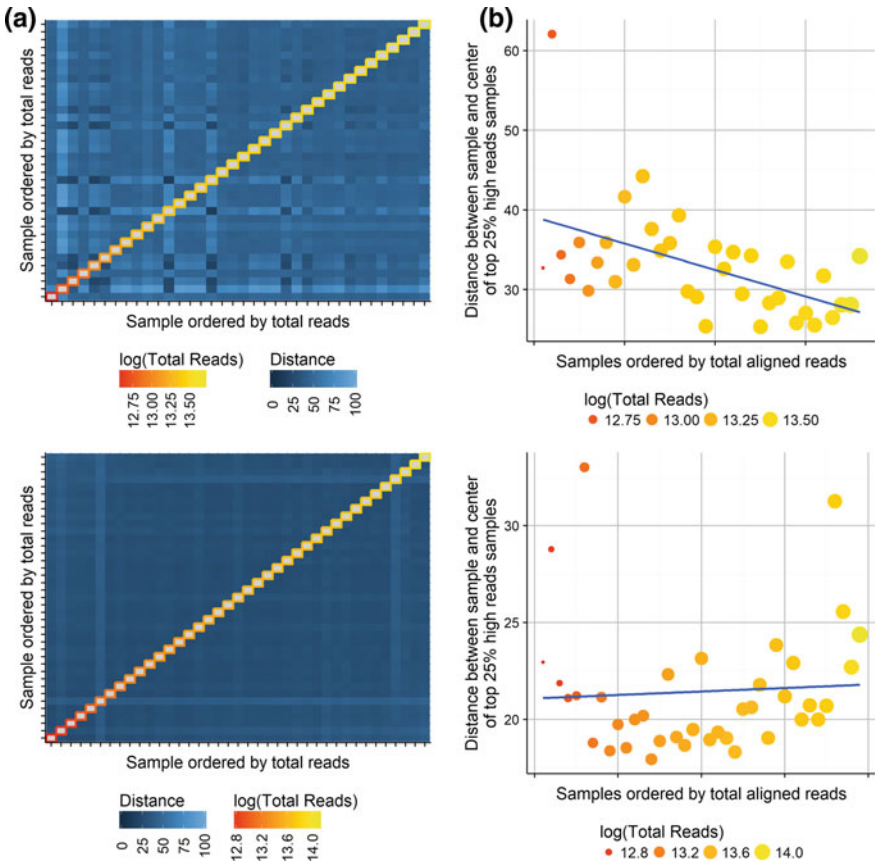


Fig. 3 Two sets of miRNA samples with samples in (1) showing a violation of compositional invariance and (2) showing compositional invariance

samples (Fig. 3b). The mRNA samples are composed of (A) 16 technical replicates of diseased pancreas tissue and (B) 16 technical replicates of normal pancreas tissue. In the diseased pancreas samples there is a clear gradient with low total aligned read samples more distant from samples with greater total aligned reads (Fig. 4(1)). This indicates that the composition is dependent on the total aligned reads, a violation of compositional invariance for these samples. In contrast, the normal pancreas samples show no such pattern related to total aligned reads (Fig. 4(2)).

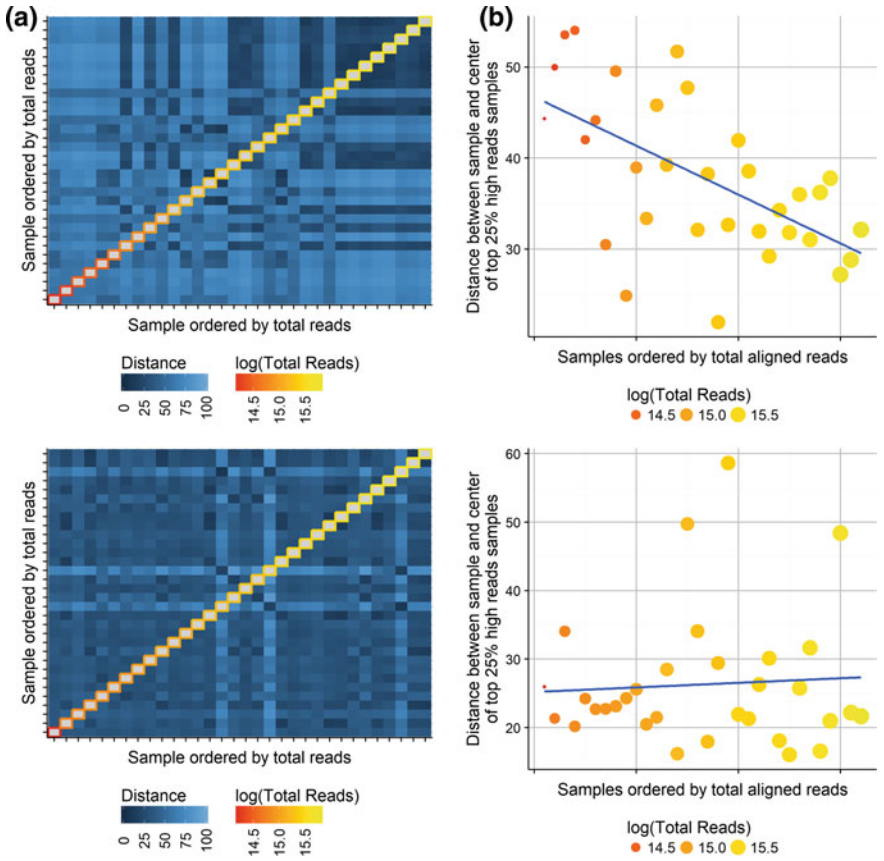


Fig. 4 Two sets of mRNA samples with samples in (1) showing a violation of compositional invariance and (2) showing compositional invariance

5 Batch Effects and Normalization

Batch effects arising from differing laboratory conditions or operator differences have been identified as a problem in high-throughput measurement systems [10, 14]. Identifying and controlling for batch effects is a critical step in the transition of RNA-Seq from the lab to the clinic. Batch effects are typically identified with a hierarchical clustering (HC) method or principal components analysis (PCA). For both methods, the multivariate distance between the samples is visualized, either in a biplot for PCA or a dendrogram for HC, to check for the existence of clusters of samples related to batch. The compositional nature of RNA-Seq data has important implications for the detection of batch effects due to the incompatibility with standard measures of distance between compositions as noted above [1, 19].

Principle components analysis is sensitive to differences in scale (total number of reads) among the variables, failure to remove these difference can mask potential batch effects and leave unwanted technical variation in the data. As noted above, most normalization methods use a scaling factor calculated for each sample to re-scale the read count for each gene within the sample [11]. The CLR transformation can similarly be viewed as a scaling normalization (with the scale factor chosen as the inverse of the geometric mean $1/g(x)$). Unlike other normalization methods, the CLR transformation has the added benefit of being applied at the individual sample level, not experiment wise like quantile or median normalization [9], and requires no assumptions about differential expression among samples like quantile or median ratio normalization [6, 22]. This makes it particularly well suited for the clinic where there are generally no reference samples for normalization. Most importantly, CLR transformation allows the use of Euclidean geometry, such as Euclidean or Mahalanobis distance, so that standard PCA or HC applied to transformed samples can be interpreted in the traditional way [3] .

We demonstrate the use of the compositional biplot to detect batch effects using 120 technical replicates of three sample types: brain, plasma, and FFPE. Samples were prepared using the EdgeSeq Whole Transcriptome miRNA assay which measures 2,280 targets including including 11 control probes and 2,269 unique miRNA probes. All sequencing was performed on an Illumina Mi-seq$^{(tm)}$ sequencer.

