Trends in Mathematics
Research Perspectives CRM Barcelona Vol.7

Elizabeth A. Ainsbury
M.Luz Calle
Elisabeth Cardis
Jochen Einbeck
Guadalupe Gómez
Pere Puig
Editors

Extended Abstracts Fall 2015

Biomedical Big Data

Statistics for Low Dose Radiation Research

Birkhäuser

Trends in Mathematics

Research Perspectives CRM Barcelona

Volume 7

Series editors

Enric Ventura
Antoni Guillamon

Since 1984 the Centre de Recerca Matemàtica (CRM) has been organizing scientific events such as conferences or workshops which span a wide range of cutting-edge topics in mathematics and present outstanding new results. In the fall of 2012, the CRM decided to publish extended conference abstracts originating from scientific events hosted at the center. The aim of this initiative is to quickly communicate new achievements, contribute to a fluent update of the state of the art, and enhance the scientific benefit of the CRM meetings. The extended abstracts are published in the subseries Research Perspectives CRM Barcelona within the Trends in Mathematics series. Volumes in the subseries will include a collection of revised written versions of the communications, grouped by events.

More information about this series at http://www.springer.com/series/13332

Elizabeth A. Ainsbury · M.Luz Calle
Elisabeth Cardis · Jochen Einbeck
Guadalupe Gómez · Pere Puig
Editors

Extended Abstracts Fall 2015

Biomedical Big Data

Guadalupe Gómez
Pere Puig
M.Luz Calle
Editors

Statistics for Low Dose Radiation Research

Elizabeth A. Ainsbury
Elisabeth Cardis
Pere Puig
Jochen Einbeck
Editors

Birkhäuser

Editors
Elizabeth A. Ainsbury
Chemical and Environmental Hazards
Public Health England
Chilton
UK

Jochen Einbeck
Mathematical Sciences
Durham University
Durham
UK

M.Luz Calle
Departament de Biociències
Universitat Central de Catalunya
Vic
Spain

Guadalupe Gómez
Estadística i Investigació Operativa
Universitat Politècnica de Catalunya
Barcelona
Spain

Elisabeth Cardis
Campus Mar
ISGlobal
Barcelona
Spain

Pere Puig
Departament de Matemàtiques
Universitat Autònoma de Barcelona
Barcelona
Spain

ISSN 2297-0215 ISSN 2297-024X (electronic)
Trends in Mathematics
ISSN 2509-7407 ISSN 2509-7415 (electronic)
Research Perspectives CRM Barcelona
ISBN 978-3-319-55638-3 ISBN 978-3-319-55639-0 (eBook)
DOI 10.1007/978-3-319-55639-0

Library of Congress Control Number: 2017936021

Mathematics Subject Classification (2010): 62M10, 62N01, 62P10, 92B15, 92C60, 92D30

Printed on acid-free paper

This book is published under the trade name Birkhäuser, www.birkhauser-science.com
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part I
Biomedical Big Data

Foreword

In the last quarter of 2015, from September 8 to November 27, over 100 biostatisticians, statisticians and mathematicians from 45 different institutions visited the Centre de Recerca Matemàtica (CRM) in Bellaterra to participate in the Intensive Research Programme on Statistical Advances for Complex Data. The local organizers of this research semester were Alejandra Cabaña (Universitat Autònoma de Barcelona), Malu Calle (Universitat de Vic), Pedro Delicado (Universitat Politècnica de Catalunya), Anna Espinal (Universitat Autonòma de Barcelona), Guadalupe Gómez (Universitat Politècnica de Catalunya), Rosa Lamarca (Almirall SA), Pere Puig (Universitat Autonòma de Barcelona), Montserrat Rué (Universitat de Lleida), and Àlex Sánchez (Universitat de Barcelona). The program brought together scientists, from enthusiastic Ph.D. students to respected senior professors, working in relevant areas such as Modeling and analysis of biological and biomedical data, Biostatistical methods for clinical trials and for complex time-to-event data, and Statistics and Big Data. The very dynamic and productive atmosphere we enjoyed translated into equally active courses, seminars and a workshop on Biomedical (Big) Data, held on November 26 and 27, closing the program.

The workshop was a meeting point for the researchers who are members of BIOSTATNET, a Spanish pioneer network of biostatisticians. BIOSTATNET has almost two hundred members organized around eight different nodes, led by statisticians from different universities, with own research projects and teaching experience in biostatistical matters, and working closely with biomedical researchers. The workshop included five invited talks, a roundtable, eleven contributed oral presentations and ten posters.

In this volume of the subseries Research Perspectives CRM-Barcelona (published by Birkhäuser inside the series Trends in Mathematics), we present ten extended abstracts corresponding to selected talks given by participants in the workshop on Biomedical (Big) Data. The variety of topics presented bears testimony to the rich activity that made a success of the workshop, and also of the Intensive Research Programme. The selected topics include methodological biostatistical and bioinfor-

matics advances as well as relevant medical applications. Five abstracts contribute with new procedures and methods for high-dimensional data sets: from integrative analysis of omics data to statistical models for microbiome data. In three papers different approaches for the joint model between longitudinal markers and time to events are the main contribution. Sample size considerations in clinical prediction and models for under-reported time series count data are other topics addressed among the authors. As is usual in our field, most of the methodological progress comes from the hand of relevant scientific questions in the medical and biological field. Among our abstracts, we find the problems that have arisen in HIV studies where the evolution of bone mineral density measurements or the characterization of the microbiome composition of HIV-infected persons is of interest; three papers discuss applications in cancer studies focusing on the selection of genes in a cancer colon study, analyzing gene expression in ovarian cancer, or assessing breast cancer risk from the longitudinal mammographic breast densities. Other clinical studies analyzing autism multiplex families, characterizing brain tissue after ischemic stroke, or finding the association between post-operative glucose proles and insulin therapy on patients survival after an orthotopic liver transplantation are part of the problems addressed from statistical and computational perspectives. We hope that this volume will give the authors the opportunity to quickly communicate their recent research: most of the short articles here are brief and preliminary presentations of new results not yet published in regular research journals.

We would like to express our gratitude to the CRM for hosting and supporting our research program. Also our warm thanks to the CRM staff, its director, Joaquim Bruna, and all the secretaries, for providing great facilities and a very pleasant working environment. Last but not least, thanks are due to all those who attended the talks, for their interest, enthusiasm and their active participation. The program was also possible thanks to the generous support of the following research projects from the Ministerio de Economía y Competitividad (Spain): “Applied Stochastic Processes”, conducted at Universitat Autònoma de Barcelona (MTM2012-31118), “Advanced Methods in Survival Analysis: Clinical Trials, Longitudinal Data and Interval Censoring”, conducted at Universitat Politècnica de Catalunya (MTM2012-38067-C02-01), “Sampling Samples: Relevant Applications of Statistics in Digital Economy and Society”, conducted at Universitat Politècnica de Catalunya (MTM2013-43992-R), and the following one from the Departament d’Economia i Coneixement (Catalunya): “Research Group in Biostatistics and Bioinformatics” at Universitat Politècnica de Catalunya and Universitat de Barcelona (SGR 464), as well as the Simons Foundation, and the Fundación Española para la Ciencia y la Tecnología (Ministerio de Economía y Competitividad, Spain).

July 2016
Barcelona, Spain

Guadalupe Gómez
Pere Puig
M.Luz Calle

Extreme Observations in Biomedical Data

Concepción Arenas, Itziar Irigoien, Francesc Mestres, Claudio Toma, and Bru Cormand

Abstract We present a new procedure to detect extreme observations which can be applied to low or high-dimensional data sets. Continuous features, a known underlying distribution or parameter estimations are not required. The procedure offers a ranking by assigning a value to each observation that reflects its degree of outlyingness. A short computation time is needed.

1 Introduction

In current biomedical research, genetic studies are extensively used to identify the causes of human diseases and they provide insights for the eventual development of therapeutic strategies. Integration of different types of data sets, such as gene expression data, genotype data or clinical information is needed to capture information that may otherwise be lost in separate analyses. Furthermore, it is crucial to be able to detect extreme observations, since an extreme value may indicate an individual with a wrong diagnosis or presenting particular clinical features or classified in the extreme spectrum of the disease. Moreover, the usual scenario with current data is the lack of

C. Arenas (✉)
Department of Statistics, University of Barcelona, Barcelona, Spain
e-mail: carenas@ub.edu

I. Irigoien
Department of Computer Sciences and Artificial Intelligence, University of the Basque Country, Leioa, Spain
e-mail: itziar.irigoien@ehu.es

F. Mestres · B. Cormand
Department of Genetics, University of Barcelona, Barcelona, Spain
e-mail: fmestres@ub.edu

B. Cormand
e-mail: bcormand@ub.edu

C. Toma
Neuroscience Research Australia, Sydney, NSW, Australia
e-mail: c.toma@neura.edu.au

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_1

information about the underlying distribution. Thus, no parametric extreme observation detection algorithms for any type of features and for any size/dimensional data sets are desirable. We present a new procedure to detect extreme observations which can be applied to low or high-dimensional data sets. Continuous features, a known underlying distribution or parameter estimations are not required. The procedure offers, using a short computation time, a ranking by assigning to each observation a value that reflects its degree of outlyingness. The proposed method takes into account all distances between observations, not only distances between neighbours, in such a way that the relation of any observation with respect to all the other observations in the data set and the dispersion of all data are considered. To illustrate our procedure, we analyzed the data of rare genetic variants from 10 autism multiplex families and twenty-six high-dimensional class-imbalanced cancer data sets. The results showed the good performance of the procedure.

2 Methods

The starting point is an $n \times p$ data matrix (p can be much larger than the size of the sample n) where the rows correspond to observations (individuals, samples...) and the columns correspond to any kind of features to be measured which can be continuous, binary or multiattribute data (genes, clinical/pathological features,...). Let G be a group that is represented by a p-random vector $\mathbf{Y} = (Y_1, \ldots, Y_p)$, with values in a metric space $S \subset R^p$ and a probability density f with respect to a suitable measure λ. Let δ be a distance function between any pair of observations, $\delta_{ij} = \delta(\mathbf{y}_i, \mathbf{y}_j)$.

Definition 1 The geometric variability of G with respect to δ, a general measure of dispersion of G, is defined by

$$V(G) = \frac{1}{2} \int_{S \times S} \delta^2(\mathbf{y}_i, \mathbf{y}_j) f(\mathbf{y}_i) f(\mathbf{y}_j) \lambda(d\mathbf{y}_i) \lambda(d\mathbf{y}_j);$$

see [1]. When δ is the Euclidean distance, $V(G) = tr(\Sigma)$ with $\Sigma = cov(\mathbf{Y})$. The geometric variability is as a variant of Rao's diversity coefficient; see [2].

Definition 2 The proximity function of an observation $\mathbf{y}$ to G is defined by

$$\phi^2(\mathbf{y}, G) = \int_S \delta^2(\mathbf{y}, \mathbf{y}_j) f(\mathbf{y}_j) \lambda(d\mathbf{y}_j) - V(G);$$

see [1].

As in applied problems, the probability distribution for $\mathbf{Y}$ is usually unknown, estimators are needed. Given a sample of size n, $\mathbf{y}_1, \ldots, \mathbf{y}_n$, natural estimators for the geometric variability and the proximity function are

$$\hat{V}(G) = \frac{1}{2n^2} \sum_{i,j} \delta^2(\mathbf{y}_i, \mathbf{y}_j),$$

and

$$\hat{\phi}^2(\mathbf{y}, G) = \frac{1}{n} \sum_{i} \delta^2(\mathbf{y}, \mathbf{y}_i) - \hat{V}(G),$$

respectively. See [3] for a review of these concepts, and for applications see [4, 5] and references therein.

Definition 3 For each observation $\mathbf{y}_i$, the depth function $I(\mathbf{y}_i, G)$ is defined by

$$I(\mathbf{y}_i, G) = \left[1 + \frac{\phi^2(\mathbf{y}_i, G)}{V(G)}\right]^{-1}; \tag{1}$$

see [6].

Proposition 4 *Function I takes values in* $[0, 1]$ *and, according to [7], it is a type C depth function. Furthermore, it verifies the following desirable properties: (i) maximality at center; (ii) monotonicity relative to the deepest observation; (iii) vanishing at infinity; and (iv) depending on the data and the selected distance, it is affine-invariant.*

As I is a depth function, it assigns to any observation a degree of centrality, thus a small value of I, or equivalently a large value of $O = 1/I$, suggests a possible extreme observation. Note that, by (1), $\hat{O}(\mathbf{y}_i, G) = n \sum_j \delta^2(\mathbf{y}_i, \mathbf{y}_j) / \sum_{j,k} \delta^2(\mathbf{y}_j, \mathbf{y}_k)$.

However, with only one observation taking a very large value, $\hat{O}$ already gives aberrant values. For this reason, we propose the following version for $\hat{O}(\mathbf{y}_i, G)$ where, due to robustness consideration, the mean is replaced by the median.

Definition 5 For each observation $\mathbf{y}_i$ a new statistic $O_R(\mathbf{y}_i, G)$ is defined by

$$O_R(\mathbf{y}_i, G) = \frac{med_{\delta,i}}{med_{\delta}}, \tag{2}$$

where $med_{\delta} = median_{j,k}(\delta^2_{jk})$ and $med_{\delta,i} = median_j(\delta^2_{ij})$.

Proposition 6 *Let $S = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be a sample, and let $\mathbf{y}_0$ be an outlier. For a fixed observation, say $\mathbf{y}_1$, the sensitivity curve of $O_R(\mathbf{y}_1, S)$ at point $\mathbf{y}_0$, $SC(\mathbf{y}_0) = O_R(\mathbf{y}_1, S') - O_R(\mathbf{y}_1, S)$, where $S' = \{\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{y}_0\}$, is bounded, which implies the robustness of O_R.*

Proposition 7 *Let δ be a distance function such that $\delta(\mathbf{y}_i, \mathbf{y}_j) \to \infty$ when $\mathbf{y}_i$ or $\mathbf{y}_j$ takes arbitrarily large values. The breakdown point of O_R (the proportion of arbitrarily large observations that O_R can handle before giving arbitrarily large values) is $n - 1/2 - \sqrt{2n^2 - 6n + 1}/2$, with n the sample size. Note that the breakdown point of O_R is always greater than 25%.*

Note that the distribution of O_R is not symmetric.

Definition 8 Following Kimber criterion (see [8]), an observation $\mathbf{y}_i$ will be consider as an extreme observation if

$$O_R(\mathbf{y}_i) > \lambda = Q_3 + 1.5(Q_3 - M), \tag{3}$$

where Q_3 and M are the 3-th quartile and the median of all the O_R values.

Our simulation studies show that the procedure is robust in front of masking effect and it can properly identify most of the outliers when mixed data are analyzed.

3 Application to Data in Autism Multiplex Families

Now consider the following study [9] in which 10 autism multiplex families were analyzed (nine with two affected sibs and one with three affected sibs). First, in a clinical study, five features were measured in 21 affected individuals: two were continuous (age and non-verbal intelligence quotient(NVIQ)), and three were categorical (gender, language delay and autism spectrum category). Using (3) and the Gower distance [10], the threshold value was $\lambda = 1.519$, and four individuals could be considered as extreme observations. Three of them were male (13, 17 and 20 years old) with autism and language delay, and they presented NVIQ values indicative of mental retardation. The most emblematic extreme value, corresponded to another man (25 years old) also with an autism diagnosis and language delay, and presenting the smallest NVIQ value. Thus, our method highlighted the four individuals from our study that had the most severe clinical presentation of the disorder. In a second study, a genetic analysis was performed in the 21 affected individuals and in their parents. The full exome sequence (the fraction of the genome that encodes proteins, approximately 3.4×10^7 nucleotide positions from 20,000 genes) of all family members was determined. We selected those rare genetic variants (infrequent in the general population) leading to an amino acid change in the encoded protein that were transmitted from one parent to the two (or three) affected sibs. The identified mutations, an average of 36.3 per family, were ranked according to their predicted damaging effect using the SIFT and PolyPhen-2 tools. In this case, no extreme observations were detected. This result is consistent with the fact that this type of mutation may not have a major role in the aetiology of the disorder (as compared to mutations leading to truncated proteins, not considered here) in the sample of multiplex families reported previously in [9].

4 Application to Gene Expression Data in Cancer

In biomedical studies, an important task is to select informative genes that present altered expression levels in the diseases under study. Selecting adequate marker genes should may be useful in classifying new samples. We considered 26 public microarray cancer data sets (http://bioinformatics.rutgers.edu/Static/Supplements/CompCancer/datasets.htm), which are high size/dimensional class-unbalanced data sets. We compared the results of a linear discriminant analysis using the original set of genes and the extreme genes identified under criterion (3). As an evaluation criteria, we considered the rate of correct classification obtained by the leave-one-out method. Using only the extreme genes detected by (3) the rate of correct classification was, in general, maintained or even improved (see Table 1). It is important to note that the reduction in the number of marker genes facilitates the interpretation of their biological meaning with regard to the disease.

Table 1 Columns: Cancer data sets; classes (k); samples (n); original genes (p); extreme genes selected by our criterion (NG); total leave-one-out classification rate, in percentage, using all genes (CR_{all}) and using the reduced list of genes (CR_{sel})

Data set	k	n	p	NG	CR_{all}	CR_{sel}	Data set	k	n	p	NG	CR_{all}	CR_{sel}
Alizadeh-2000-v1	2	42	1095	118	90.48	92.86	Laiho-2007	2	37	2202	414	81.08	86.49
Alizadeh-2000-v2	3	62	2093	306	98.39	98.39	Lapointe-2004-v1	3	69	1625	170	81.16	72.46
Armstrong-2002-v1	2	72	1081	193	91.67	98.61	Lapointe-2004-v2	4	110	2496	249	80.91	70.00
Armstrong-2002-v2	3	72	2194	391	88.89	91.67	Liang-2005	3	37	1411	179	94.59	100.00
Bittner-2000-V1	2	38	2201	279	76.32	84.21	Nutt-2003-v1	4	50	1377	320	72.00	70.00
Bittner-2000-V2	3	38	2201	279	63.16	65.79	Nutt-2003-v2	2	28	1070	173	89.29	100.00
Bredel-2005	3	50	1739	238	84.00	84.00	Nutt-2003-v3	2	22	1152	246	100.00	90.91
Dyrskjot-2003	3	40	1203	217	75.00	82.50	Pomeroy-2002-v1	2	34	857	126	76.47	79.41
Garber-2001	4	66	4553	391	81.82	83.33	Risinger-2003	4	42	1771	255	71.43	71.43
Golub-1999-v1	2	72	1877	321	88.89	90.28	Shipp-2002-v1	2	77	798	137	85.71	75.32
Golub-1999-v2	3	72	1877	321	84.72	88.89	Tomlins-2006-v2	4	92	1288	129	83.70	84.78
Gordon-2002	2	181	1626	290	100.00	96.69	West-2001	2	49	1198	180	75.51	75.51
Khan-2001	4	83	1069	165	53.01	65.06	Yeoh-2002-v1	2	248	2526	315	87.90	95.97

Acknowledgements This research was partially supported by Grant 2014 SGR 464 (GRBIO) from the Departament d'Economia i Coneixement de la Generalitat de Catalunya; by the Basque Government Research Team Grant (IT313-10) SAIOTEK Project SA- 2013/00397; and by the University of the Basque Country UPV/EHU (Grant UFI11/45 (BAILab)).