We perform a PCA on log-transformed and CLR transformed data. We then construct form-biplots of the first two principle components for each transformed data set (Fig. 5). The differences between the 3 samples types (brain, plasma, and FFPE) dominate the first two principle components for both data sets. However, the CLR transformed data provides tighter clusters, relative to the distance between the clusters, than the log-transformed raw data. There is also a single FFPE sample which is

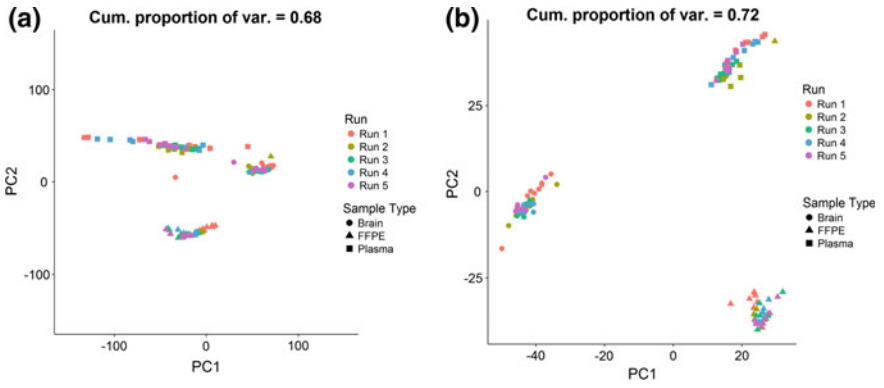


Fig. 5 Principle component analysis of (**a**) log-transformed and (**b**) CLR-transformed read count data. The differences between sample types is much greater than the batch effects in both transformation. The CLR transformation results in tighter sample type clusters resulting from less variation along the first principle component

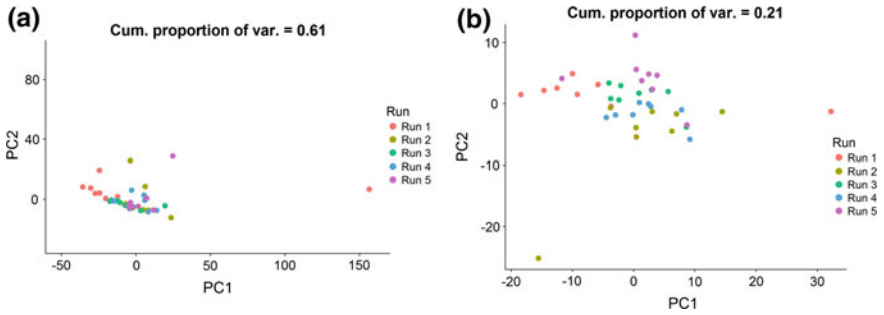


Fig. 6 Principle component analysis of only brain samples from (**a**) log-transformed and (**b**) CLR-transformed read count data. The batch effects are more easily identified in the CLR transformed data

closer to the brain samples than the other samples. It is worth noting that this sample would have been removed using our proposed quality control metric. Since the sample type differences overwhelm the potential batch effects we performed a second PCA on only the brain samples for both transformed data sets (Fig. 6). Both biplots exhibit clustering by batch but the CLR transformed data shows better separation between the batches.

6 Discussion

Our fractional read allocation metric can identify problematic samples which arise from multiple failure modes, e.g. a low quality sample or a sequencing problem. However, it is conceivable that a sample might have an unusually low (or high) number of reads and still provide quality information. In certain experimental designs one might be able to further evaluate these samples with a PCA biplot on the CLR transformed data. In our PCA analysis we identified a FFPE sample which would have failed our quality control and was clearly very different from the other technical replicates. However, if this sample had remained quite similar to the other FFPE replicates this would have provided information that the sample may still be valuable. In this way, the quality control metric and PCA biplot can be used in tandem to provide additional information about the quality of a sample.

The compositional invariance visualization is a logical extension of the sample quality control metric since the assumption of the sample quality control is that the total number of aligned reads is related to the proportional allocation of reads within the sample. As noted above samples which violate the compositional invariance property may still contain valuable information. The identification compositional invariance violations allows the investigator to account for the dependency between the total aligned reads and the relative abundance of transcripts within the samples when modelling.

The principal components analysis biplot is a well know dimension reduction visualization. For the current data the dimension is reduced from 2,280 probes to 2 principle components. The utility of the data reduction, including the quality of the approximation of the multivariate distance between the samples, is proportional to the amount of variance explained by these two principle components. In our data the first two principle components explain between 72 and 21% of the variation in the data. The analysis with the lowest percent of variation explained by the first 2 components is of the CLR-transformed brain samples. Surprisingly, batch effects are still visible in this plot, in which case they can be removed [17].

As RNA-Seq makes the transition from the research laboratory to the clinic there is a need for robust quality control metrics. The realization that RNA-Seq data are compositional opens the door to the existing body of theory and methods developed by John Atichison and others. We show that the properties of compositional data can be leveraged to develop new metrics and enhance existing methods.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall, Ltd. (1986). http://dl.acm.org/citation.cfm?id=17272
2. Aitchison, J.: On criteria for measures of compositional difference. Math Geol **24**(4), 365–379 (1992). https://doi.org/10.1007/BF00891269. http://link.springer.com/10.1007/BF00891269
3. Aitchison, J., Greenacre, M.: Biplots of compositional data. J R Stat Soc Series C (Appl Stat) **51**(4), 375–392 (2002). https://doi.org/10.1111/1467-9876.00275. http://doi.wiley.com/10.1111/1467-9876.00275
4. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V.: Logratio analysis and compositional distance. Math Geol **32**(3), 271–275 (2000). https://doi.org/10.1023/A:1007529726302
5. Aitchison, J., Shen, S.: Logistic-normal distributions: some properties and uses. Biometrika **67**(2), 261–272 (1980). https://doi.org/10.1093/biomet/67.2.261. https://www.researchgate.net/publication/229099731_Logistic-Normal_Distributions_Some_Properties_and_Uses
6. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biol **11**(10), R106 (2010). https://doi.org/10.1186/gb-2010-11-10-r106. http://www.biomedcentral.com/content/pdf/gb-2010-11-10-r106.pdf
7. Ben-Gal, I.: Outlier detection. In: Data Mining and Knowledge Discovery Handbook, pp. 117–130. Springer, US (2009). https://doi.org/10.1007/978-0-387-09823-4_7. http://link.springer.com/10.1007/978-0-387-09823-4_7
8. Billheimer, D., Guttorp, P., Fagan, W.F.: Statistical interpretation of species composition. J Am Stat Assoc **96**(456), 1205–1214 (2001). https://doi.org/10.1198/016214501753381850. http://www.jstor.org/stable/3085883
9. Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England) **19**(2), 185–193 (2003). http://www.ncbi.nlm.nih.gov/pubmed/12538238
10. Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., Liu, C.: Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One **6**(2) (2011). https://doi.org/10.1371/journal.pone.0017238

11. Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, N.S., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., Jaffrezic, F.: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefngs Bioinform **14**(6), 671–683 (2013). https://doi.org/10.1093/bib/bbs046
12. Hawkins, D.M.: Identification of Outliers. Springer Netherlands, Dordrecht (1980). https://doi.org/10.1007/978-94-015-3994-4. http://link.springer.com/10.1007/978-94-015-3994-4
13. Law, C.W., Chen, Y., Shi, W., Smyth, G.K.: voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol **15**(2), R29 (2014). https://doi.org/10.1186/gb-2014-15-2-r29. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29
14. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet **11**(10), 733–739 (2010). https://doi.org/10.1038/nrg2825. http://dx.doi.org/10.1038/nrg2825
15. Lovell, D., Müller, W., Taylor, J., Zwart, A., Helliwell, C.: Proportions, percentages, PPM: do the molecular biosciences treat compositional data right? In: Compositional Data Analysis: Theory and Applications, pp. 191–207. Wiley (2011). https://doi.org/10.1002/9781119976462.ch14. http://dx.doi.org/10.1002/9781119976462.ch14
16. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., Bähler, J.: Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol **11**(3), e1004,075 (2015). https://doi.org/10.1371/journal.pcbi.1004075. http://www.ncbi.nlm.nih.gov/pubmed/25775355
17. Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H., Zhao, C., Elloumi, F., Shi, W., Thomas, R., Lin, S., Tillinghast, G., Liu, G., Zhou, Y., Herman, D., Li, Y., Deng, Y., Fang, H., Bushel, P., Woods, M., Zhang, J.: A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J **10**(4), 278–291 (2010). https://doi.org/10.1038/tpj.2010.57. http://www.ncbi.nlm.nih.gov/pubmed/20676067www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2920074
18. Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V.: Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Math Geol **35**(3), 253–278 (2000). https://doi.org/10.1023/A:1023866030544. http://link.springer.com/article/10.1023/A%3A1023866030544
19. Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., Buccianti, A., Nardi, G., Potenza, R.: Measures of difference for compositional data and hierarchical clustering methods. In: Proceedings of IAMG, vol. 98, no. 1, pp. 526–531 (1998)
20. Martn-Fernndez, J.A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J.: Bayesian multiplicative treatment of count zeros in compositional data sets. Stat Model **15**(2), 134–158 (2015). http://ezproxy.library.arizona.edu/login?url=https://search-proquest-com.ezproxy1.library.arizona.edu/docview/1673859465?accountid=8360. (Copyright-SAGE Publications Apr 2015; Last updated 19 Sep 2015)
21. Pearson, K.: Mathematical contributions to the theory of evolution.-On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond **60**, 489–498 (1896). https://archive.org/details/philtrans00847732 (Free Download & Streaming: Internet Archive.)
22. Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol **11**(3), R25 (2010). https://doi.org/10.1186/gb-2010-11-3-r25
23. Robinson, M.D., Smyth, G.K.: Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics **9**(2), 321–332 (2007). https://doi.org/10.1093/biostatistics/kxm030. http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxm030

24. Sanford, R.F., Pierson, C.T., Crovelli, R.A.: An objective replacement method for censored geochemical data. Math Geol **25**(1), 59–80 (1993). https://doi.org/10.1007/BF00890676. http://link.springer.com/10.1007/BF00890676
25. Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P.: Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet **15**(2), 121–132 (2014). https://doi.org/10.1038/nrg3642. http://www.nature.com/doifinder/10.1038/nrg3642
26. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A.: Differential expression in RNA-seq: a matter of depth. Genome Res **21**(12), 2213–2223 (2011). https://doi.org/10.1101/gr.124321111. http://www.ncbi.nlm.nih.gov/pubmed/21903743www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3227109
27. Tukey, J.W.J.W.: Exploratory Data Analysis. Addison-Wesley Publication, Co (1977)

Part VII
Omics Data Analysis

Leveraging Omics Biomarker Data in Drug Development: With a GWAS Case Study

Weidong Zhang

Abstract Biomarkers have proven powerful for target identification, understanding disease progression, drug safety and treatment responses in drug development. Recent development of omics technology has offered great opportunities for identifications of omics biomarkers at low cost. Although biomarkers have brought many promises to drug development, steep challenges arise due to high dimensionality of data, complexity of technology and lack of full understanding of biology. In this article, the application of omics data in drug development will be reviewed. A genome wide association study (GWAS) will be presented.

Keywords Biomarker · Omics · Simulation · GWAS

1 Introduction

1.1 Overview of Biomarker in Drug Development

Precision medicine has gained great popularity in the last decade. In 2015, a total of $215 million investment was budgeted to develop national databases after President Barack Obama announced a 'Precision Medicine Initiative'. The goals of this initiative are two-folds: (a) to focus on precise cancer drug development and (b) to build a database with knowledge of biomarkers that can be used for a broader range of diseases [4].

Biomarkers are indispensable assets to precision medicine and overall drug development. A biomarker can be defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [3]. Biomarkers have been identified as important factors to improve probability of success in drug development. From a recent analysis performed by Thomas et al. 9,985 phase transition trials from

W. Zhang (✉)
Pfizer Inc., Cambridge, MA, USA
e-mail: weidong.zhang2@pfizer.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_22

2006 to 2015 were analyzed. Phase transitions are defined as either a drug candidate advances into the next phase of development or is suspended by the sponsor. It was shown that the success rate from Phase I to approval was increased to ~25% when selection biomarkers were used as compared to ~8% for those programs without [2].

In this article, an overview of the biomarker discovery and omics biomarker technologies will be presented. The statistical considerations in omics biomarker analysis will be discussed. A GWAS case study will be presented for illustration of application of omics technology.

1.2 Classification of Biomarkers

Depending on their functions, biomarkers can be classified into predictive biomarkers [8], prognostic biomarkers, pharmacodynamic (PD) biomarkers and surrogate biomarkers. A predictive biomarker predicts a patient's clinical response to the treatment he/she received. Predictive biomarkers are of particular interest in precision medicine due to the fact that a predictive biomarker can be used to identify a patient population that potentially respond or respond better to the new treatment or avoid side effects of a treatment. A recent successful story was reported by Tesaro, Inc, in which there was a study that patients who carried the germline BRCA mutation had progression-free survival (PFS) of 21 months after receiving niraparib as compared to 5.5 months in the control group (Tesaro 2017). A prognostic biomarker, however, can predict a patient's clinical outcome in a way that is independent of any treatment. An example of a prognostic biomarker can be found in a report by Paik et al. in which case a 21-gene recurrence score was used to predict breast cancer recurrence and overall survival in node-negative, tamoxifen-treated breast cancer [15]. A prognostic biomarker may not be used to predict treatment response. However, it may be helpful to a physician to decide whether chemotherapy should be prescribed for high risk patients or avoided by low risk patients. Many biomarkers, however, may be both prognostic and predictive biomarkers in nature, for example, in breast cancer estrogen receptor (ER) can be used as a prognostic biomarker because ER negative patients have a higher risk of relapse than ER-positive patients. On the other hand, the anti-estrogen tamoxifen is more effective in preventing breast cancer recurrences in ER-positive patients than in ER-negative patients, which constitutes ER as a predictive biomarker. Predictive biomarkers will be focused in most of the discussions of this article due to their unique value in patient stratification in clinical trial design.

A PD biomarker can be used to quantify drug modulation and demonstrate principle of mechanism. Frequently, PD biomarkers are useful tools in early clinical trials such as phase 1 to provide guidance for dose selection. PD biomarkers are critical to demonstrate three pillars (target exposure, target binding and target modulation) in drug discovery. It was shown that trials with successful demonstration of these three pillars had much high overall successful rate in the subsequent proof of concept (POC) studies [14].

A surrogate biomarker may be used as a substitute for a clinical endpoint of interest. According to the Biomarker Working Group [3], a surrogate endpoint is defined as "a biomarker intended to substitute for a clinical endpoint. A clinical investigator uses epidemiological, therapeutic, pathophysiological, or other scientific evidence to select a surrogate endpoint that is expected to predict clinical benefit, harm, or lack of benefit or harm". For example, many imaging markers such as total brain volume, hippocampal volume, etc. have been used as surrogate markers in Alzheimer's disease since those imaging markers seem to correlate well with disease progression [11]. However, Fleming and DeMets [7] pointed out that correlation does not automatically guarantee a surrogate status. In some circumstances, a drug may be efficacious on the marker that correlates well with the clinical endpoint but may not have any effect on the clinical endpoint of interest.

1.3 Overview of Omics Biomarker and Cutting-Edge Technologies

Omics technologies refer to the new advanced technologies that are primarily used for the global detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in a specific biological sample. Omics biomarkers are typically high-dimensional as illustrated in Fig. 1.