References

1. C. Arenas and C.M. Cuadras, "Some recent statistical methods based on distances", *Cont. Sci.* **2** (2002), 183–191.
2. C.M. Cuadras and J. Fortiana, "A continuous metric scaling solution for a random variable", *J. Multivariate Ana.* **32** (1995), 1–14.
3. J.C. Gower, "A general coefficient of similarity and some of its properties", *Biometrics* **27** (1971), 857–871.
4. I. Irigoien and C. Arenas, "INCA: new statistics for estimating the number of clusters and identifying atypical units", *Stat. Med.* **27**, (2008), 2948–2973.
5. I. Irigoien, F. Mestres, and C. Arenas, "The depth problem: identifying the most representative units in a data group", *IEEE ACM T Comput Bi* **10** (2013), 161–172.
6. I. Irigoien, B. Sierra, and C. Arenas, "ICGE: an R package for detecting relevant clusters and atypical units in gene expression", *BMC Bioinformatics* **13** (2013), 30–41.
7. A.C. Kimber, "Exploratory data analysis for possibly censored data from skewed distributions", *Appl. Stat.* **39** (1990), 21–30.
8. C.R. Rao, "Diversity: its measurement decomposition apportionment and analysis", *Sankhya Indian J. Stat.* **44** (1982), 1–22.
9. C. Toma, B. Torrico, A. Hervs, R. Valdés-Mas, A. Tristán-Noguero, V. Padillo, M. Maristany, M. Salgado, C. Arenas, X.S. Puente, M. Bayés, and B. Cormand, "Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations", *Mol Psychiatr* **19** (2014), 784–790.
10. R. Serfling and S. Zuo, "General notions of statistical depth function", *Ann. Stat.* **28** (2000), 461–482.

An Ordinal Joint Model for Breast Cancer

Carmen Armero, Carles Forné, Montserrat Rué, Anabel Forte, Hector Perpiñán, Guadalupe Gómez and Marisa Baré

Abstract We propose a Bayesian joint model to analyze the association between longitudinal measurements of an ordinal marker and time to a relevant event. The longitudinal process is defined in terms of a proportional-odds cumulative logit model and the time-to-event process through a left-truncated Cox proportional hazards model with information of the longitudinal marker and baseline covariates. Both longitudinal and survival processes are connected by a common vector of random effects.

C. Armero (✉) · A. Forte · H. Perpiñán
Department of Statistics and Operational Research, Universitat de València, València, Spain
e-mail: carmen.armero@uv.es

A. Forte
e-mail: anabel.forte@uv.es

H. Perpiñán
e-mail: hector.perpinan@uv.es

C. Forné
Department of Basic Medical Sciences, Universitat de Lleida, IRB-Lleida and Oblikue Consulting, Lleida, Spain
e-mail: carles.forne@gmail.com

M. Rué
Department of Basic Medical Sciences, Universitat de Lleida, IRB-Lleida, Lleida, Spain
e-mail: montse.rue@cmb.udl.cat

H. Perpiñán
Fundación para el Fomento de la Investigación Sanitaria y Biomédica (FISABIO), Generalitat Valenciana, València, Spain

G. Gómez
Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain
e-mail: lupe.gomez@upc.edu

M. Baré
Clinical Epidemiology and Cancer Screening, Corporació Sanitària Parc Taulí-UAB, Sabadell, Spain
e-mail: mbare@tauli.cat

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_2

1 Introduction

Joint modeling of longitudinal and time-to-event data is an increasing area of statistical research devoted to jointly analyze longitudinal and survival processes. It enhances longitudinal modeling by allowing for the inclusion of non-ignorable dropout mechanisms, and survival modeling by the inclusion of internal time-dependent covariates. Shared-parameter models are joint models connecting the longitudinal and time-to-event processes by means of common subject-specific random effects which, in the presence of covariates and parameters, endow both processes with conditional independence; see [4]. They can quantify both the population and individual effects of the longitudinal outcomes on the risk of an event, and obtain individualized dynamic predictions.

When longitudinal outcomes are ordinal, the non-linear nature of the data produce a complex likelihood function which is difficult to maximize under the frequentist paradigm. This paper discusses a Bayesian joint model for the association between longitudinal measures of an ordinal marker and a time-to-event outcome; see [1] for more details about the model. We propose a proportional-odds cumulative logit model [3] for the ordinal measurements based on the idea of a continuous latent variable, and a Cox proportional hazards model with left truncation for the time-to-event of interest with information of the longitudinal process. The model is applied to estimate the risk of breast cancer in women attending a population-based screening program with regard to longitudinal measurements of mammographic breast density.

2 A Bayesian Joint Model for Ordinal Longitudinal and Left Truncated Survival Data

Let $\{D_1, \ldots, D_K\}$ be the set of ordinal categories and y_{ij} the category of individual i, $i = 1, \ldots, n$, at time t_{ij}, $j = 1, \ldots, n_i$. We assume an underlying continuous latent variable y_{ij}^* that determines the ordinal category of individual i at time t_{ij}. This latent variable has no interest *per se* but it is useful for motivating and interpreting the longitudinal model. The relationship between y_{ij} and y_{ij}^* is the following

$$y_{ij} = D_k \;\Leftrightarrow y_{ij}^* \in (\gamma_{k-1},\, \gamma_k],\;\; k = 1, \ldots, K,$$

where $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_K = \infty$ are unknown cutpoints. We choose a logistic distribution for y_{ij}^*, $\text{Lo}(m_{ij}, s = 1)$, with mean m_{ij} and scale parameter $s = 1$. The choice of that distribution implies a logit link for the cumulative probabilities as follows

$$q_{ijk} = P(y_{ij} > D_k) = P(y_{ij}^* > \gamma_k) = \frac{1}{1 + \exp(\gamma_k - m_{ij})}. \tag{1}$$

Despite $s = 1$ in the logistic distribution, the model is overparameterized. To obtain an identifiable model, we arbitrarily introduced a reference point on the latent scale, in particular $\gamma_{K/2} = 0$ if K is even and $\gamma_{(K-1)/2} = 0$ or $\gamma_{(K+1)/2} = 0$ if K is odd.

We consider a mixed-effects model to describe the subject-specific time trajectories of the latent variable

$$y_{ij}^* = m_{ij} + \epsilon_{ij} = \boldsymbol{x}_{ij}^{(l)\prime}\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{b}_i + \epsilon_{ij}, \tag{2}$$

where $\boldsymbol{x}_{ij}^{(l)}$ is a vector of covariates associated to individual i at time t_{ij} with regression coefficients vector $\boldsymbol{\beta}$; $\boldsymbol{z}_i$ a vector of explanatory variables attached to the random effects $\boldsymbol{b}_i$ for the i-th individual; and ϵ_{ij} an error term for the i-th individual at time t_{ij}, modeled in terms of the logistic distribution Lo(0, 1). The random effects $\boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)^T$ are conditionally i.i.d. $(\boldsymbol{b}_i \mid \boldsymbol{\phi}) \sim f(\boldsymbol{b}_i \mid \boldsymbol{\phi})$, where $f(\boldsymbol{b}_i \mid \boldsymbol{\phi})$ is usually taken to be a Multivariate Normal distribution with mean 0 and unknown covariance matrix.

Let T_i, $i = 1, \ldots, n$, be the observed event time for the i-th subject, obtained as the minimum between the true failure time, T_i^*, and the right-censoring time, C_i, $T_i = \min(T_i^*, C_i)$. The event indicator $\delta_i = I(T_i^* \leq C_i)$ takes the value 1 if the observed time is a true event time, and 0 otherwise. Event times corresponding to individuals who enter the study at delayed entry times introduce left-truncation. We define the hazard function of T_i^* in terms of the left-truncated Cox proportional hazard model [5]

$$h_i(t) = h_0(t \mid \boldsymbol{\lambda}) \exp\{\boldsymbol{x}_i^{(s)\prime}\boldsymbol{\eta} + \alpha\, m_{it}\}, \;\; t > a_i, \tag{3}$$

and zero otherwise, where $h_0(t \mid \boldsymbol{\lambda})$ is the baseline risk function with parameters $\boldsymbol{\lambda}$; $\boldsymbol{x}_i^{(s)}$ is the vector of baseline covariates with coefficients $\boldsymbol{\eta}$; α assesses the effect of the longitudinal marker of subject i on the event of interest in terms of the latent variable mean; and a_i is the delayed entry time of individual i.

We complete the Bayesian modeling eliciting a prior distribution, $\pi(\boldsymbol{\theta})$, for all the unknown parameters and hyperparameters of the model. From a Bayesian perspective, $\pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D})$, where $\mathcal{D}$ represents all the data collected from the longitudinal and the survival processes, is the joint posterior distribution of the parameters, hyperparameters, and random effects, which can be obtained by hierarchical modeling.

3 Breast Cancer and Mammographic Breast Density

The joint model is applied to the assessment of breast cancer risk in women attending a population-based screening program including 13760 women aged 50–69 years who participated in the breast cancer early-detection program in the Vallès Occidental Est area in Catalonia (Spain), between October 1995 and June 1998; see [2].

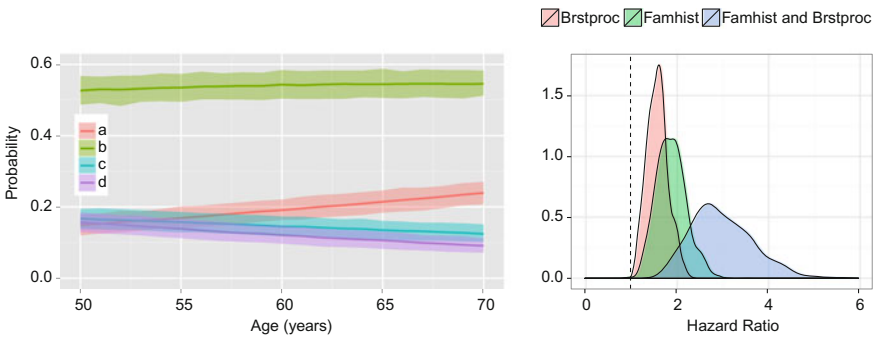


Fig. 1 Posterior mean and 95% credible band of the probability associated to each BI-RADS category with respect to age (*left*) and posterior distribution of the hazard ratios associated to family history of breast cancer, prior breast procedures, and both together (*right*)

The longitudinal ordinal marker is mammographic breast density in the scale BI-RADS, with a total of 81621 screening exams. The BI-RADS scale is ordinal with categories $\{a, b, c, d\}$, which represent low, medium, high, and very high breast density. The survival process focuses on time to a breast cancer diagnosis and incorporates *family history of breast cancer (Famhist)* and *prior breast procedures (Brstproc)* as dychotomous baseline covariates.

The posterior distribution is computed using Markov Chain Monte Carlo (MCMC) methods through the JAGS software. In particular, we run three MCMC chains with 100000 iterations, 10000 of which were used for the burn-in period. The chains were thinned by only storing every 270-th iteration to reduce autocorrelation in the saved sample. Convergence was assessed through the potential scale reduction factor and the effective number of independent simulation draws.

Figure 1 on the left shows the posterior mean and 95% credible interval of the posterior distribution associated to each BI-RADS category for a generic woman in the study. Probabilities associated to category b are always higher than 0.5, and grow slightly with age. Probabilities for categories a, c, and d are initially very similar, but categories c and d decrease with age following a similar pattern while category a increases. The credible intervals indicate high precision in the estimated means. Relevant hazard ratios (HRs) arise from the combination of covariate categories. Figure 1 on the right shows the posterior distribution of the HRs of a breast cancer diagnosis for *Famhist*, *Brstproc*, and both risk factors, with posterior means 1.864, 1.574, and 2.934, respectively. The marginal effects of each covariate are relevant, with posterior probabilities 0.998 and 1.000 to HR values greater than 1 for *Famhist* and *Brstproc*, respectively.

Figure 2 shows the posterior mean and 95% credible band for breast cancer-free survival for four women without cancer at the end of follow-up. Women 942, with stable very high breast density, is cancer-free at 68 years old and her predicted disease-free survival is higher than for women 9672, who has experienced a decrease in breast density and reaches as well 68 year old being cancer-free. The different density

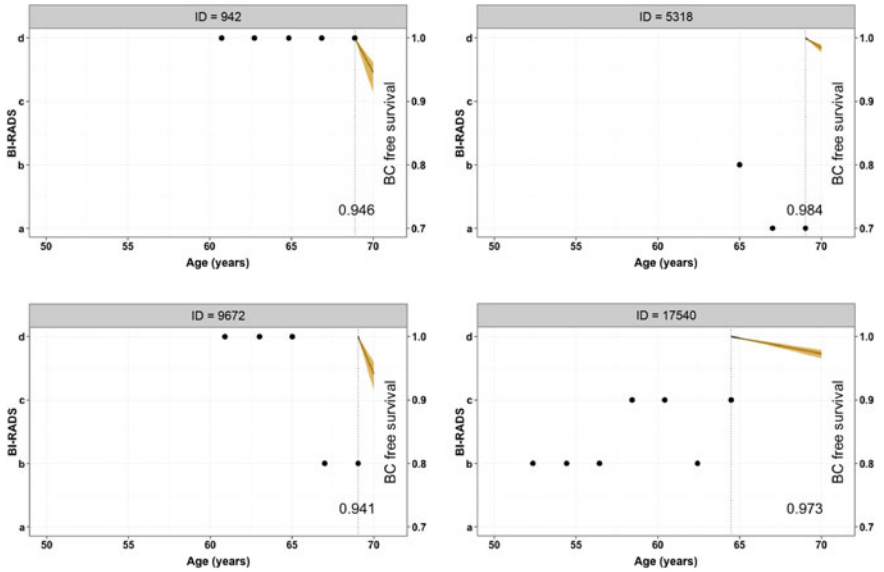


Fig. 2 Posterior mean and 95% credible band of the probability of a breast cancer-free diagnosis for women IDs 942, 5318, 9672 and 17540 without breast cancer at the end of the follow-up

behaviour might be attributed to the presence of prior breast procedures in woman 9672 and absence of them in woman 942. In general, breast cancer-free survival stays with high values, above 0.9, though they decrease with age. Furthermore, women with higher breast density values tend to have lower cancer-free survival and these probabilities depend in part of the corresponding baseline risk factors.

Acknowledgements This paper was partially supported by the research grants MTM2013-42323-P, MTM2012-38067-C02-1, PI14/00113 from the Spanish Ministry of Economy and Competitiveness, ACOMP/2015/202 from the Generalitat Valenciana, and GRBIO-2014-SGR464 and GRAES-2014-SGR978 from the Generalitat de Catalunya.

References

1. C. Armero, C. Forné, M. Rué, A. Forte, H. Perpiñán, G. Gómez, and M. Baré, "Bayesian joint ordinal and survival modeling for breast cancer risk assessment", *Stat. Med.* **35**(28) (2016), 5267–5282.
2. M. Baré, M. Sentís, J. Galcerán, A. Ameijide, X. Andreu, S. Ganau, L. Tortajada, and J. Planas, "Interval breast cancers in a community screening programme: frequency, radiological classification and prognostic factors", *Eur. J. Cancer Prev.* **17**(5) (2008), 414–421.
3. D.J. Lunn, J. Wakefield, and A. Racine-Poon, "Cumulative logit models for ordinal data: a case study involving allergic rhinitis severity scores", *Statistics in Medicine* **20** (2001), 2261–2285.
4. D. Rizopoulos, "Joint models for longitudinal and time-to-event data with applications in R", CRC Press, Biostatistics Series: Boca Raton, FL, 2012.
5. U. Uzunoḡullari and J.L. Wang, "Comparison of hazard rate estimators for left truncated and right censored data", *Biometrika* **79**(2) (1992), 297–310.

Sample Size Impact on the Categorisation of Continuous Variables in Clinical Prediction

Irantzu Barrio, Inmaculada Arostegui, and María-Xosé Rodríguez-Álvarez

Abstract Recent advances in information technologies are generating a growth in the amount of available biomedical data. In this paper, we studied the impact sample size may have on the categorisation of a continuous predictor variable in a logistic regression setting. Two different approaches to categorise predictor variables were compared.

1 Motivation

Recent advances in information technologies are generating a growth in the amount of available biomedical information and data, what is known as *Big Data*. This fact makes that the data available in some biomedical research studies is getting larger in the last years. The collection and analysis of this data allows improving the quality and efficiency of health care services and enhance the quality and longevity of life; see [9]. Furthermore, the development of prediction models to estimate the risk of developing a particular disease are nowadays relevant in the decision making process [6], with a significant growth in the number of predictive models developed in the last years. When developing prediction models to be used in clinical practice, categorised versions of continuous predictor variables are commonly used by clinical researchers; see [8]. It is possible that research studies with a big amount of data may

I. Barrio (✉)
Departamento de M.A. y Estadística e I.O., Universidad del País Vasco UPV/EHU, Leioa, Spain
e-mail: irantzu.barrio@ehu.eus

I. Arostegui
Departamento de M.A. y Estadística e I.O., BCAM-Basque Center for Applied Mathematics, Universidad del País Vasco UPV/EHU, Leioa, Spain
e-mail: inmaculada.arostegui@ehu.eus

M.-X. Rodríguez-Álvarez
Departamento de Estadística e I.O., Universidade de Vigo, Vigo, Spain
e-mail: mxrodriguez@uvigo.es

M.-X. Rodríguez-Álvarez
Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo, Spain

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_3

require or use a categorisation of the continuous predictor variables and hence, we think it is necessary to evaluate the impact the sample size may have on the selection of the cut points to categorise the predictor variable.

Recently, two different methodologies have been proposed to categorise a continuous predictor variable in a logistic regression setting; see [3, 7]. The first approach is based on a graphical display using generalised additive models (GAM, [1]) with P-spline smoothers to determine the relationship between the continuous predictor and the outcome. The second approach proposes to select the optimal cut points based on the maximisation of the area under the ROC curve (AUC) of the logistic regression model for the categorised variable. When developing a prediction model to predict the risk of poor evolution of patients with chronic obstructive pulmonary disease (COPD) in the IRYSS-COPD study [4], we categorised the predictor variable respiratory rate into three categories using both methods. The same categorisation proposal was obtained with both methods, being 20 and 24 the cut points. In this case, the sample size we had was of 1350 patients. However, we wondered whether same results would be obtained with higher sample sizes. Therefore, this question motivated the work presented in this paper, where the aim is to study how sample is related to the location of the cut points to categorise a predictor variable in a logistic regression setting.