For example, gene expression profile technologies can measure abundance of all the genes (~25 k) in the transcriptome for each sample, which gives scientists an unbiased view of the global biological landscape. The omics technologies started to emerge from the late 20th century when Microarray was first available for gene profiling of whole transcriptome and whole genome genotyping. The early DNA microarray consists of a solid glass surface and a collection of DNA fragments, known as probes or oligos attached to the surface. A probe is a fragment of a section of a gene that can be used to uniquely hybridize a cDNA or cRNA from a fluorescent molecule labeled target sample. The fluorescent intensity of a probe-target hybridization is quantified to determine the abundance of DNA molecules in the target sample. The microarray technology has evolved greatly over the last decade; however, it suffers from major drawback such as dependence on known genes, relatively low sensitivity and low dynamic range. Early in the 21st century, the next generation sequencing (NGS) technologies started to show new promises by offering variety of novel methods for genomics study. Over the last decade, turnaround time and cost of sequencing have been substantially reduced as a result of the advancement of this new technology. It was estimated that the cost of sequencing a genome dropped from $100 million in 2001 to $1,245 in 2015 Wetterstrand [22], and the turnaround time was shortened from years in the late 90 s to days including analysis in 2016 [13]. As of today, NGS technology has been widely applied to a variety of biomedical research areas including transcriptome profiling, identification of new RNA splice variant, genome-wide genetic variants identification, genome-wise epigenetic modification

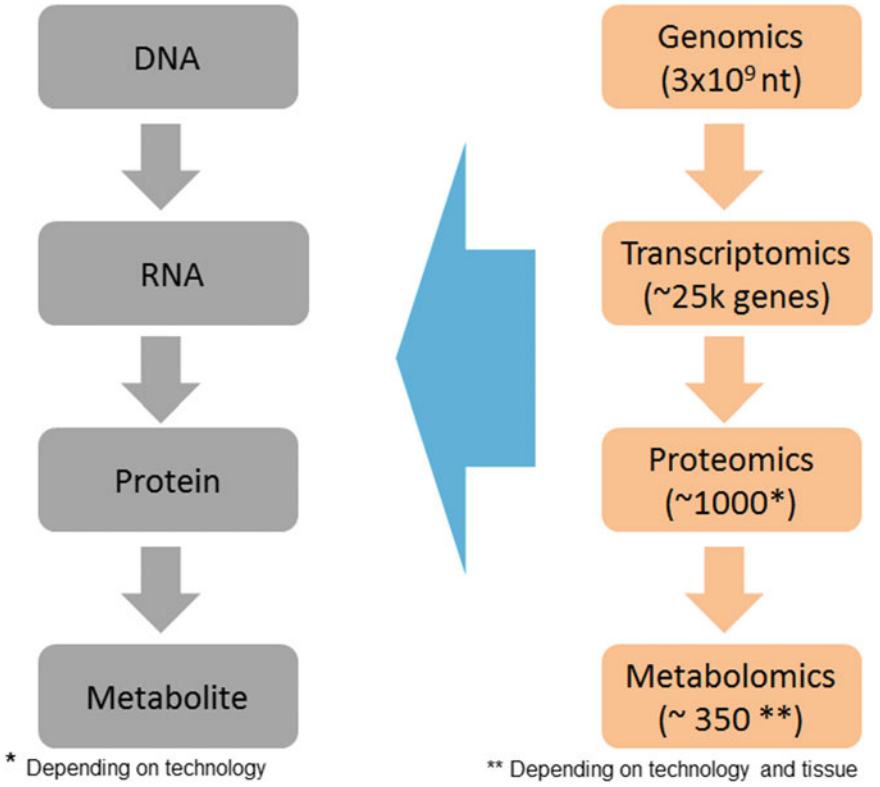


Fig. 1 Omics provide paramount view of biological cascade

and DNA methylation profiling etc. In particular, NGS technology is a promising tool for cancer research, given the "disorder of genome" nature of cancer disease. In cancer research, NGS has significantly enhanced our ability to conduct comprehensive characterization of cancer genome to identify novel genetic alterations, and has significantly helped to dissect tumor complexity. Coupling with sophisticated computational tools and algorithms, significant achievements have been accomplished for breast cancer, ovarian cancer, colorectal cancer, lung cancer, liver cancer, kidney cancer, head and neck cancer, melanoma, acute myeloid leukemia (AML) etc. [18].

Choice of technologies should be made based on the goal of the study. Unbiased high dimensional technology gives maximum information but may not be an efficient choice if the pathway under study is relatively well understood. For example, in oncology, many times scientists want to focus on a select set of genes, gene regions, or amplicons that have known associations with cancer, in which case targeted sequencing panel may be used instead of whole exome or whole genome. In pharmacology, genes from a specific pathway, e.g. JAK-STAT pathway may be of

interest to study drug modulation for JAK inhibitors, and a Taqman low density array (TLDA) panel may be sufficient instead of whole transcriptome.

2 Considerations of Statistical Analysis

Analysis of high dimensional omics biomarker needs special statistical considerations. Conventional statistics focus on problems with large number of experimental units (n) as compared to small number of features or variables (p) measured from each unit. High dimensional biomarker data are often large in p and small in n. For example, in GWAS in a clinical trial, about one million single nucleotide polymorphisms (SNPs) can be collected using microarray from each subject with the number of subjects ranging from dozens to hundreds. Many statistical methods have been developed in analysis of high-dimensional omics data. Typical methods include clustering analysis for pattern discovery, and univariate or multivariate regression and supervised and unsupervised classification analysis to predict disease status [9]. For expression based omics data such as gene expression, proteomics, metabolomics etc., dimension reduction is considered as the first step before subsequent analysis. Dimension reduction techniques include descriptive statistical approach such as coefficient variation (CV) filtering, by which biomarkers with low CV are removed from subsequent regression/ANOVA analysis. This approach is particularly useful when computing power is limited. However, the CV filtering step is typically skipped with today's high computational capacity, and instead, a univariate regression analysis is used for both dimension reduction and inference.

Univariate single biomarker analysis is popular due to the simplicity and interpretation benefit, but is often criticized for being oversimplifying biology by including only one biomarker in the regression model. Multivariate and multiple regressions consider multiple biomarkers in a model become more popular for being able to take into account (1) Complexity of disease mechanism requires an integrated information from multiple biomarkers to explain more biological variations. (2) Interactions between biomarkers cannot be modeled with single biomarker analysis. (3) Correlation and dependency among biomarkers cannot be handled with single biomarker analysis.

Another challenging area in statistical analysis of high-dimensional omics data is how to control false discovery rate (FDR), especially with presence of correlation structure among biomarkers. Family-wise error rate (FWER) adjustment techniques such as Bonferroni correction calculate the probability of making at least one type I error, often considered too conservative. FDR based approaches control the probability of false discoveries from the "positive" findings (rejected null hypotheses). Therefore, FDR procedures are more powerful than FWER but at the cost of high type I errors. Common FDR based methods include Benjamini and Hochberg (BH) method [1] and q value method [19]. The BH method first finds the largest k such that $\mathrm{P}(m) \leq k/m * \alpha$, where m stands for m tests and α is a predefined FDR level. Second, the null hypothesis for each H(i)with $\mathrm{i} = 1 \ldots \mathrm{k}$ are rejected. The q value

method calculates q values that are considered as quantification of false discovery rate. Both the qvalue and BH methods allow dependence of testing. However, the qvalue method may provide more power than BH method, and has been widely used in many omics studies [19].

For GWAS, determination of genome-wide significance threshold is difficult due to as many as millions of statistical testing and complex genetic linkage disequilibrium (LD) structures. Many procedures have been proposed including Bonferroni, FDR, Sidak, and permutation etc. however, it was suggested that a $p = 5 \times 10^{-8}$ can be used for genome wide significance and $p = 1 \times 10^{-7}$ can be used as a suggestive threshold at practical level [16, 17]. Fadista et al. recently studied different scenarios and suggested that P-value thresholds should take into account impact of LD thresholds, MAF and ancestry characteristics. Further, they confirmed a p value threshold of 5×10^{-8} was appropriate for European population with MAF > 5%. However, they suggested that the P-value threshold needs to be more stringent with European ancestry with low MAF (3×10^{-8} for MAF >= 1%) due to the increasing number of variants and the lower LD between less frequent variants [6].

3 A Case Study—A Novel Bootstrap Based Model Average Approach for GWAS Using Outbred Mice

In a study conducted by Zhang et al. [23], a total of 288 outbred mice were used to identify genetic polymorphisms that may be associated with phenotypes such as High-density lipoprotein (HDL), Systolic blood pressure (SBP), Triglyceride (TG), Glucose (GLU) and Albumin Creatinine Ratio (ACR). Outbred mice are similar to human population with regard to genetic diversity but offer great accessibility. The genotype were measured using Affymetrix® Mouse Diversity Array covering ~620 k SNPs. Population structure was first evaluated by calculating correlations between SNP pairs within 50 Mb sliding window across the whole genome. A kinship matrix between the individual animals was calculated based on identity by state among the 44,428 SNPs using Efficient Mixed-Model Association (EMMA) [10]. Single-locus association genome scans were performed by ANOVA and EMMA taking into account population structures. To assess genome-wide significance of the association statistics, a novel simulation technique was used as illustrated in the following steps:

(1) Each phenotype was transformed using van derWaerden's scores [5].
(2) Genetic and residual variances of the transformed data for each phenotype were estimated using EMMA. For each phenotype, 288 trait values were generated by sampling from a multivariate normal distribution using the mvrnorm function in R with covariance matrix defined by the estimated kinship.
(3) The observed trait values were reordered based on the rank orders of the simulated values. By doing so, permutation was performed on the original data that retains the correlation structure implied by the kinship matrix.