2 Methods

In this section, we briefly describe the two methods considered for the categorisation of a continuous predictor variable. Let us assume that there is a dichotomous response variable Y and a continuous predictor variable X which we wish to categorise in a logistic regression setting.

2.1 Categorisation Proposal Based on GAM with P-Spline Smoothers

Barrio et al. [3] proposed a methodology to categorise a continuous predictor variable which consists of creating at least one average-risk category along with high- and low-risk categories based on a GAM with P-spline smoothers. Let $logit(p) = \beta_0 + f(X)$ be the logistic GAM for X, where $p = P(Y = 1|X)$ and $f()$ is the smooth function of the GAM regression model. The average-risk category is created by building an interval around the point $x_0 \in X$ for which $f(x_0) = 0$. Let us denote $\hat{\pi}_0 = logit^{-1}(\beta_0 + f(x_0)) = logit^{-1}(\beta_0)$ and $\left(\hat{\pi}_{0_{inf}}, \hat{\pi}_{0_{sup}}\right)$ its 95% confidence interval. The average-risk category $\left(x_{0_{inf}}, x_{0_{sup}}\right)$ is obtained as $f^{-1}(logit(\hat{\pi}_{0_{inf}}) - \beta_0) = x_{0_{inf}}$ and $f^{-1}(logit(\hat{\pi}_{0_{sup}}) - \beta_0) = x_{0_{sup}}$. Thus, this categorisation proposal considers at least three categories. In a context in which only two categories are considered,

Hin–Lau–Rogers–Chang [2] proposed to dichotomise a continuous variable with x_0 as the optimal cut point.

2.2 Optimal Categorisation Based on the Maximisation of the AUC

Given $k = 2$ the number of cut points set for categorising X in 3 intervals, let us denote as X_{cat} the categorised variable taking values from 0 to 2. Barrio–Arostegui–Rodríguez-Álvarez–Quintana [7] proposed that the vector of 2 cut points $\boldsymbol{v} = (\mathsf{x}_1, \mathsf{x}_2)$ which maximises the AUC of the logistic regression model

$$P(Y = 1|X_{cat}) = logit^{-1}(\beta_0 + \beta_1 1_{\{X_{cat}=1\}} + \beta_2 1_{\{X_{cat}=2\}})$$

is thus the vector of the optimal cut points. In general, this method allows to search for any possible number of cut points, nevertheless in order to compare both method we will focus on $k = 2$ number of cut points.

For ease of notation and interpretation we will refer to these two approaches as the "GAM approach" and the "AUC approach", respectively.

3 Simulation Study

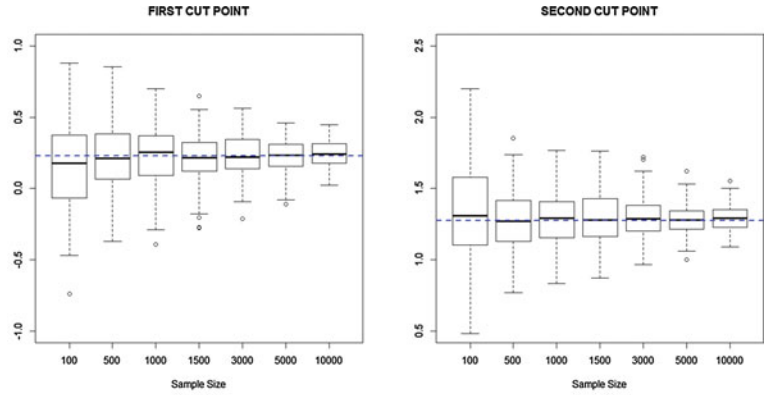
A simulation study was performed under known theoretical conditions that verify linear effects in the logistic regression model. We used this setting to evaluate the performance of the GAM approach and the AUC approach when different sample sizes were used.

3.1 Scenarios and Set Up

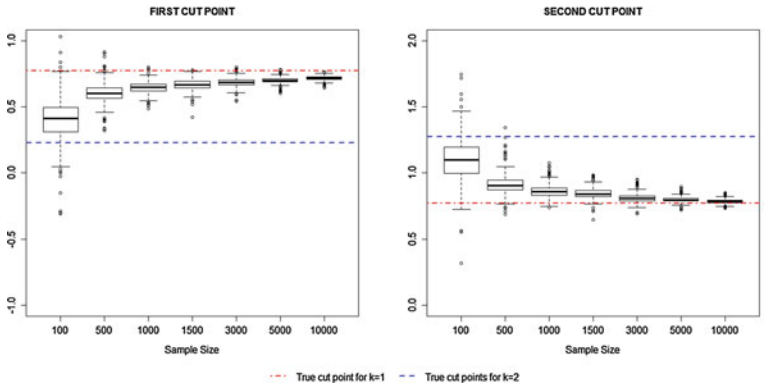
The predictor variable X was simulated from a normal distribution separately in each of the populations defined by the outcome ($Y = 0$ and $Y = 1$). Specifically, we considered $X|(Y = 0) \simeq N(0, 1)$ and $X|(Y = 1) \simeq N(1.5, 1)$. When the aim is to maximise the AUC, the theoretical location of cut points to categorise the predictor variable is known [5], as well as the AUC associated with the corresponding categorical covariate. The simulations were performed for different sample sizes assuming the same number of individuals in $Y = 0$ and $Y = 1$. As far as the number of cut points is concerned, $k = 2$ were considered. When using the AUC approach, we considered the *Genetic* algorithm to estimate the optimal number of cut points.

3.2 Results

Figure 1 depicts the boxplot of the estimated optimal cut points over 200 simulated data sets for the different sample sizes, and each of the categorisation approaches considered. Different results were obtained when the sample size was increased with the two categorisation proposals considered. The AUC approach identified the optimal cut points for any sample size considered (see Fig. 1a). Under this scenario the theoretical cut points were 0.227 and 1.274.



(a) AUC Approach



(b) GAM Approach

Fig. 1 Boxplot of the estimated cut points for $k = 2$, based on 200 simulated data sets for different sample sizes ($N = 100$, $N = 500$, $N = 1000$, $N = 1500$, $N = 3000$, $N = 5000$, and $N = 10000$). From *top* to *bottom*: results obtained with the AUC approach and the GAM approach, respectively. True cut points are represented with a *dashed line* which are $\boldsymbol{v_2} = (0.227, 1.274)$ for $k = 2$, and $\boldsymbol{v_1} = (0.773)$ for $k = 1$

Table 1 Numerical results obtained over 200 simulated data sets when the GAM approach was used to estimate the cut points. Mean, standard deviation and median for the estimated x_0 point together with the mean and standard deviation for the average-risk category values are reported

Sample size	x_0		Average-risk category	
			Low limit	Upper limit
	mean (sd)	median	mean (sd)	mean (sd)
100	0.751 (0.198)	0.759	0.395 (0.183)	1.099 (0.181)
500	0.760 (0.087)	0.752	0.605 (0.085)	0.915 (0.087)
1000	0.754 (0.050)	0.754	0.645 (0.048)	0.863 (0.054)
1500	0.755 (0.046)	0.755	0.665 (0.046)	0.845 (0.048)
3000	0.749 (0.036)	0.749	0.685 (0.036)	0.813 (0.037)
5000	0.751 (0.026)	0.748	0.701 (0.026)	0.800 (0.027)
10000	0.751 (0.019)	0.751	0.716 (0.019)	0.786 (0.019)

On the other hand, the cut points obtained with the GAM approach differed more from those theoretical cut points as the sample size was increased (see Fig. 1b). In fact, for larger sample sizes, the average-risk category obtained converged to the point x_0 for which $f(x_0) = 0$, which turned out to be close to the theoretical cut point for $k = 1$, i.e., 0.773. This results can be seen in Table 1, where the numerical results obtained with the GAM approach are shown.

4 Conclusions

To summarise, we have seen that the sample size has an impact on the categorisation of a continuous predictor variable. For a large sample size, the GAM approach leads to a very narrow average-risk category which can be interpreted as a unique cut point, being thus equivalent to the proposal of Hin–Lau–Rogers–Chang [2], this is, a dichotomisation of the continuous variable. On the other hand, the AUC approach performs satisfactorily in large sample sizes when looking for two cut points, i.e., categorising the predictor variable into three categories. In general, as long as it is feasible, we recommend the use of the AUC approach. Otherwise, we should take into account that for large sample sizes the GAM approach does not provide an optimal categorisation when the goal is to categorise the predictor variable into three categories.

Acknowledgements This study was supported by the grants IT620-13, MTM2011-28285-C02-01, UFI11/52, MTM2013-40941-P, and MTM2014-55966-P, and by the Agrupamento INBIOMED from DXPCTSUG-FEDER unha maneira de facer Europa (2012/273).

References

1. I. Barrio, I. Arostegui, J.M. Quintana, and IRYSS-COPD group, "Use of generalised additive models to categorise continuous variables in clinical prediction", *BMC Med Res Methodol* **13** (2013), 83.
2. I. Barrio, I. Arostegui, M.X. Rodríguez-Álvarez, and J.M. Quintana, "A new approach to categorising continuous variables in prediction models: proposal and validation", accepted in *Stat Methods Med Res* (2015).
3. FF. Costa, "Big data in biomedicine", *Drug Discov Today*, **19** (2014), 433–440.
4. T. Hastie and R. Tibshirani, "Generalized additive models", London, Chapman & Hall, 1990.
5. L.Y. Hin, T.K. Lau, M.S. Rogers, A.M.Z. Chang, "Dichotomization of continuous measurements using generalized additive modelling - application in predicting intrapartum caesarean delivery", *Stat Med* **18** (1999), 1101–1110.
6. J.M. Quintana, C. Esteban, I. Barrio, and the IRYSS-COPD group, "The IRYSS-COPD appropriateness study: objectives, methodology, and description of the prospective cohort", *BMC Health Serv Res* **11** (2011), 322.
7. EW. Steyerberg, "Clinical prediction models. A practical approach to development, validation, and updating", New York, Springer, 2009.
8. H. Tsuruta and L. Bax, "Polychotomization of continuous variables in regression models based on the overall C index", *BMC Med Inform Decis Mak* **6** (2006), 41.
9. E. Turner, J. Dobson, J. Pocock, "Categorisation of continuous risk factors in epidemiological publications: a survey of current practice", *Epidemiol Perspect Innov* **7**(9) (2010).

Integrative Analysis of Transcriptomics and Proteomics Data for the Characterization of Brain Tissue After Ischemic Stroke

Ferran Briansó, Teresa García-Berrocoso, Joan Montaner, and Alex Sánchez-Pla

Abstract Many diseases such as ischemic stroke, have a multigenic origin and are affected by environmental factors. Integrative omics approaches, targetting the different levels of the omics cascade are particularly appropriate in these cases and many multivariate methods have been adapted or developed in recent years to taggle this approach. In this work Multiple Co-inertia and Regularized Canonical Correlation with Sparse Partial Least Squares Regression have been applied for an integrative analysis of transcriptomics and proteomics data for the characterization of brain tissue after ischemic stroke.

1 Introduction and Objectives

The molecular basis for the genetic risk of ischemic stroke is likely to be multigenic and influenced by environmental factors; see [11]. For that reason, an integrative, multi-omics approach can be very useful to gain deeper knowledge of the genetic components of these injuries.

F. Briansó (✉) · A. Sánchez-Pla
Statistics and Bioinformatics Unit, Vall d'Hebron Institut de Recerca, Barcelona, Spain
e-mail: ferran.brianso@vhir.org

A. Sánchez-Pla
e-mail: asanchez@ub.edu

F. Briansó
Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Spain

T. García-Berrocoso · J. Montaner
Neurovascular Diseases Laboratory, Vall d'Hebron Institut de Recerca, Barcelona, Spain
e-mail: teresa.garcia@vhir.org

J. Montaner
e-mail: joan.montaner@vhir.org

A. Sánchez-Pla
Statistics Department, Universitat de Barcelona, Barcelona, Spain

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_4

Over the past decade, major advance in omics technologies have facilitated a high-throughput monitoring and understanding of a variety of molecular and organismal processes. These techniques, and its translational adoption, have been widely applied to identify biological agents and to characterize biochemical systems, often focusing on the discovery of therapeutic targets and biomarkers with application to specific diseases. While many analysis tools target comprehensive analysis of genes (genomics), mRNA (transcriptomics), or proteins (proteomics), there is still a long way to go in the field of omics data integration in order to provide a better understanding of the biological systems. To address this challenge, several multivariate statistical techniques have been revised and proposed in the last years: some of them based on classical dimension reduction, such as Principal Component Analysis, but others based on more "novel" approaches, such as Regularized Canonical Correlation Analysis (rCCA) and Multiple Co-inertia Analysis (MCIA); [10] is a recent review of them.

The main objectives of this work are: (i) to perform an integrative analysis of transcriptomics and proteomics data of an study on ischemic stroke, and (ii) to investigate the possibility of combining two multivariate approaches for omics data integration.

2 Methods

For this study, human brain tissue samples, collected by the Neurovascular Diseases Laboratory at Vall d'Hebron Research Institute, have been processed to obtain mRNA and protein expression values. Each dataset was first analysed independently using standard bioinformatics protocols [6]. These analyses allowed to select subsets of relevant features, for each type of data, to be used in the integrative analysis.

Among all available options, we decided to use two distinct and complementary approaches: (i) Multiple Co-inertia Analysis, implemented in Bioconductor packages `made4` [2] and `mogsa` [8]; and (ii) Regularized Canonical Correlation Analysis with Sparse Partial Least Squares regression (sPLS), provided by `mixomics` R package [3].

MCIA [9] is a technique allowing to combine two or more dimension reduction analyses by searching pairs of axes (one in each analysis) that maximize covariance. rCCA [7] allows to analyse two groups of variables by forming linear combinations in each group that maximize the correlation between them. While standard CCA can not be used where the number of samples is far lower than the number of variables, rCCA is useful in these cases. PLS [5] is similar to CCA, but it is assumed that one group of variables depends on the other group and is computed by finding linear combinations of the variables of each group that maximize the covariance.

MCIA and rCCA were applied to the proteomics and transcriptomics data sets to find out relations within and between both groups of variables. Finally, the biological annotations of each feature to the Gene Ontology database [1] have been used, with the MCIA approach, in order to obtain a better understanding of the biological differences between infarcted and healthy brain areas under study.

3 Results

The application of rCCA and sPLS to both datasets allow to visualize the results of the analysis through clustered image maps and relevance networks (Figs. 1 and 2); see [4]. These pictures depict relations between genes and proteins that could not be determined neither from separate multivariate analyses nor from the previously known gene-protein associations. Four major clusters of gene-protein associations showing positive and negative co-expressions, coloured in red and blue, respectively, are shown in Fig. 1. A network resulting from keeping only the most correlated elements (absolute coefficient value bigger than or equal to 0.85), with red- and green-coloured edges, for top positive and negative correlations, respectively, is included in Fig. 2.

MCIA was used through mogsa package in order to measure how experimental samples are related between them, in terms of the projection of gene and protein components within the same bi-dimensional space (Fig. 3). This shows that two samples have an ambiguous behaviour, being grouped with the other condition from one of the two, gene or protein, points of view, so having their arrow lines crossing the vertical axis of the scatterplot. In a complementary way, a Circle Correlation Plot (Fig. 4) allows to identify among genes (light blue), some which are not allocated with any of the protein clusters (orange). These genes have special interest because, although they are differentially expressed between ischemic stroke and healthy samples, they

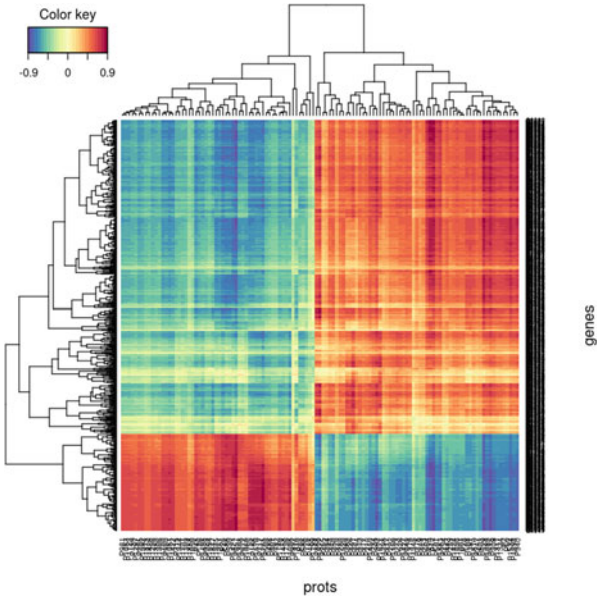


Fig. 1 Clustered image map (heat map of correlation coefficients between genes and proteins)

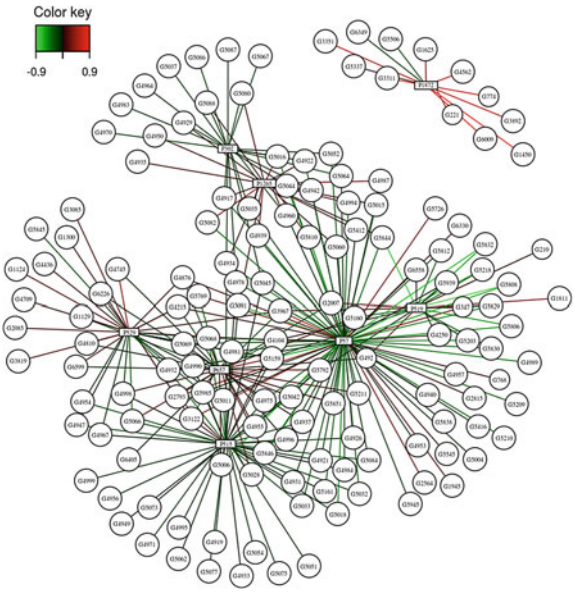


Fig. 2 Relevance network of top correlations between genes (*circles*) and proteins (*rectangles*)

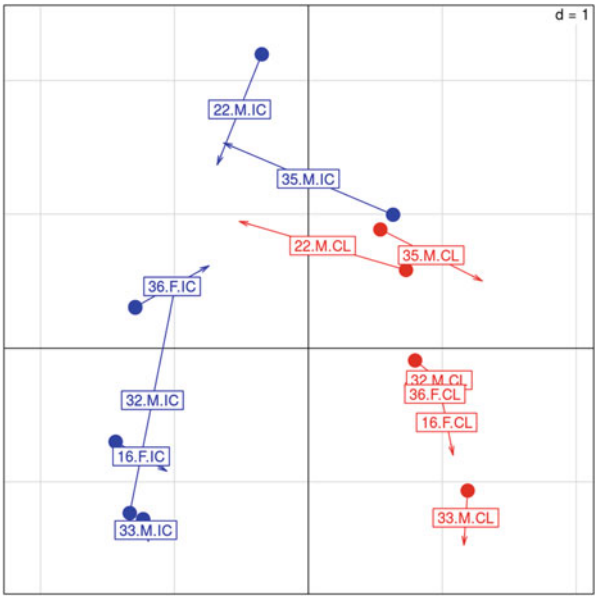


Fig. 3 Distribution of samples based on their transcriptomic (*dot*) and proteomic (*arrow*) information for healthy (*red*) and affected (*blue*) samples

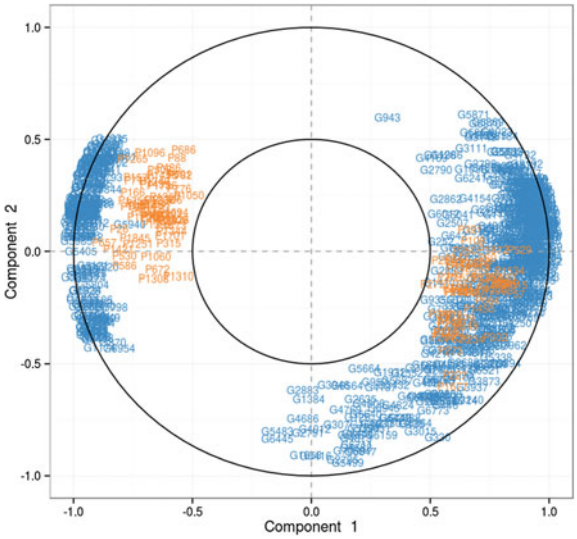


Fig. 4 Correlation *circle plot* allocating gene and protein features along the main components derived from the integration of both data sets

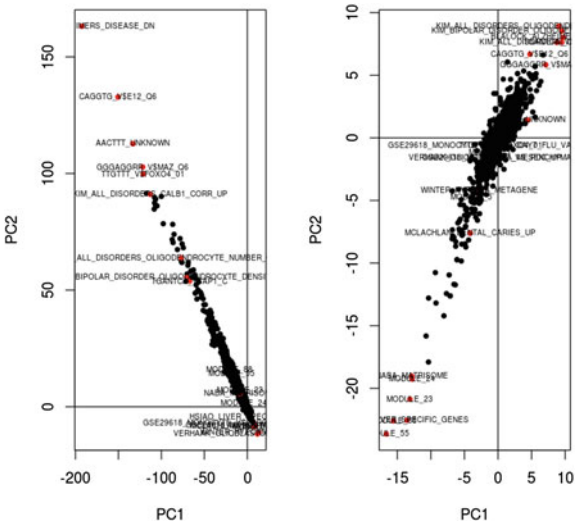


Fig. 5 Gene set plots overimposing biological information in the form of Gene Ontology annotations to genes (*left*) or proteins (*right*) representations

are not directly related with the first gene-protein derived component, that mainly distinguishes between both experimental conditions.