(4) A genome scan using the permuted trait values and recorded the largest –log(p) scores. This was repeated 100 times. A generalized extreme value distribution was fitted to these scores and significance thresholds were derived from the quantiles of this distribution [11].

It is well known that the biological process is a complex system that involves multiple components. To obtain realistic estimates of effect sizes, multilocus analysis was performed using forward stepwise regression with bootstrap resampling [21]. First, 100 data sets were generated by sampling with replacement from the 288 animals. Forward stepwise regression on each resampled data set was performed to obtain a multilocus regression model with 20 SNPs. The choice of 20 is arbitrary just to ensure that the number of SNPs in the regression model is more than the number that could significantly influence the phenotype. A resample model inclusion probabilities (RMIP) for each SNP m was calculated as

$$RMIP_m = \frac{1}{R}\sum_{r=1}^{R} i_{rm}$$

where R = 100 is the number of resampled data sets $i_{rm} = 1$ if at least one SNP within ±w Mb of SNP m was included in the model of sample r, otherwise $i_{rm} = 0$. We varied the window size w from ±0.5 Mb to ±4 Mb.

Precision of the locations of the GWAS hits was not well understood. A simulation approach was used in this study to assess the genome-wide average precision of mapping in this population. The steps are illustrated as follows:

(1) A SNP was randomly selected from the genome and trait values were simulated assuming that SNP selected was the causal locus.
(2) Simulate an effect size corresponding to the same percentage of total variance explained as the HDL QTL on Chromosome 1. Phenotype values were sampled from a multivariate normal distribution using mvrnorm in R with correlation structure defined by the kinship matrix and the genetic and residual variances were the same as those estimated for HDL.
(3) The selected SNP was removed from the data and a genome scan was performed using EMMA. The distance between the SNP with highest –log(p) and the target SNP was recorded.
(4) The process from (2) to (3) was repeated 1000 times, and the distribution of distances from the peak to the target SNP was computed.

The significance thresholds were evaluated by simulation and unrestricted permutation, and was applied to each of the following three methods for measuring association: the trend test, ANOVA test and EMMA. The estimated genome-wide significance thresholds for glucose, HDL cholesterol, systolic blood pressure, and triglycerides were similar across all of these combinations. Values ranged from 5.12 to 5.90, but no single method or trait was consistently higher or lower than another.

Two highly significant loci associated with HDL were identified from chromosome 1 and 5. There seemed to be an association with SBP on proximal Chromosome

10 at 7 Mb that exceeded the genome-wide 0.05 thresholds for the simple trend and ANOVA tests, however, it was not significant for the EMMA test. The logACR trait was the most variable of the five traits two loci seemed to be significant on Chromosome 5 at 147 Mb and Chromosome 11 at 88 Mb using the 0.05 thresholds from either simple trend test or the ANOVA test. The results from multilocus genome-wide scans using forward stepwise variable selection on bootstrapped samples showed that RMIP for the two loci Chromosome 1 at 173 Mb and Chromosome 5 at 126 Mb for HLD were 100% but the hit on Chromosome 1 at 181 Mb was never included as an independent QTL in the multilocus analysis, which indicate this method may be useful for prioritization of GWAS hits.

The simulated precision analysis showed that a GWAS hit in this population with a large effect, e.g. as large as the effect of the HDL hit on chromosome 1, can be localized within 1.34 Mb of the greatest association peak. This approach could be expanded to a range of effect sizes in any genotyped population sample including human GWA studies.

This study demonstrates that the GWA analysis employed here can be successfully applied to outbred mice populations to identify genetic variants underlying complex traits.

4 Summary

Omics technology and genomics data have proven to be powerful tools in drug development. Complexity of the biology, technology and high dimensionality of omics data require extensive attention on novel analytical methodology development. Using an example in GWAS, it can be shown that simulation-based method offers many advantages in regards to prioritizing multiple GWAS hits, determination of genome wide threshold considering population structure, and estimation of precision of GWAS hits. With whole-genome sequencing becoming a new norm for genotyping, transcriptome profiling and many other genomic quantification applications, additional challenges associated with handling data quality control, interaction modeling and integration of multiple types of biomarkers will manifold more complex. With collective efforts from the statistical and other analytical communities, significant progresses have been made and will greatly facilitate using these omics information to elucidate disease mechanisms.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. **57**(1), 289–300 (1995)
2. Biomarkers Definition Working Group Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin Pharmacol Therapeutics. **69**, 89–95 (2001)

3. Collins, F.S., Varmus, H.: A new initiative on precision medicine. N. Engl. J. Med. **372**(9), 793–795 (2015)
4. Conover, W.J.: Practical Nonparametric Statistics. John Wiley Chichester, New York (1999)
5. Fadista, J., Manning, A., Florez, J., Groop, L.: The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur. J. Hum. Genet. **24**, 1202–1205 (2016)
6. Fleming, T.R., DeMets, D.L.: Surrogate end points in clinical trials: are we being misled? Ann. Intern. Med. **125**(7), 605–613 (1996)
7. Gosho, M., Nagashima, K., Sato, Y.: Study designs and statistical analyses for biomarker research. Sensors **12**, 8966–8986 (2012)
8. Johnstone, I., Titterington, D.: Statistical challenges of high-dimensional data. Phil. Trans. R. Soc. A **367**, 4237–4253 (2009)
9. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., et al.: Efficient control of population structure in model organism association mapping. Genetics **178**, 1709–1723 (2008)
10. Katz, R.: Biomarkers and Surrogate Markers: an FDA Perspective. NeuroRx **1**(2), 189–195 (2004)
11. Knijnenburg, T.A., Wessels, L.F., Reinders, M.J., Shmulevich, I.: Fewer permutations, more accurate P-values. Bioinformatics **25**, i161–i168 (2009)
12. Meienberg, J., Bruggmann, R., Oexle, K., Matyas, G.: Clinical sequencing: is WGS the better WES? Hum. Genet. **135**, 359–362 (2016)
13. Morgan, P., Van Der Graaf, P.H., Arrowsmith, J., Feltner, D.E., Drummond, K.S., Wegner, C.D., Street, S.D.: Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival. Drug. Discov. Today **17**, 419–424 (2012)
14. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., et al.: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N. Engl. J. Med. **351**(27), 2817–2826 (2004)
15. Panagiotou, O.A., Ioannidis, J.P.: Genome-wide significance project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int. J. Epidemiol. **41**(1), 273–86 (2012)
16. Pe'er, I., Yelensky, R., Altshuler, D., Daly, M.: Estimation of the multiple testing burden for Genomewide association studies of nearly all common variants. Genet. Epidemiol. **32**, 381–385 (2008)
17. Shyr, D., Liu, Q.: Next generation sequencing in cancer research and clinical application. Biol. Proced. Online **15**, 4 (2013)
18. Storey, J.D.: A direct approach to false discovery rates. J. Roy. Stat. Soc. **64**, 479–498 (2002)
19. TESARO's Niraparib Significantly Improved Progression-Free Survival for Patients With Ovarian Cancer in Both Cohorts of the Phase 3 NOVA Trial (2016). http://ir.tesarobio.com/releasedetail.cfm?releaseid=977524
20. Thomas, D.W., Burns, J., Audette, J., Carroll, A., Dow-Hygelund, C., Hay, M.: Clinical Development Success Rates 2006–2015. June 2016. https://www.bio.org/sites/default/files/Clinical%20Development%20Success%20Rates%202006-2015%20-%20BIO,%20Biomedtracker,%20Amplion%202016.pdf
21. Valdar, W., Holmes, C.C., Mott, R., Flint, J.: Mapping in structured populations by resample model averaging. Genetics **182**, 1263–1277 (2009)
22. Wetterstrand, K.A.: DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) (2016). www.genome.gov/sequencingcostsdata. Accessed 23 Dec 2016
23. Zhang, W., Korstanje, R., Thaisz, J., Staedtler, F., Harttman, N., Xu, L., Feng, M., Yanas, L., Yang, H., Valdar, W., Churchill, G.A., DiPetrillo, K.: Genome-wide association mapping of quantitative traits in outbred mice. G3: Genes Genomes Genet. **2**(2), 167–174 (2012)

A Simulation Study Comparing SNP Based Prediction Models of Drug Response

Wencan Zhang, Pingye Zhang, Feng Gao, Yonghong Zhu and Ray Liu

Abstract Lack of replication on findings and missing heritability are two of the major challenges in Pharmacogenetics (PGx) studies. Recently developed statistical methods for genome-wide association studies offer greater power both to identify relevant genetic markers and to predict drug response or phenotype based on these markers. However, the relative performance of these methods has not been thoroughly studied. Here, we present several simulations to compare the performance of these analysis methods. In our first simulation, we compared five different approaches: Elastic Net (EN), Genome-wide Association Study (GWAS)+EN, Principal Component Regression (PCR), Random Forest (RF) and Support Vector Machine (SVM). The results showed that EN has the smallest test mean squared error (MSE) and the highest portion of causal SNPs among identified SNPs. In the second simulation, we compared three approaches, GWAS+EN, GWAS+RF and GWAS+SVM. The GWAS+RF has the smallest test MSE and the highest causal percent. In the third simulation study, we compared two cross validation procedures: GWAS+EN versus modified learn and confirm cross validation GWAS+EN. The latter approach demonstrated better prediction accuracy at the expense of greatly increased computational time.

W. Zhang (✉)
Takeda Develop Center, B3 4202A. One Takeda PKWY, Deerfield, IL 60015, USA
e-mail: Wencan88@msn.com

P. Zhang
Merck, 126 E. Lincoln Avenue, Rahway, NJ 07065, USA
e-mail: pingyezhangusc@gmail.com

F. Gao
Biogen, 300 Binney Street, Cambridge, MA 02142, USA
e-mail: gaof60@gmail.com

Y. Zhu
Shanghai Henlius Biotech Inc, Building B #806, No.1289 Yishan Road, Shanghai 200233, China
e-mail: yonghongny@yahoo.com

R. Liu
Takeda, 40 Landsdowne Street, Cambridge, MA 02139, USA
e-mail: Ray.Liu@Takeda.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_23

Keywords Genomics · GWAS · Predictive modeling · Machine learning · Cross validation

1 Introduction

Over the last decade, many clinically important single nucleotide polymorphisms (SNPs) and SNP-harboring genomic regions have been identified by Genome-Wide Association Studies (GWAS). These variants may be used as biomarkers predictive of disease susceptibility or treatment response, which can support both clinical decision making and drug development [1]. However, non-reproducible findings and missing heritability [2–9] are two of the major barriers to the application of pharmacogenomics findings in clinical practice. These problems exist because of the large number of SNPs in the genome that are not associated with the outcome. In addition, common diseases usually involve many SNPs with a small effect size at the single SNP level.

Recent advances in statistical methodology have improved the power to identify relevant SNPs and predict the outcome for a patient based on their SNPs. These methods include techniques from machine learning, such as random forests and support vector machines [10–18]. For example, Cosgun et al. [11] applied three machine learning approaches: Random Forest Regression (RFR), Boosted Regression Tree (BRT) and Support Vector Regression (SVR) to the prediction of warfarin maintenance dose in a cohort of African Americans [11]. They showed that even though all three methods achieved better performance than the previously published reports, RFR had the best accuracy.

Building a predictive model from genomic SNP data usually involves two steps. The first step, called feature selection, is to rank individual SNPs by their association with the outcome of interest and select the top SNPs. The ranking is usually done by fitting a simple logistic regression model or performing a trend test for a binary response, such as response/non-response to a drug treatment or presence/absence of an adverse event. The ranking can also be determined by fitting a generalized liner model for a continuous response, such as change from baseline for an efficacy measurement. The second step is to build a predictive model based on the selected SNPs [10]. In some cases, the method for building the predictive model performs feature selection as part of the fitting process (e.g. with elastic nets). With such fitting procedures, it may be possible to omit the prior feature selection step, and instead perform the analysis in a single step (e.g. a 1-step procedure).

Innovations in statistical procedures and methodologies could be helpful to understand and meet those challenges in predictive model building. The objectives of the current study are to compare these new methods. To do so, we developed three different simulations. In the first simulation, we compared five approaches: 1-step Elastic Net (EN), 2-step genome-wide association study (GWAS)+EN, 1-step Principal Component Regression (PCR), 1-step Random Forest (RF) and 1-step Support Vector Machine (SVM). For the second simulation, we compared three 2-step

procedures, GWAS+EN, GWAS+RF and GWAS+SVM. In the third simulation, we compared two cross validation approaches: GWAS+EN and a modified learn and confirm cross validated GWAS+EN (i.e. Modified CV GWAS+EN).

2 Materials and Methods

2.1 Introduction to the Statistical Methods

2.1.1 Univariate Association Analysis

For all 2-step approaches used in this study, the first step was a univariate linear association analysis to select the top SNPs in Genome-wide association study (GWAS). For each SNP, we fit the model

$$y_i = \beta x_i + \varepsilon_i$$

Here, y_i is the observed patient outcome for the ith patient, β is the coefficient for the SNP, and x_i is copy number of the SNP (0, 1, or 2) in the ith patient. ε_i is a normally distributed error term with mean 0. The null hypothesis was H_0: $\beta = 0$. We order SNPs by their P-Values and selected SNPs lower than a specified genome-wide significance level of approximate P-value threshold of $5 \times 10(-8)$.

2.1.2 Elastic Net (EN)

In the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods [19]. The decorrelation step leads to grouping effect and better prediction accuracy [19]. In addition, the decorrelation allows the fitting procedure to succeed even the number of SNPs (M) is greater than the number of patients (N).

We used EN in our simulations and the hyper-parameters associated with L1 and L2 penalties were trained using a five-fold cross validation external to the predictive model building process to avoid potential bias in estimated test errors [20]. In the 1-step approach, we directly applied EN to all the SNPs to build the predictive model. EN was used in both the one and two step simulations.

2.1.3 Random Forest (RF)

The Random Forest is a group of trees based on bootstrapped datasets [21]. We used the RF method in both the one and two step simulations. The out of bag (OOB) error

for each tree was computed based on samples not used in the bootstrapped dataset. First, we generated a variable importance list for all SNPs. We then iteratively fit the data with RF, each time building a new forest after discarding the lowest 30% of the SNPs used in the previous iteration. OOB error was computed for each iteration. Our final prediction model was the one with the smallest number of SNPs whose OOB error was within 1 standard error of the smallest OOB error of all forests.

2.1.4 Principal Component Regression (PCR)

Principal components regression (PCR) uses principal components analysis (PCA) to decompose the independent (x) variables into an orthogonal basis (the principal components), and select a subset of those components as the variables to predicty [22]. We only used PCR in an one step analysis in the first simulation. Here, the principal components (PC) were linear combinations of the SNPs. The principal components can be ranked by the amount of variance in the SNPs that each principal component explains. Only the top 3000 SNPs ranked by p-value were use are in the PCR method. In the one step PCR approach, we applied PCA on all M SNPs and pick the top k PCs. We then used the k PCs for prediction.

2.1.5 Support Vector Machine (SVM)

In machine learning, support vector machines (SVMs) are supervised learning models used for classification and regression analysis [23]. As part of the fitting procedure, SVM assigns a weight to each feature (which in our case were SNPs). We used a linear SVM to fit a predictive model for the patient outcomes. We computed the weight for each SNP and ordered the SNPs by the weight. We iteratively fit a SVM, each time discarding the lowest 10% of the SNPs used in the previous iteration. A 5-fold cross validation (CV) was used to get a CV error for each iteration. The final model chosen was the one with the smallest number of SNPs whose CV error was within 1 standard error of the smallest CV error of all SVMs. The SVM was used in both one and two step simulations.