In addition to that, MCIA allowed the annotation of the genes and proteins analysed with this method against the gene sets extracted from the Gene Ontology data base, which have been used to directly overlap the main enriched biological gene sets in the plane formed by the corresponding principal components. Figure 5 above shows a projection of gene sets on the main components for gene (left) and

protein (right) datasets. Among the top gene sets found with mogsa functions is worth to notice two cases, one related with Alzheimer's disease and other with some psychiatric disorders, respectively.

4 Conclusions

Two distinct approaches for projection-based multivariate analysis of gene and protein data have been applied to characterize human brain tissue samples, also including the annotation to standard biological databases as a method for merging information in a common space. The first approach (rCCA and sPLS with `mixomics`) provided a good visualization of individual relationships between features (shown in Figs. 1 and 2), and could be used for variable selection, but did not allow the addition of biological information. On the other hand, the other approach (MCIA and Gene Set Analysis with `mogsa`) could not perform variable selection, but has shown to be useful for presenting samples, features and its associated biological information (Figs. 3, 4, 5, respectively) in a common projection space. In summary, both approaches have been able to show aspects of relations between genes and proteins (Figs. 2 and 4) that could not have been unveiled separately, which is the main goal of integrative analysis. Some aspects were similar but others were not, showing the complementarity between the approaches.

This study is a first step in a larger collaborative project, on which more omics data have been collected. The preliminary analysis suggests how to perform the integration with other omics, and our aim is now to merge these approaches in a wrapper pipeline, which will be used in order not only to characterize the experimental conditions with an integrative method, but also to find ways to present the results in a more "understandable way", from the point of view of its biomedical and translational interpretation.

Acknowledgements We wish to acknowledge GRBIO (2014 SGR 464), for funding part of this work.

References

1. "Gene ontology consortium: going forward", *Nucleic Acids Res*, **43** (2015), (Database issue):D1049–D1056.
2. A.C. Culhane, J. Thioulouse, G. Perrière, and D.G. Higgins, "Made4: an R package for multivariate analysis of gene expression data", *Bioinformatics* **21**(11) (2005), 2789–2790.
3. I. González, K.A. Lê Cao, and S. Déjean, "mixOmics: Omics data integration project" (2011), availabe at url:http://www.mixomics.org.
4. I. González, K.A. Lê Cao, M.J. Davis, and S. Déjean, "Visualising associations between paired 'omics' data sets", *BioData Mining* **5** (2012), 19.
5. T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning", Springer, New York Inc. 2001.

6. W. Huber, V.J. Carey, R. Gentleman, S. Anders, M. Carlson, B.S. Carvalho, H.C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K.D. Hansen, R.A. Irizarry, *et.al*., "Orchestrating high-throughput genomic analysis with Bioconductor", *Nature Methods* **12**(2) (2015), 115–121.
7. K.A. Lê Cao, I. González, and S. Déjean, "integrOmics: An R package to unravel relationships between two omics datasets", *Bioinformatics* **25**(21) (2009), 2855–2856.
8. C. Meng, "mogsa: Multiple omics data integrative clustering and gene set analysis", R package v1.6.2.
9. C. Meng, B. Kuster, A.C. Culhane, and A.M. Gholami, "A multivariate approach to the integration of multi-omics datasets", *BMC Bioinformatics* **15** (2014), 162.
10. C. Meng, O.A. Zeleznik, G.G. Thallinger, B. Kuster, A.M. Gholami, and A.C. Culhane, "Dimension reduction techniques for the integrative analysis of multi-omics data", *Brief Bioinform* **1** (2016), 14.
11. J.F. Meschia, T.G. Brott, R.D. Brown, R.J.P. Crook, M. Frankel, J. Hardy, J.G. Merino, S.S. Rich, S. Silliman, and B.B. Worrall, "The ischemic stroke genetics study (ISGS) protocol", *BMC Neurology* **3** (2003), 4.

Applying INAR-Hidden Markov Chains in the Analysis of Under-Reported Data

Amanda Fernández-Fontelo, Alejandra Cabaña, Pedro Puig, and David Moriña

Abstract We present a model for under-reported time series count data in which the underlying process satisfy an INAR(1) structure. Parameters are estimated through a naïve method based on the theoretical expression of the autocorrelation function of the underlying process, and also by means of the forward algorithm. The hidden process is reconstructed using the Viterbi algorithm, and a real data example is discussed.

1 Introduction

Time series analysis is an old discipline. However, dealing with count data is relatively recent. Under-reported phenomena are present in almost any field, but are specially interesting when counts are low such as the number of reported cases of a rare disease; see [1, 3]. A good real example to show the applicability of the model is the number of weekly cases of *Human papillomavirus* (HPV) in Girona from 2010 to 2014,

A. Fernández-Fontelo (✉) · A. Cabaña · P. Puig
Departament de Matemàtiques, Universitat Autònoma de Barcelona, Bellaterra, Spain
e-mail: amanda@mat.uab.cat

A. Cabaña
e-mail: acabana@mat.uab.cat

P. Puig
e-mail: ppuig@mat.uab.cat

D. Moriña
Unit of Infections and Cancer (UNIC), Cancer Epidemiology Research Program (CERP), Catalan Institute of Oncology (ICO)-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain
e-mail: dmorina@creal.cat

D. Moriña
ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain

D. Moriña
Universitat Pompeu Fabra (UPF), Barcelona, Spain

D. Moriña
CIBER Salud Pública y Epidemiología, Barcelona, Spain

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_5

which is a stationary series. This disease is a very common sexual infection in such a way that nearly all sexual active people have the infection at some point in their lives; see [4]. However, the infection disappears on its own in most cases, and only in some of them becomes a more serious disease. For that reason, it seems sensible to consider that the HPV can be severely under-reported.

The simplest model for completely observed count series data is an INAR(1) of the form

$$X_n = \alpha \circ X_{n-1} + W_n, \tag{1}$$

where $\alpha \in (0, 1)$ is a fixed parameter and the operator $\circ$ is the *binomial thinning* operator, such that $\alpha \circ X_{n-1} = \sum_{i=1}^{X_{n-1}} Z_i(\alpha)$, where Z_i are i.i.d Bernoulli random variables with success probability α. Here, W_n is assumed to be Poisson(λ) distributed. The probabilistic and statistical properties of these models are widely studied; see [2, 6].

2 The Model

Let X_n be the unobserved process satisfying an INAR(1) as in Eq. (1). Hence, if the observed process Y_n is under-reported, it can be written as

$$Y_n = \begin{cases} X_n & \text{with probability } 1-\omega, \\ q \circ X_n & \text{with probability } \omega. \end{cases} \tag{2}$$

That is Y_n, which is a *binomial thinning* of the underlying process X_n, coincides with X_n with probability $1-\omega$, implying that the observed count at time n is not under-reported. It is important to notice that ω is defined as the proportion of times that Y_n does not coincide with X_n, that is, ω is the frequency of the under-reported phenomenon. Parameter q is the intensity of the under-reportation in the sense that this phenomenon will be more intense for small values of q. Observe that Y_n can be understood as a hidden Markov chain with an infinite number of states.

Taking into account that X_n follows an INAR(1) process with Poisson innovations, it is easy to see that the marginal distribution of X_n is Poisson($\lambda/(1-\alpha)$), and its auto-correlation function is such that $\rho_X(k) = \alpha^{|k|}$, as in the classical AR(1) model. According to (2), the mean of Y_n is $(1-\omega(1-q))\frac{\lambda}{(1-\alpha)}$, and its variance $\frac{\lambda^2}{(1-\alpha)^2}\omega(1-\omega)(1-q)^2 + \frac{\lambda}{(1-\alpha)}(1-\omega(1-q))$. In [5] it is shown that the auto-correlation function of Y_n is

$$\rho_Y(k) = \frac{(1-\alpha)(1-\omega(1-q))^2}{(1-\alpha)(1-\omega(1-q)) + \lambda(\omega(1-\omega)(1-q)^2)}\alpha^{|k|} = c(\alpha, \lambda, \omega, q)\alpha^{|k|}. \tag{3}$$

2.1 Parameter Estimation

The marginal distribution of Y_n is a mixture of two Poisson distributions

$$Y_n \sim (1-\omega)\text{Poisson}(\lambda/(1-\alpha)) + \omega\text{Poisson}(q\lambda/(1-\alpha)), \tag{4}$$

since the marginal distribution of X_n is Poisson$(\lambda/(1-\alpha))$. Accordingly, applying the EM algorithm to fit Poisson mixture distributions, we obtain estimates for ω, $(\lambda/(1-\alpha))$ and $(q\lambda/(1-\alpha))$, and using the last two, q is directly obtained. Then, α can be estimated by using the theoretical expression of the ACF of Y_n, that is, replacing in (3) the parameters ω and q for their estimates, λ for its estimated value from $(\lambda/(1-\alpha))$, and equalling this ACF to $\widehat{\rho}_1$. Once $\widehat{\alpha}$ is obtained, λ is immediately computed using, for example, $(\lambda/(1-\alpha))$. These estimates can be used as initial values in the algorithm that maximises the likelihood function of Y_n. In addition, it is important to remark that this naïve method can produce values out of its corresponding domain, but this fact is not unusual since the method of moments has the same limitation.

The parameters of the model can also be estimated by means of maximum likelihood method. In that case, if $Y_n = (Y_1, Y_2, \ldots, Y_n)$ is the observed process and $X_n = (X_1, X_2, \ldots, X_n)$ the underlying process, then the likelihood function of the model is:

$$P(Y) = P(Y_1, Y_2, \ldots, Y_n) = \sum_X P(X, Y) = \sum_x P(Y \mid X = x)P(X = x). \tag{5}$$

However, the function (5) is intractable since Y_n can be thought of as a hidden Markov chain with an infinite number of states, and then the problem becomes computationally unsolvable. The forward algorithm, which is used in the context of hidden Markov chains, can be an appropriate option in order to compute the likelihood function of Y_n. According to that, we have modified this algorithm to fit our needs. Straightforward computations lead to the following expression for the forward probabilities:

$$\alpha_k(X_k) = P(Y_1, \ldots, Y_k, X_k) = P(Y_k|X_k) \sum_{X_{k-1}} P(X_k \mid X_{k-1})\alpha_{k-1}(X_{k-1}). \tag{6}$$

Hence, the likelihood function of Y_n is computed by $P(Y) = \sum_{X_n} \alpha_n(X_n)$, according to (6) and assuming that $\alpha_1(X_1) = P(Y_1, X_1) = P(X_1)P(Y_1 \mid X_1)$.

The expression (6) is based on the transition and emission probabilities. In our case, the transition probabilities are defined by means of the probability distribution of an INAR(1) model as detailed in [5, 7], while the emission probabilities are equal to

$$P(Y_i = j \mid X_i = k) = \begin{cases} 0 & \text{if } k < j, \\ (1-\omega) + \omega q^k & \text{if } k = j, \\ \omega \binom{k}{j} q^j (1-q)^{k-j} & \text{if } k \geq j. \end{cases} \tag{7}$$

2.2 Reconstruction of the Underlying Process

In order to reconstruct the underlying process X_n, we apply the Viterbi algorithm [8]. Given the observed process $Y_{1:n} = (Y_1, Y_2, \ldots, Y_n)$ and the hidden process $X_{1:n} = (X_1, X_2, \ldots, X_n)$, the aim of the algorithm is to maximise $P(X_{1:n}, Y_{1:n})$. That is, let $P(X_{1:n}|Y_{1:n})$ be the likelihood function of Y_n, then it is enough to maximise $P(X_{1:n}, Y_{1:n})$, since $P(Y_{1:n})$ does not depend on the underlying process. Finally, the reconstructed chain is obtained by using $X^* = \arg\max_X P(X_{1:n}, Y_{1:n})$.

2.3 Model Selection and Goodness of Fit

Model selection is based on both the statistical significance of the model parameters and also some criteria of parsimony such as the Akaike information criterion (AIC). Model validation is performed using the mid-normal-pseudo-residuals based on the normal-pseudo-residuals-segments according to the discreteness of the variables. That is, if the model is valid, the mid-normal-pseudo-residuals might be similar to white noise; see [9].

3 Example of Application

The data set ranges from 0 to 6 cases per week, with a mean and a median of 1 case per week and a variance of 1.61. A slight overdispersion is present, which is consistent since a mixture of Poisson distributions is always overdispersed. Figure 1 shows the behaviour of the series during the period of study, and it is immediate to see that the series has no pattern of trend and/or seasonality.

The naïve method described in Sect. 2 can be used to evaluate whether the series can actually be under-reported in two different ways. Firstly, the observed process Y_n can be modelled by means of a Poisson distribution or a mixture of two Poisson distributions, and then both models can be compared using the AIC. Here, the mixture of two Poisson distributions seems to fit the phenomenon better since the AIC of this mixture is smaller (783.843) than the AIC of the Poisson distribution (787.075), indicating that the observed process can be under-reported. On the other hand, we can evaluate whether Y_n is under-reported by fitting the following model: $\log(\rho_k) = \log(c(\alpha, \lambda, \omega, q)) + \mathrm{k}\log(\alpha)$ and studying whether its intercept is statistically significant (under-reported). In our case the intercept is statistically significant

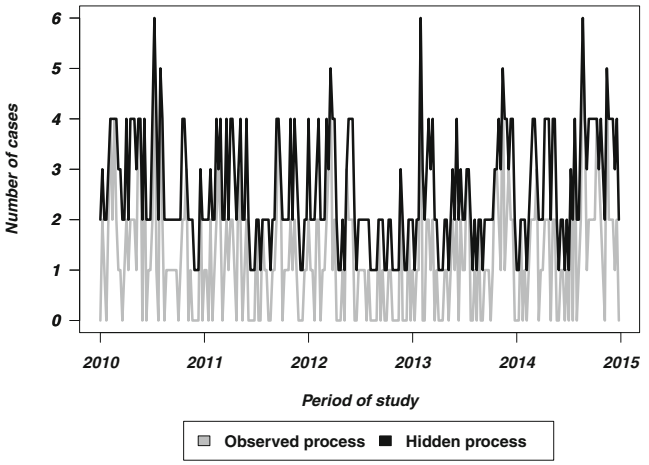


Fig. 1 The observed process Y_n and the most probable reconstructed chain X_n

Table 1 MLE (and standard errors) of the model

Parameter	MLE	s.e.
$\hat{\alpha}$	0.517	0.227
$\hat{\lambda}$	1.623	0.616
$\hat{\omega}$	0.922	0.073
$\hat{q}$	0.326	0.085

(p-value $= 0.002$) leading to an under-reported observed process. Table 1 shows the maximum likelihood estimators of the parameters. Finally, the unobserved process is reconstructed as shown in Fig. 1 and the model is validated (residuals are similar to white noise).

HPV in Girona seems to be severely under-reported since the frequency (ω) of the phenomenon is in (0.78, 1.00), and the intensity (q) in (0.153, 0.487). It is interesting to remark that all the observed zeroes are under-reported.

References

1. J.H. Alfonso, E.K. Løvseth, Y. Samant, and J.Ø. Holm, "Work-related skin diseases in Norway may be underreported: data from 2000 to 2013", *Contact Dermatitis* **72**(6) (2015), 409–412.
2. M.A. Al-Osh and A.A. Alzaid, "First-order integer-valued autoregressive (INAR(1)) process'", *Journal of Time Series Analysis* **8**(3) (1987), 261–275.
3. M. Boulanger, F. Morlais, V. Bouvier, F. Galateau-Salle, L. Guittet, M.F. Marquignon, C. Paris, *et al.*, "Digestive cancers and occupational asbestos exposure: incidence study in a cohort of asbestos plant workers", *Occupational and Environmental Medicine* **72**(11) (2015), 792–797.

4. E.F. Dunne, L.E. Markowitz, M. Saraiya, S. Stokley, A. Middleman, E.R. Unger, A. Williams, and J. Skander, "CDC grand rounds: reducing the burden of HPV-associated cancer and disease", *Morbidity and Mortality Weekly* **63**(4) (2014), 69–72.
5. A. Fernández-Fontelo, A. Cabaña, P. Puig, and D. Moriña, "Under-reported data analysis with INAR-hidden Markov chains", *Statistics in Medicine* **35**(26) (2016), 4875–4890.
6. E. McKenzie, "Some simple models for discrete variate time series 1", *Journal of the American Water Resources Association* **21**(4) (1985), 645–650.
7. D. Moriña, P. Puig, J. Ríos, A. Vilella, and A. Trilla, "A statistical model for hospital admissions caused by seasonal diseases", *Statistics in Medicine* **30**(26) (2011), 3125–3136.
8. A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", *IEEE Transactions on Information Theory* **13**(2) (1967), 260–269.
9. W. Zucchini and I.L. MacDonald, "Hidden Markov models for time series: an introduction using R" (2009).