2.1.6 Cross Validation

A cross validation procedure was used in our simulation. In the model building process, we have one sample of individuals (training sample) to "learn" the prediction model. We can use another independent sample of individuals (testing sample) to evaluate how well the prediction power (test error) is for our prediction model. Cross validation (CV) can be used to estimate the test error using the training sample. It's just a technique to assess the prediction performance, we still use the entire training sample to train your prediction model and we do not waste any data.

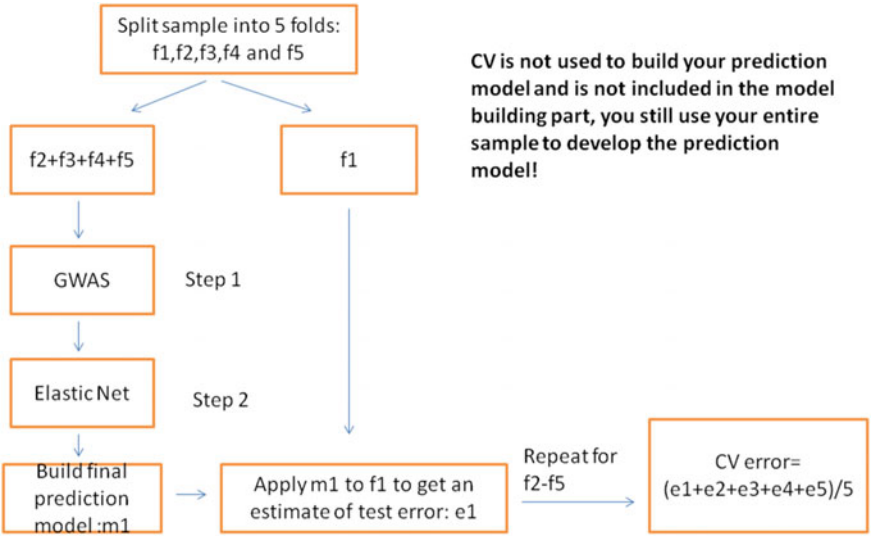


Fig. 1 Standard cross validation flow chart

A fivefold cross validation has been used in all simulations in this study. In our first and second simulations, we had following cross validation chart (Fig. 1):

2.2 Simulation One: One and Two Step Comparisons

In our first simulation study, we compared five approaches: 1-step elastic net (EN), 2-step genome-wide association study (GWAS)+EN, 1-step principal component regression (PCR), 1-step random forest (RF) and 1-step support vector machine (SVM). In all 1-step approaches, the EN, PCR, RF and SVM were directly used for both feature selection and predictive model building with all SNP variants.

2.2.1 Settings for Simulation One

For our first simulation, we used real SNP data from chromosome 1 from 535 patients (from a Takeda clinical study with PGx samples), as measured by the Illumina Human Omni5Exome array. This dataset contained 9,968 SNPs after QC. We randomly selected 5 SNPs to be causal variants and used them to generate a simulated phenotype y:

$$\mathbf{y} = \boldsymbol{\beta}_1\mathbf{x}_1 + \boldsymbol{\beta}_2\mathbf{x}_2 + \boldsymbol{\beta}_3\mathbf{x}_3 + \boldsymbol{\beta}_4\mathbf{x}_4 + \boldsymbol{\beta}_5\mathbf{x}_5 + \mathbf{e}$$

$$e \sim N(0, \sigma^2)$$

Since the simulation did not modify the SNP data, the original LD structure was maintained. The minor allele frequency (MAF) for the causal variants was 5, 7, 8, 9 and 16%. The coefficients for causal variants were set to 0.5, 0.75, 1, 1.25, and 1.5 and the variance of the noise term was set so that together the 5 associated variants together explained 20% of the total variance in y. Two hundred fifty such datasets were simulated and for each replicated dataset, 300 patients were randomly selected for training sample and the other 235 patients for testing sample.

The top 100 SNPs found by GWAS were used in the GWAS+EN analysis. For PCR, the top 25 PCs were used, which accounted for more than 99% of the total original variance. Five-fold CV was used to estimate the test error across 250 simulated datasets.

2.3 Simulation Two: 2-Step Strategy Comparisons

The second simulation compared three procedures, GWAS+EN, GWAS+RF and GWAS+SVM. The following settings were used.

2.3.1 Settings for Simulation Two and Simulation Three

Simulations two and three used the same model as above to generate y. However, in this case, the SNP values were simulated as well. In each replicated dataset, the number of total genotyped SNPs (M) was set to 10,000. The number of causal variants (m) was set to 5, all with MAF = 0.165. The coefficients for causal variants were set to 0.5, 0.75, 1, 1.25, and 1.5. As before, the variance of the noise term was set so that together, the SNPs explained 20% of the total variance. The remaining null markers were generated with MAF following a uniform distribution U(0.1, 0.4). The training sample size and testing sample size were both set to be 300. The top 100 SNPs found by GWAS were used in the two step methods. We selected the top 10 PCs for PCR. As before, we generated 250 simulated data sets and used 5 fold CV to estimate the test error. These settings were used in both simulations two and three.

2.4 Simulation Three: Additional Cross Validation Considerations

A more sophisticated "learn and confirm" strategy was compared in simulation three. The purpose is finding a better way to conduct cross validation by having an extra validation (confirm) on the top SNPs already identified from the first step GWAS

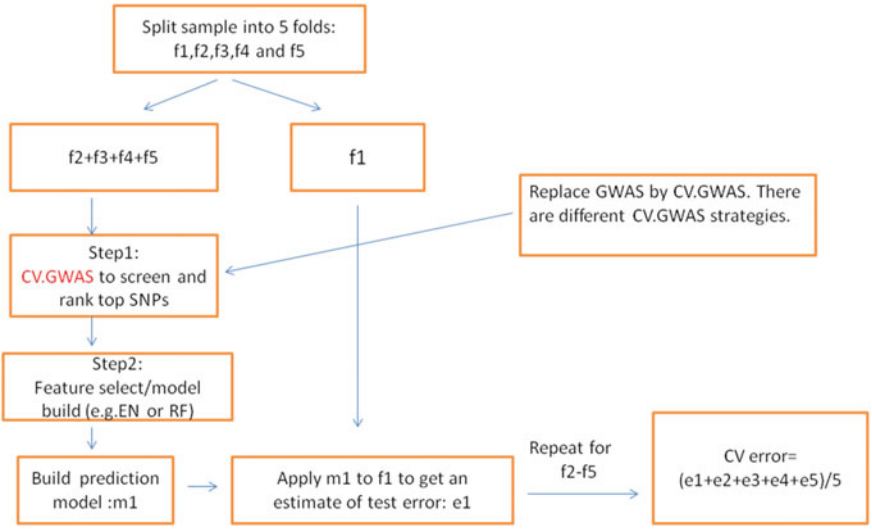


Fig. 2 Modified cross validation flow chart

(learn), before a second step EN on model building. We considered two cross validation approaches. 1. GWAS+EN (as shown in the Fig. 1) and 2. A modified CV GWAS+EN. We used cross validation along with GWAS to stable the feature selection for the top variants (Fig. 2).

3 Results and Discussion

3.1 Results and Discussion for Simulation One

As shown in Table 1, the comparison of the five approaches identified that the 1-step EN had the smallest test MSE (4.49) and the highest percent associated with the causal variants (0.14). However it also came up with a relatively higher training error (3.51). SVM had the highest sensitivity (0.74) and the smallest training error (0.04). Random forest approach had the second smallest test MSE (4.78). In addition, it had the second highest causal % (0.09). The 2-step approach (GWAS+CN) had the second highest sensitivity (0.64). The training error was biasedly down warded (underestimated) for SVM (0.04).

In Simulation One, with manageable number of SNP variants (<10,000) and sample size (535), one step methods, especially the EN and RF showed some advantages over the rest of the other methods. When the phenotype-genotype model is generated by a linear model: EN and RF had better prediction accuracy than GWAS+EN.