Joint Modelling for Flexible Multivariate Longitudinal and Survival Data: Application in Orthotopic Liver Transplantation

Ipek Guler, Christel Faes, Carmen Cadarso-Suárez, and Francisco Gude

Abstract Orthotopic liver transplantation (OLT) is the established treatment for end-stage liver disease and acute fulminant hepatic failure. The clinical interest lies on the association between post-operative glucose profiles, daily therapy with insulin and the risk of death. We propose a two-staged model based approach for flexible modelling of multivariate longitudinal and survival data to study these associations.

1 Introduction

Orthotopic liver transplantation (OLT) data includes patients who underwent OLT in the Hospital Clínico Universitario de Santiago, between July 1994 and July 2011. Alterations in glucose metabolism are common among patients undergoing surgery, and are associated with increased risk of mortality and morbidity. To maintain glucose levels within the range of normality, implementation of different protocols have been developed, but such strict control does not necessarily entail a decrease in mortality. The interest is to study the association between post-operative glucose profiles, daily therapy with insulin and the risk of death. For this aim, it is important to use appropriate statistical tools to study this association when the daily insulin

I. Guler (✉) · C. Cadarso-Suárez
Department of Statistics and Operations Research, University of Santiago de Compostela, Santiago de Compostela, Spain
e-mail: ipek@usc.es

C. Cadarso-Suárez
e-mail: cadarso@usc.es

C. Faes
Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, Belgium
e-mail: christel.faes@uhasselt.be

F. Gude
Clinical Epidemiology Unit, Hospital Clínico Universitario de Santiago, Santiago de Compostela, Spain
e-mail: francisco.gude.sampedro@sergas.es

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_6

therapy and glucose measurements are highly correlated to each other. In addition, both glucose and insulin measurements have non-linear profiles over time.

An appropriate approach should be a joint modelling of longitudinal and survival data maximizing the joint likelihood of both longitudinal and survival processes. These models based on conditional distributions where a shared random effect underlines both longitudinal and survival process; see Wulfsohn–Tsiatis [8]. However, these approaches are difficult to implement when the number of longitudinal biomarkers is large and their profiles are non-linear. Due to the computational complexity, it is hard to evaluate the joint likelihood in these situations.

Fieuws–Verbeke [3] and Fieuws–Verbeke–Molenberghs [4] have proposed an approach for the multivariate longitudinal processes in which all the possible pairs of longitudinal data are separately modelled and then combined in a final step. We use the latter approach to fit a flexible multivariate longitudinal model using spline smoothing for longitudinal profiles; Ruppert–Wand–Carroll [6]. Furthermore, we use the true unobserved value of longitudinal biomarkers to incorporate them into the survival process as a second stage.

2 Orthotopic Liver Transplantation Data

A total of 644 patients were available for this study. The participants were followed up until the primary event (death) happens. Median follow-up was 5.6 years (range: [0.1, 17.5]). Patients were classified as patients with known diabetes and the study includes only those patients without diabetes who received insulin during their first

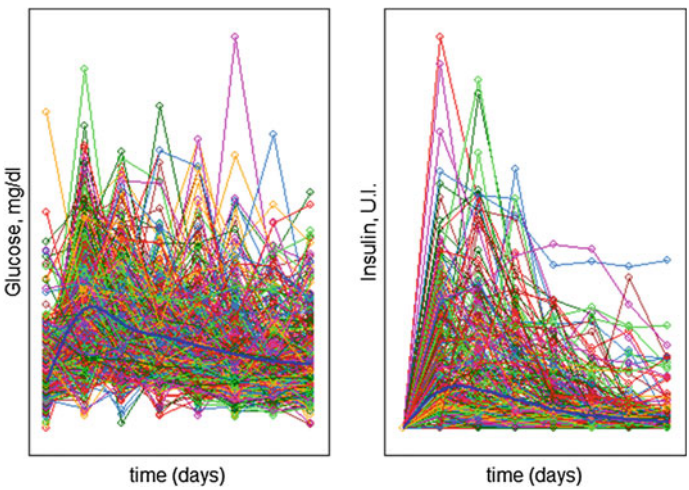


Fig. 1 Subject specific trends of glucose and insulin measurements with corresponding overall trends using a natural cubic spline smoothing

post-operative week. The variability of subject specific profiles of both glucose and insulin measurements can be observed in Fig. 1a. Figure 1b represents the overall profiles using a natural cubic spline smoothing; see Boor [2].

3 Two-Stage Model Based Proposal

The main idea of this two-stage model based proposal is coming from the initial approaches of simple longitudinal and time-to-event data (see Tsiatis–DeGruttola–Wulfsohn [7], among others) where the likelihood calculation is divided into two stages, separately for the single longitudinal outcome and survival process. Extending this idea, we can use the pairwise modeling approach to study the multivariate longitudinal data introduced by Fieuws–Verbeke [3] in the first stage. This approach is taking into account the correlation structure of the longitudinal biomarkers, assuming a multivariate normal distribution for random effects. To obtain flexibility, a spline smoothing for longitudinal profiles can be used; see Ruppert–Wand–Carroll [6]. Furthermore, the predictions of this model are incorporated into the survival model in Stage 2. This model proposal is composed by two stages: (1) a flexible multivariate longitudinal model; and (2) a proportional hazard regression to study the survival.

3.1 Stage 1: Flexible Multivariate Longitudinal Data

In the first stage of the model proposal, we introduce a flexible multivariate longitudinal model. Let $Y_{i,j,k}$ be the k-th longitudinal biomarker for subject i at time j, being $i = 1, \ldots, N$. We have

$$\log(Y_{i,j,k}) = \log(\beta_{0,k} + f_{i,j,k}(time) + u_{0,i,k} + \epsilon_{i,j,k}),$$

where k is the number of longitudinal biomarkers for glucose and insulin measurements (in our application, $k = 2$), $u_{0,i,k}$ is a random intercept effect, $f_{i,k}(time)$ is the smooth function of time with truncated spline basis, which is represented as a linear mixed model (see Ruppert–Wand–Carroll [6]), and finally $\epsilon_{i,j,k}$ is the error term. The function is

$$f_{i,k}(time) = \beta_{0,k} + \beta_{1,k} * time + \beta_{2,k} * time^2 + \beta_{3,k} * time^3 + \sum_{i=1}^{L} U_k (x_i - \kappa_l)^3{}_+,$$

where $\beta_{0,k}, \ldots, \beta_{3,k}$ represent the fixed effects of the linear mixed model representation of the smoothing term, and U_k represents the random part of this representation with $(x_i - \kappa_l)_+$ quadratic spline basis with knots $\kappa_1, \ldots, \kappa_l$. The degree of truncated power basis is equal to 3 in this case, which represents cubic splines. In the

pairwise fitting approach, the log likelihood of the following form will be maximized separately:

$$\sum_{i=1}^{N} l_{pi}(\theta_p),$$

where $p = 1, \ldots, P$ with $\mathrm{P} = m(m-1)/2$ indicating the total number of possible pairs (in this case we have only one pair), and let θ then be the stacked vector combining all pair-specific parameter vectors θ_p. Estimates for the elements in θ are obtained by maximizing each of the P likelihood separately; see Fieuws–Verbeke [3] and Fieuws–Verbeke–Molenberghs [4].

Inference for θ

Although in the pairwise approach each likelihood is maximized separately, the approach fits within the pseudo-likelihood framework. Indeed, fitting all possible pairwise models is equivalent to maximizing a pseudo-likelihood function of the form

$$pl(\theta) = l(Y_1, Y_2/\theta_{1,2}) + l(Y_1, Y_3/\theta_{1,3}) + l(Y_2, Y_3/\theta_{2,3}).$$

The asymptotic multivariate normal distribution for θ is given by

$$\sqrt{N}(\theta - \theta) \approx MV(0, J^{-1} K J^{-1}),$$

where J is a block-diagonal matrix with diagonal blocks $J_{p,p}$, and where K is a symmetric matrix containing blocks $K_{p,q}$. In the final step, estimates for the parameters can be calculated by taking averages over all pairs.

3.2 Stage 2: Survival Model

In the second stage, we introduce the following Cox model, including the unobserved values of glucose and insulin measurements obtained estimated in stage 1:

$$h_i(t) = h_0(t) \exp(age_i(t)\gamma_1 + gender_i(t)\gamma_2 + \alpha_1 m_{i1}(t) + \alpha_2 m_{i2}(t)),$$

where $h_0(t)$ is the baseline hazard function. Therefore, α_1 and α_2 represent the coefficient of association between the longitudinal markers and patients' survival. And $m_{i,1}(t)$ and $m_{i,2}(t)$ are the true unobserved values of glucose and insulin measurements at time t, respectively, $m_{i,k}(t) = \log(\beta_{0,k} + f_{i,j,k}(time) + u_{0,i,k})$.

Table 1 Results of the survival model in stage 2

Survival model				
		Coef (Std.Error)	p-value	HR (95% CI)
Fixed effects	Women (γ_1)	0.20 (0.06)	< 0.01	1.22 (1.09 − 1.38)
	Age (yr) (γ_1)	0.02 (0.002)	< 0.01	1.02 (1.02 − 1.02)
Association	log(Glucose) (α_1)	See Fig. 1b	< 0.01	See Fig. 1b
	log(Insulin) (α_2)	See Fig. 1b	< 0.01	See Fig. 1b
$logLik$		−12923		

4 Conclusions

In patients without diabetes who underwent liver transplantation, glycemic levels display a marked rise at 24–48 h, and then subsequently declined (all within the context of insulin being administered via continuous perfusion). This behaviour could reflect glycemic response to stress; see Fig. 1b. Blood glucose profiles were observed to be statistically associated with long-term mortality among patients without diabetes ($p < 0.01$), despite insulin being administered via continuous perfusion to maintain glycemia figures between normative ranges. Due to having non-linear trends for longitudinal biomarkers, the interpretation of the coefficients of association (α_1 and α_2) becomes compromised. Thus, the overall glucose and insulin profiles are shown in Fig. 1b.

Our proposed two-stage based model allows flexibility on both longitudinal and survival models and avoid computational problems in case of having large number of longitudinal biomarkers or non-linear profiles. The limitation of this proposed model could be the unignorable informative censoring on the longitudinal model cause of drop-out process. A regression calibration approach can be used to account for informative drop-out in the longitudinal part as Albert–Shih [1] presented in their approach. However, in some cases, for external longitudinal measurements, the informative censoring can be ignored; see Murawska–Rizopoulos–Lesare [5] (Table 1).

Acknowledgements This work was supported by the project MTM2014-52975-C2-1-R: "Inference in Structured Additive Regression (STAR) Models with Extensions to Multivariate Responses. Applications in Biomedicine", cofinanced by the Ministry of Economy and Competitiveness (SPAIN) and by the European Regional Development Fund (FEDER).

References

1. P.S. Albert and J.H. Shih, "On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure", *Biometrics* **3** (2010), 983–987.
2. C. de Boor, "*A practical guide to splines*", Springer (2001).

3. S. Fieuws and G. Verbeke, “Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles”, *Journal of Statistical Software* **62**(2) (2006), 424–431.
4. S. Fieuws, G. Verbeke, and G. Molenberghs, “Random-effects models for multivariate repeated measure”, *Statistical Methods in Medical Research* **16** (2007), 387–397.
5. M. Murawska, D. Rizopoulos, and E. Lesaffre, “A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies”, *Journal of Probability and Statistics*, article ID 194194 (2012), 18 pages.
6. D. Ruppert, M. Wand, and R. Carroll, “Semiparametric Regression”, Cambridge University Press, Cambridge, UK, (2003).
7. A.A. Tsiatis, V. DeGruttola, and M.S. Wulfsohn, “Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS”, *Journal of American Statistical Association* **90** (1995), 27–37.
8. M.S. Wulfsohn and A.A. Tsiatis, “A joint model for survival and longitudinal data measured with error”, *Biometrics* **53** (1997), 330–339.

A Multi-state Model for the Progression to Osteopenia and Osteoporosis Among HIV-Infected Patients

Klaus Langohr, Nuria Pérez-Álvarez, Eugenia Negredo, Anna Bonjoch, Montserrat Rué, Ronald Geskus, and Guadalupe Gómez

Abstract We model the evolution of bone mineral density measurements in a cohort of HIV-infected persons by DXA scans. We define the minimum T-score (MTS) from the DXA measures at four different sites and propose a disease progression model for the transitions between three different health states: normal, osteopenia, and osteoporosis. A linear mixed model for the MTS is fitted, the estimated ages at osteopenia and osteoporosis onset are imputed, and the transition probabilities between the states are estimated.

1 Motivation

Bone mineral density (BMD) measurements are used to determine bone health and can help identifying subjects at risk of fracture. The most widely recognized BMD scan, which measures bone density at hip and spine, is called dual-energy

K. Langohr (✉)
Universitat Politècnica de Catalunya, Barcelona, Spain
e-mail: klaus.langohr@upc.edu

N. Pérez-Álvarez · E. Negredo · A. Bonjoch
Fundació Lluita contra la SIDA, Badalona, Spain
e-mail: nperez@flsida.org

E. Negredo
e-mail: enegredo@flsida.org

A. Bonjoch
e-mail: abonjoch@flsida.org

M. Rué
Universitat de Lleida, Lleida, Spain
e-mail: montse.rue@cmb.udl.cat

R. Geskus
Academic Medical Center, Amsterdam, The Netherlands
e-mail: statistics@inter.nl.net

G. Gómez
Universitat Politècnica de Catalunya, Barcelona, Spain
e-mail: lupe.gomez@upc.edu

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_7

x-ray absorptiometry (DXA). The DXA measures are compared to the BMD of a healthy 30-years-old adult of the same gender and converted into T-scores: T-scores above -1 are considered normal, values between -1 and -2.5 indicate low bone mass (osteopenia), and values below -2.5 indicate osteoporosis. Even though no bone density test is 100% accurate, the BMD test is an important predictor for the risk of a fracture; see [1].

The main goal of this study is to determine the evolution of BMD in a cohort of nearly 1300 HIV-infected persons with a total of 3610 DXA scans as a function of gender and age. In particular, we are interested in estimating the transition probabilities from normal BMD to osteopenia and osteoporosis. The clinical relevance of building a model for such purpose is to guide the clinical practice by acting on the risk factors in these patients at higher risk of progression, and the rationalization of DXA scans measurements. This study follows previous analyses, on a subset of these data, where the focus was the estimation of the distribution of the time from normal to osteopenia and from osteopenia to osteoporosis; see [4].

2 Multi-state Model

We propose the use of a disease progression model as shown in Fig. 1, which considers three different health states defined by the minimum T-score (MTS) from the DXA measures at four different sites in the femur and the lumbar region: normal BMD (MTS ≥ -1), osteopenia ($-1 >$ MTS ≥ -2.5), and osteoporosis (MTS < -2.5). This model assumes that BMD can only deteriorate because this is the natural evolution of BMD over time, i.e., that T-scores are monotonically decreasing.

2.1 Notation

We denote by $\{X(a),\ a \in A\}$ the multi-state process of the disease progression model of interest with finite state space $\mathcal{S} = \{0, 1, 2\} = \{\text{Normal, Osteopenia, Osteoporosis}\}$. Therein, a refers to the patient's age at each DXA scan, $X(a) = s \in \mathcal{S}$ to the patient's state at age a, and $A = [18, \infty)$ to the set of possible ages. In addition, let $\mathcal{H}_{a-}$ be the process history over $[0, a)$ which includes, among other variables of interest, gender, risk group, and antiretroviral treatments.

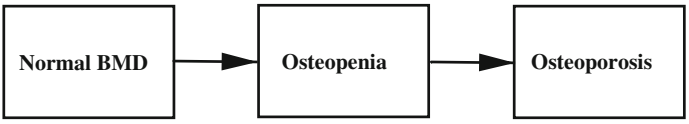


Fig. 1 Scheme of the disease progression model for bone mineral density

The model can be characterized by the transition probabilities (1) or the transition intensities (2):

$$P_{hj}(a,b;\mathcal{H}_{a-}) = P(X(b)=j|X(a)=h;\mathcal{H}_{a-}) = P_{hj}(a,b) \quad \text{(for short)}, \tag{1}$$

$$\alpha_{hj}(a) = \lim_{\Delta a \downarrow 0} \frac{P_{hj}(a, a+\Delta a;\mathcal{H}_{a-})}{\Delta a}, \tag{2}$$

where $18 \leq a < b$ and $h, j \in \mathcal{S} = \{0, 1, 2\}$.

The corresponding transition probability matrix of the disease progression model is

$$P(a,b) = \begin{pmatrix} P_{00}(a,b) & P_{01}(a,b) & P_{02}(a,b) \\ 0 & P_{11}(a,b) & P_{12}(a,b) \\ 0 & 0 & 1 \end{pmatrix}, \tag{3}$$

where the 0 cells indicate that no transition is possible between the corresponding two states, and $P_{22}(a,b) = 1$ because osteoporosis is an absorbing state.

We assume that the Markov property holds, that is, that the process after age a depends only on the state occupied at a. In addition, we assume a piecewise constant intensity model: $\alpha_{hj}(a) = \alpha_{hj}^m$ for $\theta_{m-1} < a \leq \theta_m$, $m = 1, \ldots, M$, where the θ_m provides an age partition. That is, disease progression is assumed to remain constant within and to vary between the age intervals.

2.2 Estimation Method

In order to fit the multi-state model, a follow-up of the patients is needed. In an ideal case, the exact transition times would be known. Here, however, progression times to osteopenia and osteoporosis are either right or interval-censored since osteopenia and osteoporosis onset cannot be determined exactly. A possible approach would be to fit the multi-state model using these interval-censored data.

A different approach we present herein consists of the following: (i) we use a longitudinal mixed model to fit the minimum T-scores (MTS) as a function of age and gender and to predict the MTS at each DXA scan; (ii) we estimate, for each patient, the times where the MTS are equal to -1 and -2.5; and (iii) we fit a disease progression model with the transition probability matrix (3) based on the imputed exact transition times:

(i) Denote by $Y(a)$ the true MTS pattern at age a of a given patient, and by Y_{ij} the observed j-th MTS for patient i. A linear mixed model considering random intercept and slope for variable age is fitted as follows:

$$Y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1}) \cdot \text{Age}_{ij} + \beta_2 \cdot \text{Gender}_i + \epsilon_{ij}, \tag{4}$$

where $i = 1, \ldots, 1293$, $j = 1, \ldots, n_i$, the random effects $\boldsymbol{b}_i$ are assumed $\mathcal{N}(\mathbf{0}, D)$, and the random error $\boldsymbol{\epsilon}_i$ is $\mathcal{N}(\mathbf{0}, \Sigma_i)$. The R package `lme4` from [2] used to fit this model provides parameter estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ as well as predicted values $\hat{b}_{i0}$ and $\hat{b}_{i1}$ for each patient.

(ii) Denoting by $\hat{Y}_{ij}$ the fitted MTS values for each patient, we consider patients with $\hat{Y}_{i1} \geq -1$ and $\hat{Y}_{i,n_i} < -1$, and for these we obtain by interpolation, following (4), the age at which the patient would have crossed the -1 boundary. Analogously, we proceed for osteoporosis and the -2.5 boundary.

(iii) The disease progression model based on (3) is fitted using both the estimated and the right-censored transition times of the patients with at least two DXA scans. To fit the model, we used the R package `msm` from [3], which provides maximum likelihood estimates of the transition probabilities $P_{hj}(a, b)$.

3 Results

3.1 Descriptive Analysis

The complete data set included a total of 3610 DXA scans from 1293 patients performed between 1999 and 2014. The mean (median; range) number of DXA scans per patient was 2.8 (2; 1 − 17); 55.2% ($n = 714$) of the patients had two or more DXA scans, for whom the mean (median; range) number of DXA scans per patient was 4.2 (3; 2 − 17). 73.7% of the patients were males and the mean age was 42.8 (42.3; 20.1 − 77.6) years at the first DXA scan. Among those patients with at least two DXA scans, 73.1% were males and mean age was 42.1 (41.5; 21.1 − 77.6) years.

3.2 Estimated Transition Probabilities

The fit of the linear mixed model (4), which was based on the data of all 1293 patients, provided the following parameter estimates: $\hat{\beta}_0 = -0.548$ (s.e.: 0.122); $\hat{\beta}_1 = -0.016$ (0.002); $\hat{\beta}_2 = -0.333$ (0.065). The negative signs of $\hat{\beta}_1$ and $\hat{\beta}_2$ indicate, respectively, that MTS decrease with age and that MTS are lower for men.

Next, the disease progression model was fitted using the data of the 714 patients with follow-up data. For the age partition, we chose $\theta_1 = 40$, and $\theta_2 = 50$ years. Figure 2 shows the estimated transition probabilities from normal BMD to osteopenia starting at different ages. For instance, the probability that a 40 years old man with normal BMD has osteopenia after 5 years is 0.12, that is, $\hat{P}_{01}(40, 45; \text{Male}) = 0.12$.

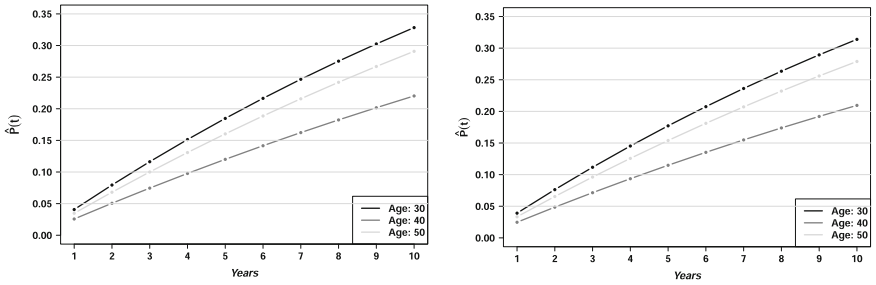


Fig. 2 Estimated transition probabilities from normal BMD to osteopenia among male (*left*) and female (*right*) HIV-infected persons

4 Discussion

The single imputation of the time to osteopenia and osteoporosis is certainly a limitation of this study. We plan to work with a multiple imputation model that will account for the randomness of the intercept and slope in the linear mixed model. The present approach is to be compared to a genuine method where the interval-censored nature of the data is taken into consideration. In addition, we plan to assess the impact of some specific antiretroviral treatments on bone loss by adding this new variable into the disease progression model. Finally, the data suggested a potential bidirectional disease progression model since the estimated slope of 28.4% of the patients was positive. This new model would take into account the potential recovery of bone loss due to the action of therapies and is in mind for future research.

Acknowledgements This work is partially supported by grants MTM2012-38067-C02-01 of the Spanish Ministry of Economy and Competitiveness and 2014 SGR 464 from the *Departament d'Economia i Coneixement de la Generalitat de Catalunya*.

References

1. NIH Osteoporosis and related bone diseases national resource center: bone mass measurement: what the numbers mean, available at http://www.niams.nih.gov/Health_Info/Bone/Bone_Health/bone_mass_measure.asp.
2. D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using `lme4`", *Journal of Statistical Software* **67**(1) (2015), 1–48.
3. C. Jackson, "Multi-state models for panel data: the msm package for R". *Journal of Statistical Software* **38**(8) (2011), 1–29.
4. E. Negredo, A. Bonjoch, M. Gómez-Mateu, C. Estany, J. Puig, N. Pérez-Álvarez, J. Rosales, S. di Gregorio, L. del Rio, G. Gómez, and B. Clotet, "Time of progression to osteopenia/osteoporosis in chronically HIV-infected patients: screening DXA scan", *PLoS One* **7**(10) (2012), e46031.

Statistical Challenges for Human Microbiome Analysis

Javier Rivera-Pinto, Carla Estany, Roger Paredes, M.Luz Calle, Marc Noguera-Julián and the MetaHIV-Pheno Study Group

Abstract DNA sequencing technologies have revolutionized microbiome studies. In this work we analyze microbiome data from an HIV study focused on the characterization of microbiome composition in HIV-1 infected patients. A 155 cohort of HIV infected and non-infected individuals is analyzed to characterize dietary and gut microbiome association in this group of patients. A penalized Dirichlet Multinomial regression model has been considered. The assumed underlying Dirichlet distribution in this modelization provides additional flexibility to the multinomial model which results in a better fit of the typically overdispersed microbiome data.

1 Introduction

Until recently, the composition and properties of the human microbiome were largely unknown, since the study was limited to in vitro cultivation of some specific microorganisms. Currently, high-throughput DNA sequencing technologies have revolutionized this field, allowing the study of the genomes of all microorganisms of a given environment. Metagenomics is the massive study of the genomes of the microorganisms and represents a breakthrough in the study of the relationship between the human microbiome and our health. The data from these studies provide valuable information about the composition and functional properties of microbial communities.

However, microbiome data analysis poses important statistical challenges. After DNA sequencing data analysis, microbiome data consists of a count matrix repre-

J. Rivera-Pinto (✉) · R. Paredes · M. Noguera-Julián
IrsiCaixa AIDS Research Institute, Barcelona, Spain
e-mail: jrivera@irsicaixa.es

C. Estany · R. Paredes
HIV Unit & Lluita contra la SIDA Foundation, Hospital Universitari Germans Trias i Pujol, Barcelona, Spain

R. Paredes
Universitat Autònoma de Barcelona, Catalonia, Spain

J. Rivera-Pinto · R. Paredes · M.Luz Calle · M. Noguera-Julián
University of Vic - Central University of Catalonia, Barcelona, Spain

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_8

senting the number of sequences corresponding to a specific bacterial taxa for each individual. Statistical techniques assuming the normal distribution are usually not appropriate. Instead, specific distributions for count data are required. An additional important feature of microbiome data is zero inflation (a large proportion of zero counts corresponding to taxa that are only present in some subjects) and the overdispersion in the rest of values. Since the total number of counts is not equal for every subject, there is the possibility of working with compositional data by dividing each count by the total number of counts giving the proportion that each taxa represents for each individual. In this case, appropriate methods for compositional data analysis are required.

In this work we analyze microbiome data from an HIV study focused on the characterization of microbiome composition along the different inflammatory profiles in healthy individuals and HIV-1 infected patients. HIV-linked chronic inflammation is associated with metabolic disorders, cardiovascular disease, immune senescence, premature aging and other inflammatory diseases. The role of the intestinal microbiome in these inflammatory processes has shown to be relevant. Interestingly, HIV infection clinical course, even when treated, is accompanied by an increase in gut permeability, bacterial translocation and low-level chronic inflammation. However, the precise effects of HIV-1 and related factor on the human gut microbiome are not well understood. It has been shown that diet has an important effect on gut microbiome composition; see [2, 6]. Therefore, it was important to characterize dietary-gut microbiome associations in this cohort. Available information was obtained from IrsiCaixa retrovirology laboratory, where microbiome and dietary information was collected from healthy and HIV infected patients showing different immune and inflammatory profiles and clinical outcomes.

First results of this project have been published in Noguera-Julián et al. [4].

2 Methods

Microbiome information was derived from 16s gene next generation sequencing from fecal samples of 155 subjects. Each one of them fulfilled both a nutrient and food portion independent diet questionnaires whose information was standardized (see Willet–Howe–Kushi [5]) to have total energy intake into account and apply the analysis over energy-relative information and not over raw data which could lead to erroneous conclusions. The standardization was made taking the residuals of a linear regression over total energy intake as new variable values.

The analysis of dietary-gut microbiome associations involves multivariate multiple regression between two matrices: $\mathbf{X}$, of size $n \times p$, and $\mathbf{Y}$, of size $n \times q$. Matrix $\mathbf{X}$ contains dietary information for p different nutrient and $\mathbf{Y}$ the microbiome abundance (*count data*) for q bacterial taxa, being n the total number of individuals.

The previously proposed penalized Dirichlet-Multinomial (DM) regression model (see [3]) was used to analyze the associations. This regression model addresses the overdispersion present in microbiome data by considering the DM distribution, with density function

$$f_{DM}(y_1, y_2, \ldots, y_q; \gamma) = \binom{y_+}{y} \frac{\Gamma(y_+ + 1)\Gamma(y_+)}{\Gamma(y_+ + \gamma_+)} \prod_{j=1}^{q} \frac{\Gamma(y_j + \gamma_j)}{\Gamma(\gamma_j)\Gamma(y_j + 1)}, \quad (1)$$

where $(y_1, \ldots, y_q)$ represents the counts for each genus, $y_+ = \sum_{j=1}^{q} y_j$, $\gamma = (\gamma_1, \ldots, \gamma_q)$ are parameters associated with the mean and variance of each genus, and $\gamma_+ = \sum_{j=1}^{q} \gamma_j$ is controling the degree of overdispersion, with a larger value indicating less overdispersion. In this modelization, the counts of the different taxa are assumed to follow a Dirichlet Multinomial distribution (see [1, 3]), which corresponds to a multinomial distribution

$$f_M(y_1, y_2, \ldots, y_q, \pi) = \binom{y_+}{y} \prod_{j=1}^{q} \pi_j^{y_j}, \quad (2)$$

with random underlying probability vectors following a Dirichlet distribution

$$f_D(\pi_1, \pi_2, \ldots, \pi_q; \gamma) = \frac{\Gamma(\gamma_+)}{\prod_{j=1}^{q} \Gamma(\gamma_j)} \prod_{j=1}^{q} \pi_j^{\gamma_j - 1}, \quad (3)$$

where $\pi = (\pi_1, \ldots, \pi_q)$ are the probabilities for a certain count to belong to the corresponding genus ($\sum_{i=1}^{q} \pi_i = 1$).

Penalized maximum likelihood estimation jointly performs model fitting and variable selection. As a result, the algorithm returns a matrix $\mathbf{C}$ of size $p \times q$, where c_{ij} represents the association between the i-th nutrient and the j-th genus. The penalization used in DM-regression assings zeroes to some coefficients selecting only the strongest associations.

3 Results

DM-regression provides the strongest associations between nutritional and genus composition information as a first step for deeper analysis. In the analyzed cohort, both *Prevotella* and *Bacteroides* are the genus with the strongest associations with nutrition parameters but in an inverse way. *Prevotella* is positively linked specially with *water* and *iron* and negatively associated with *saturated fat*. In the other hand, *Bacteroides* is negatively associated with *water* and *iron* (Fig. 1).

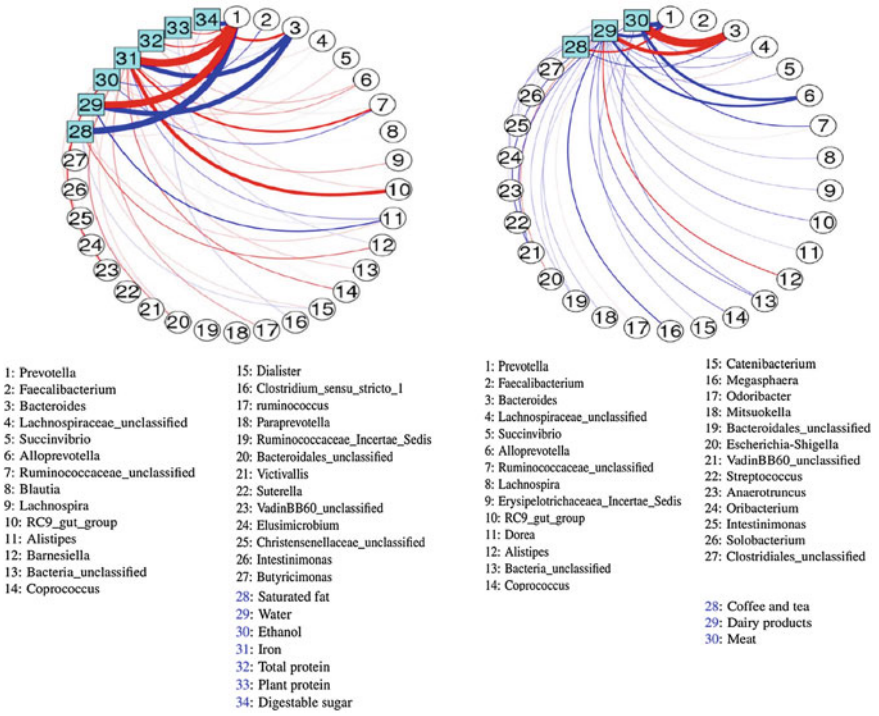


Fig. 1 Results with DM-regression model after penalization both for Nutrients (*left*) and Portions (*right*). *Red* lines represent positive relationship, while *blues* negative associations

4 Conclusions

DM-regression model allows to link two multivariate data matrices, one of them a count matrix. In this analysis those matrices where composed by genus counts after 16s rRNA sequencing and by the nutritional information of the individuals. DM distribution over the counts, has the overdispersion into account and links better with the nature of the data. In the other hand, the penalization included in the regression model selects only the strongest associations between genus and nutrients, allowing to the user to get more interpretable results.

Acknowledgements This study was mainly funded through philanthropy and private donations, which had no influence on its contents. Funds were obtained from a personal donation from Mr. Rafael Punter, the "Gala contra la SIDA" 2013 and 2014 editions, the "Nit per la Recerca a la Catalunya Central" 2015 edition, and by grant MTM2012-38067-C02-02 from the Ministerio de Economía e Innovación (Spain).

The first author is supported through a grant for doctoral studies from Noel Alimentaria to the University of Vic (UVic-UCC).

Other members of MetaHIV-Pheno Study Group were supported trough FI-DGR grant (FI-B00184) from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) at the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya.

References

1. J. Chen and H. Li, "Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis", *Annals of Applied Statistics* **7**(1) (2013), 418–442.
2. L. David *et al.*, "Diet rapidly and reproducibly alters the human gut microbiome", *Nature* **505** (2014), 559–563.
3. P.S. La Rosa, J.P. Brooks, E. Deych, E.L. Boone, D.J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, and W.D. Shannon, "Hypothesis testing and power calculations for taxonomic-based human microbiome data", *PLoS One* **7**(12) (2012), e52078.
4. Noguera-Julián *et al.*, "Gut microbiota linked to sexual preference and HIV infection", *Ebiomedicine* **5** (2016), 135–146.
5. W.C. Willet, G.R. Howe, and L.H. Kushi, "Adjustment for total energy intake in epidemiologic studies", *Am. J. Clin. Nutr.* **65**, (1997), 1220S–1228S; discussion 1229S–1231S.
6. G.D. Wu *et al.*, "Linking long-term dietary patterns with gut microbial enterotypes", *Science* **334** (2011), 105–108.

Integrative Analysis to Select Genes Regulated by Methylation in a Cancer Colon Study

Alex Sánchez-Pla, M. Carme Ruíz de Villa, Francesc Carmona, Sarah Bazzoco, and Diego Arango del Corro

Abstract Methylation is a regulatory mechanism known to be associated with tumour initiation and progression. Finding genes regulated by methylation is a first step to develop therapies that target these genes, for instance to inhibit tumor development. This study addresses this problem by comparing two methods, one based on mutual information, and a new one based on clustering the coefficients of fitted curves. The methods are tested on a Cancer Colon study and the biological analysis of the resulting lists suggests that at least some of the genes selected are indeed genes regulated by methylation, opening the door to an automatic mining method.

1 Introduction and Objectives

Methylation of CpG dinucleotides in the promoter of genes involved in the oncogenic process has been shown to be a key process contributing to tumor initiation and/or progression; see Sadikovic–Al-Romaih–Squire–Zielenska [5]. Finding genes regulated by methylation can lead to a better understanding of such processes and also be a guide to finding new drug targets.

This study originates in a work aiming at the identification of biomarkers for chemotherapy sensitivity in colorectal cancer (CRC). A panel of 50 cell lines derived

A. Sánchez-Pla (✉)
Statistics and Bioinformatics Unit, Vall d'Hebron Research Institute (VHIR), Barcelona, Spain
e-mail: asanchez@ub.edu

A. Sánchez-Pla · M.C. Ruíz de Villa · F. Carmona
Statistics Department, Universitat de Barcelona, Barcelona, Spain
e-mail: mruiz_de_villa@ub.edu

F. Carmona
e-mail: fcarmona@ub.edu

S. Bazzoco · D.A. del Corro
CIBBIM-Nanomedicine, Vall d'Hebron Research Institute (VHIR), Barcelona, Spain
e-mail: sarah.bazzocco@vhir.org

D.A. del Corro
e-mail: diego.arango@vhir.org

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_9

from colorectal tumors characterized by increasing sensitivity to several chemotherapy drugs was analyzed using different high-throughput data generation methods. Finding genes regulated by methylation was one of the approaches adopted in the search of candidate genes for new therapies.

In cancer–related genes it is relatively common to observe a decrease in gene expression associated with hypermethylation. Indeed, methylation is often described as a binary on-off signal (see Liu–Ji–Qiu [4]) that is, when methylation is "off" the gene can express normally and its expression will be intermediate or high, whereas when methylation is "on", the expression of the gene will be *repressed* and its values will tend to be low.

As a consequence of this *high-methylation/low-expression* and *low-methylation/high-expression* relation plots depicting methylation and expression will show L–shape patterns so the strategy adopted will be to mine such plots and select those that have such a shape.

The main objectives of this work are: (i) to select an appropriate method for scatterplot clustering that can be used to mine a multiple high-throughput dataset formed by expression and methylation data and extract the desired patterns, (ii) to test the methods selected on a colon cancer dataset formed by a panel of cell lines derived from colorectal tumors.