Table 1 Results of simulation one

Approaches	Test MSE	Sensitivity	Causal percent
EN	4.49	0.61	0.14
GWAS+EN	5.55	0.64	0.04
PCR	5.29	NA	NA
RF	4.78	0.51	0.09
SVM	5.41	0.74	0.01

Test MSE: mean squared error on testing sample
Sensitivity: number of causal SNPs in final set/number of causal SNPs
Causal Percent: number of causal SNPs in final set/number of selected SNPs in final set

GWAS+EN approach may preferentially select in associated variants with the price of bringing more noise than EN. The training error was biasedly down warded (underestimated) for SVM. The cross validation error was a good estimate of the true test error.

Cosgun et al. [11] applied three machine learning approaches: Random Forest Regression (RFR), Boosted Regression Tree (BRT) and Support Vector Regression (SVR) to the prediction of warfarin maintenance dose in a cohort of African Americans [11] and found R^2 between the predicted and actual square root of warfarin dose in this model was on average 66.4% for RFR, 57.8% for SVR and 56.9% for BRT. Thus RFR had the best accuracy. Our results were consistent with Cosgun's study and had confirmed that RF is one of the better methods in prediction model building.

3.2 Results and Discussion for Simulation Two

Table 2 shown the results from the second simulation. GWAS+Random Forest gives the best prediction accuracy among all 2-step strategies. GWAS+Random Forest tends to select SNPs with higher accuracy than the others with the smallest test MSE (7.51) and causal % (0.09). This results again confirmed findings from Cosgun et al. [11] and even in the 2-step model building procedure, GWAS+RF had better accuracy than other methods. On the other hand, the GWAS+EN had the highest sensitivity (0.65).

3.3 Results and Discussion on Simulation Three

The results of three procedure comparisons GWAS+EN and Modified CV GWAS+EN are shown in Table 3. The results showed that the Modified CV

Table 2 Results of simulation two on two stage approaches

Procedure	Test MSE	Sensitivity	Causal percent
GWAS+EN	8.78	0.65	0.04
GWAS+RF	7.51	0.52	0.09
GWAS+SVM	10.02	0.48	0.05

Test MSE: mean squared error on testing sample
Sensitivity: number of causal SNPs in final set/number of causal SNPs
Causal Percent: number of causal SNPs in final set/number of selected SNPs in final set

Table 3 Results of simulation three on different cross validation considerations

Procedure	Test MSE	Sensitivity	Causal percent
GWAS+EN	8.79	0.66	0.04
Modified CV GWAS+EN	8.12	0.51	0.04

Test MSE: mean squared error on testing sample
Sensitivity: number of causal SNPs in final set/number of causal SNPs
Causal Percent: number of causal SNPs in final set/number of selected SNPs in final set

GWAS+EN have better prediction accuracy than GWAS+EN (MSE of 8.12 vs. 8.79), and modest training error as well (3.1 vs. 1.42)). Nevertheless, it came with more computational burden. GWAS +EN had higher sensitivity than Modified CV GWAS+EN (0.66 vs. 0.51).

The GWAS+EN procedure was a standard one (as shown in the Fig. 1). The difference for the Modified CV GWAS+EN was that we used a "learn and confirm" cross validation procedure along with GWAS to stable the feature selection for the top variants (Fig. 2). The "learn and confirm" procedure was with an additional confirmation step on the selected top ranked SNPs in a different data set before building up the models. This strategy would be very similar to the model building procedure by Shigemizu et al. [12] with real type 2 diabetes data [12], in which an extra validation on the top identified SNPs was implemented before predictive model building. We recommend this procedure as it came with higher accuracy and gave additional stability on the SNPs for the predictive model building.

4 Conclusions

In our simulation of genotype-phenotype association:

1. One step EN showed better prediction accuracy than GWAS+EN.
2. GWAS+EN identified more total causal SNPs than EN, but the portion of causal SNPs among the identified SNPs was lower.

3. GWAS+RF gave the highest prediction accuracy among all two-step strategies.
4. Modified CV GWAS+EN (learn and confirm) had better prediction accuracy than GWAS+EN, with the burden of extra computational cost.

Acknowledgements Useful discussions with Dr. Zheng Zha and reviews by Dr. Yu-chen Su at Takeda Pharmaceutical Develop Center are highly appreciated.

Conflict of Interest The project was carried out while Dr. Pingye Zhang was a summer intern at Takeda develop center at Deerfield, IL. USA. All other authors were Takeda employees at the time. The nature of the research is comparison of statistical methodologies and cross validation procedures, there is no conflict of interests.

References

1. Schilsky, R.L.: Personalized medicine in oncology: the future is now. Nat. Rev. Drug. Discov. **9**(5), 363–366 (2010)
2. Schrodi, S.J., Mukherjee, S., Shan, Y., Tromp, G., Sninsky, J.J., Callear, A.P., et al.: Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. Front. Genet. **5** Article162. 2 (2014)
3. Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., Visscher, P.M.: Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. **14**(7), 507–515. https://doi.org/10.1038/nrg3457 (2013)
4. Lee, S.H., Wray, N.R., Goddard, M.E., Visscher, P.M.: Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. **88**, 294–305 (2011)
5. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., et al.: Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. **42**, 565–569 (2010)
6. Visscher, P.M., Yang, J., Goddard, M.E.: A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. Twin Res. Hum. Genet. **13**, 517–524 (2010)
7. Pang, G.S.Y., Wang, J., Wang, Z., Lee, C.G.L.: Predicting potentially functional SNPs in drug-response genes. Phamacogenomics **10**(4), 639–653 (2009)
8. Francis Lam, Y.W.: Scientific challenges and implementation barriers to translation of Pharmacogenomics in clinical practice. ISRN Pharm. Article ID 641089 (2013)
9. Lee, S.H., Wray, N.R., Goddard, M.E., Visscher, P.M.: Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. **88**(3), 294–305 (2011)
10. Nguyen, T.-T., Huang, J.Z., Wu, Q., Nguyen Mark, T.T., Li, J.: Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. BMC Genomics **16**(Suppl 2), S5 (2015)
11. Cosgun, E., Limdi, N.A., Duarte, C.W.: High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. Bioinformatics **27**(10), 1384–1389 (2011)
12. Shigemizu, D., Abe, T., Morizono, T., Johnson, T.A., Boroevich, K.A., Hirakaw, Y., et al.: The Construction of risk prediction models using G**W**AS data and its application to a Type 2 diabetes prospective cohort. PLoS ONE **9**(3), e9254 (2014)
13. Kooperberg, C., LeBlanc, M., Obenchain, V.: Risk prediction using genome-wide association studies. Genet Epidemiol. **34**(7), 643–652 (2010)
14. Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., et al.: Large sample size, wide variant advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am. J. Hum. Genet. **92**, 1008–1012 (2013)

15. Chen, X., Ishwaran, H.: Random forests for genomic data analysis. Genomics **99**, 323–329 (2012)
16. Schrijver, I., Aziz, N., Farkas, D.H., Furtado, M., Gonzalez, A.F., Greiner, T.C., et al.: Opportunities and challenges associated with clinical diagnostic genome sequencing. J. Mol. Diagn. **14**(6) (2012)
17. Cantor, R.M., Lange, K., Sinsheimer, J.S.: Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am. J. Hum. Genet. **86**, 6–22 (2010)
18. Li, L., Guennel, T., Marshall, S.L., Cheung, L.W.K.: A multi-marker molecular signature approach for treatment-specific subgroup identification with survival outcomes. Pharmacogen. J. **14**(5), 439–445 (2014)
19. Zou, H., Trevor, T.: Regularization and variable selection via the elastic net. J. Royal Stat. Soc. Ser. B **67**(2), 301–320 (2005)
20. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. PNAS **99**(10), 6562–6566 (2002)
21. Tin Kam, H.O.: Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August, pp. 278–282 (1995)
22. Jolliffe, I.T.: A note on the use of principal components in regression. J. Royal Stat. Soc. Ser. C. **31**(3), 300–303 (1982)
23. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)