2 Methods for L-Pattern Selection

There have been published several methods to relate methylation and expression values. These range from simple correlation analysis (see Wagner–Busche–Ge–Kwan–Pastinen–Blanchette [6]) to more sophisticated approaches such as the one proposed by Liu and Qiu [3]. However, in spite of a certain agreement that the two magnitudes are negatively correlated, there is no generally accepted approach to select genes regulated by methylation. This work intends to be one more step into this direction.

2.1 Gene Selection Based on Conditional Mutual Information

When studying methylation, we are faced with two main questions: (i) which genes exhibit an L-shape, and (ii) what is the optimal threshold for binarizing methylation data for each L-shape gene.

Liu–Qiu [3] suggests to determine whether methylation and expression of a gene exhibit an L-shape by computing the conditional Mutual Information (MI) for different choices of the threshold adopted to binarize the methylation data.

If we consider the continuous valued methylation and expression data as two random variables X and Y, and denote a nominal threshold as t, the conditional MI can be written as a weighted sum of MIs on the two sides of the threshold:

$$cMI(t) = I(X, Y|X > t)P(X > t) + I(X, Y|X \leq t)P(X \leq t).$$

For an L-shape gene, as t moves from 0 to 1, $cMI(t)$ first decreases and then increases, and its value approaches zero when t coincides with the reflection point.

The ratio $r = \min\{cMI(t)\}/cMI(0)$ for an L-shape gene is small, and the *optimal threshold* for dichotomizing the methylation data of this gene is $t^* = \text{argmin}\{cMI(t)\}$.

To estimate the MI terms we use a kernel-based estimator, which constructs a joint probability distribution by applying a Gaussian kernel to each data point:

$$I(X, Y) = \frac{1}{M}\sum_{i=1}^{M} \log \frac{M\sum_{j=1}^{M} e^{-\frac{1}{2h^2}((x_i-x_j)^2+(y_i-y_j)^2)}}{\sum_{j=1}^{M} e^{-\frac{1}{2h^2}(x_i-x_j)^2}\sum_{j=1}^{M} e^{-\frac{1}{2h^2}(y_i-y_j)^2}},$$

where h is a tuning parameter for the kernel width and empirically set $h = 0.3$.

2.2 Gene Selection Based on Spline Regression

The above approach is appealing but previous studies suggest that it works best when the number of samples is very big —perhaps hundreds or even thousand samples. This is a common sample size when working for example with TCGA samples [7], but not for individual experiments. As an alternative, we suggest to fit a curve to each scatterplot, that is to the relation between expression and methylation for each gene, and then cluster these lines and keep those clusters that can be associated with an L-pattern.

The relation between expression and methylation is weak and non-linear, so a reasonable option for modelling this type of data is *splines regression* a form of non-parametric regression that automatically models non-linearities; see Hastie–Tibshirani–Friedman [2]. *Splines* are continuous functions formed by connecting linear segments. The points where the segments connect are called the *knots* of the spline. A particularly efficient form of splines regression is B-splines [2], where the splines are B_{mp} p-th order polynomial of degree $p - 1$ with finite support over the interval and 0 everywhere else.

With this representation we have applied the following algorithm to select genes regulated by methylation:

(i) prefilter genes to be fitted, for instance select those having a significantly negative Spearman correlation coefficient;
(ii) fit a cubic regression spline to each gene and extract the spline coefficients;

(iii) use coefficients to compute a distance matrix based on a "1-correlation" distance;
(iv) perform hierarchichal clustering on this distance matrix;
(v) select clusters that visually adapt to an L–shape.

3 Results and Application: Selecting L-Shaped Genes from a Genome-Wide Analysis of Colorectal Cancer

We have applied the methods described above to the experimental data obtained from an ongoing CRC study; see Bazzocco–Alazzouzi–Ruiz de Villa–Sánchez-Pla–Mariadason–Arango [1]. The data analyzed consisted of expression and methylation values obtained respectively from Affymetrix (hgu133plus2 expression microarrays) and Illumina (256K methylation arrays). Expression and methylation data do not have a one to one correspondence so they were preprocessed separately, using standard approaches for these types of data, and then aggregated on a gene basis so they could be matched. This process yield two 30 (samples) $\times$ 11746 (genes) arrays.

3.1 Results Using the Conditional Mutual Information Approach

The data were processed using the algorithm for finding the optimal binarization threshold described above, and genes with L-shape were selected using a combination of three criteria:

(i) genes had "small" ratio between conditional mutual and overall mutual information; this was set to $r = cMI/MI < 0.25$;
(ii) the minimum value of overall mutual information was at least 0.1, that is, $cMI(0) > 0.1$;
(iii) the median expression on the left side of the optimal threshold t^* had to be higher than median expression on the right side.

Applying the above criteria yield a total of 641 genes that could be considered to have a L-shape.

3.2 Results Using Splines Regression to Select Genes

Splines regression cannot be applied to all the genes so a prefiltering step was used, and only genes showing a significant negative Spearman correlation were modelled to avoid an excess of noise that would negatively afect clustering later. A heuristic

filter was also applied to guarantee non L-shape removal. Overall, this led to keep 191 genes for which splines were fitted and clustered into 5 clusters. The first two, majoritary, clusters included 162 genes that could be considered to have a L-shape.

There were a total of 98 genes in common selected by the two methods.

4 Discussion and Conclusions

This study can still be considered preliminary but a certain number of consistent results can be highlighted:

(i) cMI based gene selection provides an intuitive approach for selecting L–shaped patterns, although it can yield a certain number of "false positives". The method, however, works well with big (hundreds) samples which makes it less reliable for normal-size (dozens) datasets.
(ii) Clustering based on the results of Splines regression is also useful in detecting L–shaped patterns. While it selects a smaller number of genes than cMI, it is not so dependent from sample size.
(iii) Biological interpretation is still ongoing but the results are consistent with the hypothesis that is, genes known to be regulated by methylation have been found with both methods.

Acknowledgements The first and third author wish to acknowledge the *Grup de Recerca en Bioestadística i Bioinformàtica* (GRBIO, Grup Consolidat 2014-SGR-464 Generalitat de Catalunya), for funding part of this work.

References

1. S. Bazzocco, H. Alazzouzi, M.C. Ruiz de Villa, A. Sánchez-Pla, J.M. Mariadason, and D. Arango, "Genome-Wide Analysis of DNA Methylation in Colorectal Cancer", *Preprint* (2016).
2. T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: data mining, inference, and prediction", Springer, New York, second edition edition, 2009.
3. Y. Liu and P. Qiu, "Integrative analysis of methylation and gene expression data in TCGA", in "Genomic Signal Processing and Statistics (GENSIPS)" (2012), 1–4.
4. Y. Liu, Y. Ji, and P. Qiu, "Identification of thresholds for dichotomizing DNA methylation data", *EURASIP Journal on Bioinformatics and Systems Biology* **1** (2013), 8.
5. B. Sadikovic, K. Al-Romaih, J.A. Squire, and M. Zielenska, "Cause and consequences of genetic and epigenetic alterations in human cancer", *Current Genomics* **9**(6) (2008), 394–408.
6. J.R. Wagner, S. Busche, B. Ge, T. Kwan, T. Pastinen, and M. Blanchette, "The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts", *Genome Biology* **15**(2) (2014), R37.
7. J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Mills Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J.M. Stuart, "The cancer genome atlas research network, (the cancer genome atlas pan-cancer analysis project)", *Nature Genetics* **45**(10) (2013), 1113–1120.

Topological Pathway Enrichment Analysis of Gene Expression in High Grade Serous Ovarian Cancer Reveals Tumor-Stoma Cross-Talk

Oana A. Zeleznik, Gerhard G. Thallinger, John Platig, and Aedín C. Culhane

Abstract Identifying the biological pathways that are significantly regulated in a given condition is a fundamental step to understanding biological phenomena. Existing pathway approaches were designed for the analysis of a single dataset and are not optimized for simultaneous analysis of multiple data sources. Increasing availability of multiple omics datasets obtained on the same sample allows for a more complete understanding of pathway behavior in human diseases. We propose a pathway analysis approach in which we integrate multiple molecular datasets using multivariate analysis and apply dynamical importance to extract topology-based pathway scores.

1 Introduction

Traditional single dataset gene set or pathway analysis often reduces the pathway network to a simple flat list of genes, ignoring biological knowledge of pathway topology, protein complexes, and functional non-equivalence of genes. While there are methods considering the rank of genes in a gene list, many weight genes equally and use variations of Fisher's Exact Test to calculate enrichment of genes in a pathway; see Khatri–Sirota–Butte [4]. A few network topology-based approaches have emerged, but these are computationally intensive, limiting in their application, and

O.A. Zeleznik · G.G. Thallinger
Computational Biotechnology and Bioinformatics, Institute of Molecular Biotechnology, Graz University of Technology, Petersgasse 14/V, 8010 Graz, Austria

O.A. Zeleznik · G.G. Thallinger
BioTechMed Omics Center Graz, Stiftingtalstrasse 24, 8010 Graz, Austria

J. Platig · A.C. Culhane (✉)
Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02115, USA
e-mail: aedin@jimmy.harvard.edu

J. Platig · A.C. Culhane
Biostatistics, Harvard T.H. Chan School of Public Health, Huntington Ave, Boston, MA 02115, USA

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_10

may not outperform simple gene lists approaches; see Bayerlová–Jung–Kramer–Klemm–Bleckmann–Beißbarth [1].

We present an integrative, network-based and pathway-centric gene set enrichment (GSE) approach which we call Integrative Pathway Enrichment Analysis (IPEA). It combines features (genes, proteins, metabolites, etc.) from multiple molecular datasets using a multivariate latent variable analysis (Multiple Co-Inertia Analysis, [5]) and correlates them with their importance scores from each biological pathway in the Reactome database [3]. The topological importance of features in a pathway is quantified by the dynamical importance score; see Restrepo–Ott–Hunt [7]. We apply this analysis to discover pathways regulated in tumor and stroma samples [6] from high grade serous ovarian cancer.

2 Methods

IPEA is structured in three steps. First, we integrate features from multiple datasets using a latent variable approach called Multiple Co-Inertia Analysis (MCIA); see Meng–Kuster–Culhane–Moghaddas-Gholami [5]. MCIA reduces the features of two or more omics datasets into the same lower dimension while maximizing the squared covariance between the eigenvectors of the initial datasets and the new co-inertia axes.

Second, features are scored based on their contribution to the information flow within pathways in the network. We reward highly linked hubs and bottleneck nodes which may have few connections but bridge different clusters within a network; see Restrepo–Ott–Hunt [7]. The information flow scores of genes in a pathway are defined here by their dynamical importance (DI) which was shown to well characterize the importance of nodes in a network. The dynamical importance I of node k, denoted I_k, is defined as the change (Δ) in the largest eigenvalue λ of the corresponding network adjacency matrix upon removal of node k, i.e., $I_k \equiv \Delta_k/\lambda$.

Finally, pathway enrichment scores are calculated by the Spearman correlation between the information flow scores of each feature in each pathway and the loadings of each feature on each principal component (PC) of the integrative MCIA analysis. Enrichment scores were calculated separately for the negative and for the positive side of each PC resulting in up- and down-regulated pathways.

3 Results

IPEA was applied to gene expression profiles of 20184 genes in paired microdissected tumor ($n = 38$) and stroma ($n = 38$) samples from high-grade serous ovarian cancer [6] to discover features important in tumor-stroma cross-talk.

First, the gene expression profiles were integrated using MCIA to extract covariant features among the datasets. Inspection of the space spanned by the first two MCIA PCs, which explained 27% of the total variance, revealed samples which had similar

tumor and stroma gene expression profiles, while other samples had discordant tumor and stroma profiles. The stroma profiles (standard deviation of 0.42 on PC1 and 0.34 on PC2; mean equal to zero on both PCs) varied more than the tumor profiles (standard deviation of 0.24 on PC1 and 0.28 on PC2; mean equal to zero on both PCs). In general we observed more diversity in stromal tissue, possibly due to greater heterogeneity of infiltrating cells.

Next, the scores of the tumor and stroma genes on the first MCIA axis were correlated with the corresponding dynamical importance scores from Reactome. The enriched pathways are displayed in Fig. 1 as a double bipartite-like graph: one can distinguish between tumor and stroma but also between up- and down-regulated pathways/genes. To facilitate the interpretation of the result, only enriched pathways which include genes with MCIA scores higher than the 75% quantile are shown.

Additionally, clinically actionable genes (drug targets) are displayed as triangles on the network of MCIA selected genes and enriched pathways. Clinically actionable genes were extracted from the TARGET[1] (tumor alterations relevant for genomics-driven therapy) database of genes that, when somatically altered in cancer, are directly linked to a clinical action in that they might be predictive of response or resistance to therapy. Four actionable target genes were present in the resulting network: CTNNB1 (catenin beta 1), ERBB4 (erb-b2 receptor tyrosine kinase 4), SMAD4 (SMAD family

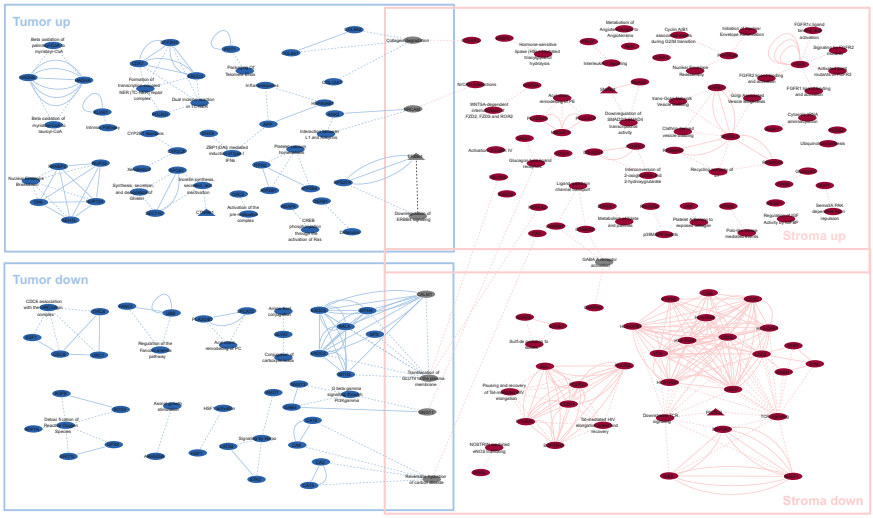


Fig. 1 Enriched pathways and corresponding genes resulting from IPEA computed from the first MCIA axis of the tumor and stroma datasets. Genes/pathways that are active in tumor are displayed as *blue ellipses* while genes/pathways that are active in stroma are displayed as *red ellipses*. *Gray ellipses* represent genes/pathways that are active in tumor and stroma. Each enriched pathway is represented by an ellipse which is linked to the genes belonging to it by *dashed lines*. *Solid lines* link genes that belong to the same pathway. Notably, there are no edges linking tumor up to tumor down nor stroma down to tumor up

[1]TARGET version 3.0 was downloaded from https://www.broadinstitute.org/cancer/cga/target.

member 4) and PIK3CB (phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit beta).

The network discovered by IPEA analysis shows that DNA repair pathways are highly regulated in ovarian tumors. We observed up-regulation of *Formation of transcription coupled NER (TC-NER) repair complex*, *Dual incision reaction in TC-NER* and the down-regulation of the *Fanconi Anemia pathway* in tumors. Regulation of DNA repair in tumors was associated (Fig. 1) with increased regulation of immune response pathways in tumor and in stroma: up-regulation of *Inflammasomes* and *ZBP1 (DAI) mediated induction of type I IFNs* in tumor, up-regulation of *Interleukin-1 signaling* and down-regulation of *TCR signaling* in stroma samples.

4 Discussion

We present a method for integrative pathway analysis of multiple molecular datasets. The result of IPEA is a network highlighting enriched pathways and activated genes.

An advantage of IPEA over traditional GSE approaches is its ability to account for the topology in biological pathway networks. We apply it to known pathway networks from Reactome, but the network topology could be experimentally or computational determined also. While we use MCIA as a multivariate latent variable analysis step of IPEA, any other multivariate analysis method (multiple factor analysis, etc.) may be used. MCIA was chosen due to its ability to capture covariant features between the datasets. Indeed, IPEA could potentially be adapted to any list of features that are ranked, e.g., differentially expressed genes or proteins. This makes IPEA versatile and suitable for a wide range of applications.

We apply IPEA to study ovarian cancer, a leading cause of cancer death in women world-wide. Most women are diagnosed with advanced stage disease and consequently have a poor probability of survival after five years. The poor outcome is attributed to the complex nature of this disease. It had been difficult to define molecular subtypes of high grade serous ovarian cancer; see [2, 8, 9]. One possible reason may be the role played not only by the tumor itself but by the tumor microenvironment, the stroma. Only a few gene expression studies microdissect tumor tissue. The proportion of stroma varies considerably both within and between studies, introducing new variance which may prevent a rigorous phenotype characterization.

We applied IPEA to the gene expression profiles of microdissected tumor and stroma to discover pathway cross-talk between tumor and its microenvironment. Therefore in this case study, we display results as a double bipartite-like graph. We showed up- and down-regulated pathways (DNA repair pathways, immune pathways) in high grade serous ovarian cancer tumor and stroma. Actionable target genes superimposed on the resulting network identified actionable genes in both tumor and stroma, providing further support that investigation of stroma gene expression and the cross-talk between tumor and stroma is needed in ovarian cancer. The new pathways together with the target genes have to be further investigated and may characterize new targets in ovarian cancer.

5 Conclusion

We introduced IPEA, a new topology-based pathway analysis approach, and applied it to investigate the complex cross-talk between biological pathways in tumor and the tumor microenvironment of high grade serous ovarian cancer.

References

1. M. Bayerlová, K. Jung, F. Kramer, F. Klemm, A. Bleckmann, and T. Beißbarth, "Comparative study on gene set and pathway topology-based enrichment methods", *BMC Bioinformatics* **16**(1) (2015), 1.
2. S. Bentink, B. Haibe-Kains, T. Risch, J.B. Fan, M. Hirsch, *et al*., "Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer", *PloS One* **7**(2) (2012), e30269.
3. D. Croft, G. OKelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, *et al*., "Reactome: a database of reactions, pathways and biological processes", *Nucleic Acids Research* **39** (2011), D691–D697.
4. P. Khatri, M. Sirota, and A.J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges", *PLoS Computational Biology* **8**(2) (2012), e1002375.
5. C. Meng, B. Kuster, A. Culhane, and A. Moghaddas-Gholami, "A multivariate approach to the integration of multi-omics datasets", *BMC Bioinformatics* **15**(1) (2014), 162–175.
6. S.C. Mok, T. Bonome, V. Vathipadiekal, A. Bell, M.E. Johnson, D.C. Park, K. Hao, D. Yip, H. Donninger, L. Ozbun, *et al*., "A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2". *Cancer Cell* **16**(6) (2009), 521–532.
7. J. Restrepo, E. Ott, and B. Hunt, "Characterizing the dynamical importance of network nodes and links", *Physics Review Letters* **97**(9) (2006), 094102.
8. The Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma", *Nature* **474**(7353) (2011), 609–615.
9. R.W. Tothill, A.V. Tinker, J. George, R. Brown, S.B. Fox, *et al*., "Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome", *Clinical Cancer Research* **14**(16) (2008), 5198–5208.

Part II
Statistics for Low Dose Radiation Research

Foreword

Uncertainties, both quantitative and conceptual in nature, have been identified as key to addressing the remaining research questions in EU low dose radiation research. From October 26 to 28 2015, EU FP7 DoReMi project collaborators from the UK Public Health England and the Spanish Centre for Research in Environmental Epidemiology (CREAL), together with colleagues from Universitat Autònoma de Barcelona (UAB) and Durham University (DU), organized a workshop to bring together researchers from the low dose radiation fields and invited expert mathematicians and statisticians with an interest in applied uncertainty analysis. The key topics of the *LD-RadStats: Workshop for statisticians interested in contributing to EU low dose radiation research* meeting were radiation biology and biomarkers of dose and effect, epidemiological elucidation of health risks at low doses, and modelling including the dose response relationship. The overall aims of the workshop were to work towards ensuring that state-of-the-art mathematical and statistical techniques are fully exploited within EU low dose radiation research.

In this special issue of Research Perspectives CRM Barcelona subseries of the Birkhuser's series Trends in Mathematics, we present eleven extended abstracts from the workshop, representing an overview of mathematical and statistical analysis techniques currently in use across the radiation research disciplines or which may help to take the research forward. Representatives from the field of radiation biology present models for consideration of traditional chromosomal biodosimetry, gene expression or general multi-level omics data analysis. Appropriate model definition and selection are recurring themes here. The complex task of appropriately characterizing uncertainties for epidemiological risk analyses is addressed through mechanistic analysis of lung cancer mortality and through provision of a new R package to integrate ERR modelling. A case study of modelling plaque overlap in radiation induced atherosclerosis further demonstrates the need for a very good mechanistic understanding to support modelling, and a general summary of analytic and stochastic approaches processes details some tools for modelling radiation induced carcinogenesis.

Before the closing of the workshop, the attendees separated into three breakout groups to discuss the way forward. In conclusion, it was decided that uncertainty analysis does require greater representation in EU radiation research, and that this can be achieved by all attendees ensuring that the issues continue to be highlighted during projects and in the wider field. The attendees agreed to create an informal network of individuals interested in uncertainty analysis in low dose radiation research, LDRadStatsNet. And it is hoped that this will also prove beneficial to these aims. The organizers also hope presentation of these extended abstracts here will further stimulate discussion and collaboration within the field and we look forward to working with all our colleagues on these exciting topics.

Finally, we would like to express our thanks to DoReMi and the Centre de Recerca Matemàtica CRM-Barcelona for funding the workshop, and say a huge thank you to our colleagues at CREAL for helping to organize and host this extremely productive meeting.

July 2016

Didcot, England — Elizabeth Ainsbury
Barcelona, Spain — Elisabeth Cardis
Barcelona, Spain — Pere Puig
Durham, England — Jochen Einbeck

Biological Dosimetry, Statistical Challenges: Biological Dosimetry After High-Dose Exposures to Ionizing Radiation

Joan Francesc Barquinero and Pere Puig

Abstract A statistical model to deal with low and high-dose exposures is presented. The model is based on a weighted Poisson distribution which allows to explain the underdispersion observed in the empirical data. A Gompertz type calibration curve is also introduced.

When a radiological accident occurs, it is very important to estimate the dose of ionizing radiation (IR) received to guide medical care. If physical measurements are not available or it is suspected that dosimeters have not been used correctly, biological dosimetry methods are necessary for a precise dose-assessment. Within the different methodologies, the most widely used is to score dicentric chromosomes in metaphases of peripheral blood lymphocytes. This method accurately estimates doses in cases of acute and recent exposures; see [1]. Currently, the majority of dose-effect curves for dicentric chromosomes include doses from 0 to 5 Gy. For this dose range and for low LET radiation types, such as X and gamma rays, the dose-effect relationship fits well to a linear-quadratic model. Additionally after whole body exposure from 0 to 5 Gy the distribution of dicentrics among cells agrees with the Poisson distribution, allowing the detection of partial body exposures when deviations of the Poisson are detected; see [3, 10].

Some accidents have demonstrated the need to evaluate exposures to high doses and if they are whole or partial body exposures; see [4, 11, 12]. Not only because they occur but also for the improvements reached in medical care after IR over-exposures [2], and the development of acute radiation syndrome mitigators [6, 9]. However, the dicentric based biodosimetry is not suitable for doses of IR higher than 5 Gy, because the number of cells able to reach metaphase decreases dramatically when the dose increases. After a high dose exposure heavily damaged cells,

J.F. Barquinero (✉)
Departament de Biologia Animal, Vegetal i Ecologia, Universitat Autònoma de Barcelona, Barcelona, Spain
e-mail: Francesc.Barquinero@uab.cat

P. Puig
Departament de Matemàtiques, Universitat Autònoma de Barcelona, Barcelona, Spain
e-mail: ppuig@mat.uab.cat

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_11

show a delay or even the impossibility of progressing through the G2/M cell cycle checkpoint to reach mitosis; see [1]. A way to overcome this problem is inhibiting this checkpoint using a caffeine treatment; see [7, 9]. Here, we show the analysis of dicentric chromosomes after irradiating at doses from 0 to 25 Gy.

As expected, a clear increase in the frequency of dicentrics was observed as the dose increased (Table 1). The agreement of dicentrics cell distribution with Poisson, tested by the normalized unit U of the dispersion index, was not rejected for six of the ten doses evaluated. However, in all cases U values were negative and, for 3, 5, 7 and 10 Gy, U values were significantly underdispersed.

Another observed result was that the frequency of dicentrics tend to saturate at highest doses. At higher doses, fewer cells with an elevated number of dicentrics were observed. This is mainly due by the limited number of chromosomes, which in human lymphocytes is 46. The theoretical maximum of possible dicentrics are 23. In addition, from 5 G to 25 Gy, the number of cells without dicentrics was lower than expected from the Poisson distribution, this last phenomena probably due to a major misrejoining probability at higher doses. The saturation and the low number of cells without dicentrics would contribute to the observed underdispersion.

A new count probability function has been considered to model our underdispersed count data, having the form

$$P(k; b, \lambda) = \frac{1 + bk^2}{1 + b(\lambda + \lambda^2)} \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k = 0, 1, 2, \ldots \tag{1}$$

This is a specific weighted Poisson distribution with a weight equal to $w(k) = 1 + bk^2$, representing the sighting mechanism. The domain of the parameters is $b \geq 0$, $\lambda > 0$, and for $b = 0$ this is just the Poisson probability function. It is immediate to verify that changing the values of the parameters b, λ, the dispersion index can take values slightly greater than 1 or values lower than 1. Therefore, the probability distribution described in (1) is useful to model count data presenting underdispersion, like that observed in our empirical distributions.

Taken into account that in biological dosimetry dose-effect calibration curves for dicentric chromosomes are linear or linear-quadratic models, the challenge was to consider parameter λ in (1) to be dependent of the dose d, using a Gompertz type curve of the form, $\lambda(d) = \beta_0 e^{-\beta_1 e^{-\beta_2 d}}$, where β_0, β_1, β_2 are suitable parameters to be estimated from the data. Moreover, parameter b in (1) must also be considered depending of the dose, and following a simple linear relationship, $b(d) = \beta_3 d$, where β_3 is another parameter. Our Gompertz type curve is very flexible, having a sigmoid profile very suitable to fit our empirical data.

The maximum likelihood method has been used to estimate the four parameters of the model. The details and an R program to fit he data can be found in [8]. This model can be also applied to partial body irradiation problems, as it is described in [8]. Gomperz type model can be also used under the Poisson assumption (in (1)), leading to a three parameter model. To fit the data in this situation, the RADIR package is a Bayesian-based suitable tool that can be downloaded from CRAN repository; see [5].

Table 1 Dicentrics observed after γ-irradiation at doses from 0 to 25 Gy (dose rate of 5.25 Gy/min). For each dose, the number of cells scored and the distribution of dicentrics among cells is indicated. Gy = Gray; Total dic = total number of dicentrics; Y = frequency of dicentrics; U = values of the u-test; u values lower than −1.96 are indicative of a significant underdispersion

Dose (Gy)	Cells	Dicentrics distribution among cells																			Total Dic	Y	U
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
0	2000	1999	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.00	–
0.1	2000	1989	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0.01	−0.17
0.5	2000	1922	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78	0.04	−1.23
1	1000	886	108	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	120	0.12	−0.43
3	500	213	192	85	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	393	0.79	−2.91
5	150	3	23	58	38	15	10	2	1	0	0	0	0	0	0	0	0	0	0	0	382	2.55	−3.29
7	150	0	4	23	35	35	29	10	9	4	1	0	0	0	0	0	0	0	0	0	604	4.03	−2.85
10	150	0	0	0	3	18	40	35	25	16	9	4	0	0	0	0	0	0	0	0	915	6.10	−5.11
15	100	0	0	0	0	3	10	12	21	10	16	7	7	7	3	1	3	0	0	0	834	8.34	−1.31
20	100	0	0	0	0	0	6	9	10	12	17	13	9	6	8	6	1	1	2	0	967	9.57	−1.17
25	100	0	0	0	0	0	4	5	5	8	18	16	12	7	3	9	4	3	3	3	1065	10.65	−0.42

References

1. International atomic energy agency (2011) cytogenetic dosimetry: applications in preparedness for and response to radiation emergencies. Vienna.
2. G.H. Anno, R.W. Young, R.M. Bloom, and J.R. Mercier, "Dose response relationships for acute ionizing-radiation lethality", *Health Physics* **84**(5) (2003), 565–575.
3. J.F. Barquinero, L. Barrios, M.R. Caballín, R. Miró, M. Ribas, *et al*., "Biological dosimetry in simulated in vitro partial irradiations", *Int. J. Radiat Biol*. **71** (1977), 435–440.
4. I. Hayata, R. Kanda, M. Minamihisamatsu, M. Furukawa, and M.S. Sasaki, "Cytogenetical dose estimation for three severely exposed patients in the JCO criticality accident in Tokai-mura", *J. Radiat Res.* **42** (2001), 149–155.
5. D. Moria, M. Higueras, P. Puig, E.A. Ainsbury, and K. Rothkamm, "Radir package: an R implementation for cytogenetic biodosimetry dose estimation", *Journal of Radiological Protection* **35**(3) (2015), 557–569.
6. R. Patil, E. Szabó, J.I. Fells, A. Balogh, K.G. Lim, Y. Fujiwara, D.D. Norman, S.C. Lee, L. Balazs, F. Thomas, S. Patil, K. Emmons-Thompson, A. Boler, J. Strobos, S.W. McCool, C.R. Yates, J. Stabenow, G.I. Byrne, D.D. Miller, and G.J. Tigyi, "Combined mitigation of the gastrointestinal and hematopoietic acute radiation syndromes by an LPA2 receptor-specific nonlipid agonist", *Chem. Biol.* **22**(2) (2015), 206–216.
7. M. Pujol, R. Puig, M.R. Caballín, L. Barrios, and J.F. Barquinero, "The use of caffeine to assess high dose exposures to ionising radiation by dicentric analysis", *Radiat. Prot. Dosimetry* **149** (2012), 392–398.
8. M. Pujol, J.F. Barquinero, P. Puig, R. Puig, M.R. Caballín, and L. Barrios, "A new model of biodosimetry to integrate low and high doses", *PLoS One* **9**(12) (2014), e114137.
9. R. Rowley, M. Zorch, and D.B. Leeper, "Effect of caffeine on radiation-induced mitotic delay: delayed expression of G2 arrest", *Radiat Res.* **97** (1984), 178–185.
10. V.A. Vinnikov, E.A. Ainsbury, N.A. Maznyk, D.C. Lloyd, and K. Rothkamm, "Limitations associated with analysis of cytogenetic data for biological dosimetry", *Radiat Res.* **174** (2010), 403–414.
11. B. Yao, B.R. Jiang, H.S. Ai, Y.F. Li, G.X. Liu, *et al*., "Biological dose estimation for two severely exposed patients in a radiation accident in Shandong Jining, China, in 2004". *Int. J. Radiat Biol.* **86** (2010), 800–808.
12. B. Yao, Y. Li, G. Liu, M. Guo, J. Bai, *et al*., "Estimation of the biological dose received by five victims of a radiation accident using three different cytogenetic tools", *Mutat Res.* **751** (2013), 66–72.

Heterogeneous Correlation of Multi-level Omics Data for the Consideration of Inter-tumoural Heterogeneity

Herbert Braselmann

Abstract In integrative radiation systems biology, relationships between variables generated from different molecular levels are investigated. Two approaches to detect correlations in subsets of bivariate continuous data are discussed. The approaches are based on two-component finite Gaussian mixture models and on parametric bootstrap of the null-hypothesis to generate a reference distribution of the likelihood ratio statistic.

1 Introduction

1.1 Radiobiological Background

There is good evidence that exposure to low doses of ionizing radiation increases cancer risk [1]. One topic in the research field of radiocarcinogenesis is the identification of genetic biomarkers in tumors of radiation exposed cohorts which can be used in epidemiological risk modelling. Among the primary radiation damages in cells of organic tissue occur chromosomal translocations, mutations or DNA copy number alterations which can be transcribed to the RNA and affect the expression of proteins. In order to corroborate the identification of candidate biomarkers, the discipline of integrative radiation systems biology investigates causal relationships between different molecular levels; see [11]. For e.g., the molecular transcriptome (mRNA) and miRNA levels in tumor cells are measured by expression microarrays which, after data preprocessing and normalization, yield to continuously distributed random variables. Relationships between variables are statistically assessed by correlation tests. It is experimentally known that miRNAs are involved in downregulation of gene expression; see [4]. This is reflected by inverse relationships between paired expression data from miRNA and their target genes; see [9]. An overexpressed miRNA in cells of a potentially radiation exposed tumor should correspond to a low

H. Braselmann (✉)
Research Unit Radiation Cytogenetics, German Research Center for Environmental Health, Neuherberg, Germany
e-mail: braselm@helmholtz-muenchen.de

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_12

expression value of its target gene. Correlation analysis between miRNA and mRNA for a candidate gene works then as a procedure for target validation.

1.2 Heterogeneous Gene Expression

Genes are possibly expressed in only a part of the cases or expressed at different levels among the cases in a statistical sample of tumors of a defined type. This decreases the power of standard two-sample comparison tests for differential gene expression as well as of correlation tests. Heterogeneity in random variables can mathematically be described by mixture models of suitable distributions. Van Wieringen et al. [12] proposed a nonparametric two-component mixture model for univariate testing of partial differential expression. In that model, the distribution H in the test group (here, exposed group) is thought as a mixture of a distribution F in the control group with subpopulation size $1-\tau$ and a shifted distribution $G \leq F$ with subpopulation size τ. A suitable weighted distance function (for e.g., L_2-distance) between the functions $(1-\theta)F$ and H is then minimized for θ, which serves as a test statistic. The population parameter θ was inspired by the theory on the estimation of mixing proportions using minimum distance estimators; see [10]. The test is implemented as a permutation test, replacing F and H by its empirical cumulative distributions.

The correlation or association between two continuous and heterogeneous molecular biological variables x and y could, in principle, be reduced to the univariate partial differential expression test when one of the two variables is taken as fixed and dichotomized, for e.g., by the median. However, it would be desirable to use perhaps more powerful bivariate correlation tests such as Pearson or Spearman. A distribution-shift model as in the univariate case does not hold here. As a first step following that line, it will be necessary to divide the data in the x-y-plane into at least two clusters, one in which the data shows correlation or covariance, and no correlation or opposite directed correlation in the other. In this presentation some possible strategies will be discussed without giving final solutions or mathematical proofs.

2 Normal Mixture Models

As a first approach, we consider bivariate normal mixture models to cluster data in the x_1-x_2-plane. These kind of models are fitted with help of the EM (Expectation Maximization) algorithm of Dempster-Laird-Rubin [2]. The number of components and shape parameters in the these models is often optimized with the BIC (Bayesian Information Criterion). However, for an inferential test, a null-hypohtesis H_0 and an alternative hypothesis H_1 with fixed parameter sets have to be formulated. In the simplest case with one against two normal components and centered x, i.e., $E(x) = 0$, we set H_0 to be

$$f_0(x, \theta) = N(0, \Sigma)$$

and H_1 to be

$$f_1(x, \theta) = (1 - \pi)N(\mu_1, \Sigma_1) + \pi N(\mu_2, \Sigma_2),$$

where $x \in \mathbb{R}^2$, $N(\mu_i, \Sigma_i)$ is the bivariate Gaussian density with mean value μ_i and variance-covariance matrix Σ_i, $i = 1, 2$, $\pi \in [0, 1]$, and $(1 - \pi)\mu_1 + \pi\mu_2 = 0$. As a test quantity, the likelihood ratio statistic λ between the two models seems to be adequate. It was applied in [8] as a formal test for clustering of microarray expression data. However, as noted from the authors *"the situation is not straightforward since regularity conditions do not hold for the asymptotic null distribution of* $-2\log(\lambda)$ *to be chi-squared"*. Therefore, McLachlan [7] proposed parametric bootstrap sampling under the null model to build a computerized reference distribution of the likelihood ratio. In Feng and McCulloch [3] it is demonstated that the true parameters of the alternative model lay in the boundary of the parameter space, and H_0 corresponds to a nonidentifiable subset. For the univariate case, the authors could show that the maximum likelihood estimator satisfies an extended consistency condition and that "test sizes and coverage probabilities of bootstrap methods match the nominal levels well in simulation studies, when the null hypothesis is true". An example is shown in Fig. 1, left side.

3 Regression Clustering

Clusterwise regression, also called latent class regression, is a special case in the framework of finite mixture models introduced in Leisch [6]. For linear regression analysis, we consider $(x, y) \in \mathbb{R}^2$, where y is taken as the dependent variable. Hypotheses corresponding to (1) and (2) become

$$H_0 : f_0(y|x, \theta) = N(\beta_0 + \beta_1 x, \sigma^2)$$

and

$$H_1 : f_1(y|x, \theta) = (1 - \pi)N(\beta_{10} + \beta_{11}x, \sigma_1^2) + \pi N(\beta_{20} + \beta_{21}x, \sigma_2^2),$$

where N here represents the univariate Gaussian density, symbols β_{ij} are the regression line coefficients, σ_i^2 are the variances, and $\pi \in [0, 1]$ is the mixing proportion. Using the likelihood ratio statistic together with its bootstrapped distribution as a test quantity, similar limitations as with the bivariate Gaussian mixture will occur. An example is shown in Fig. 1, right side.

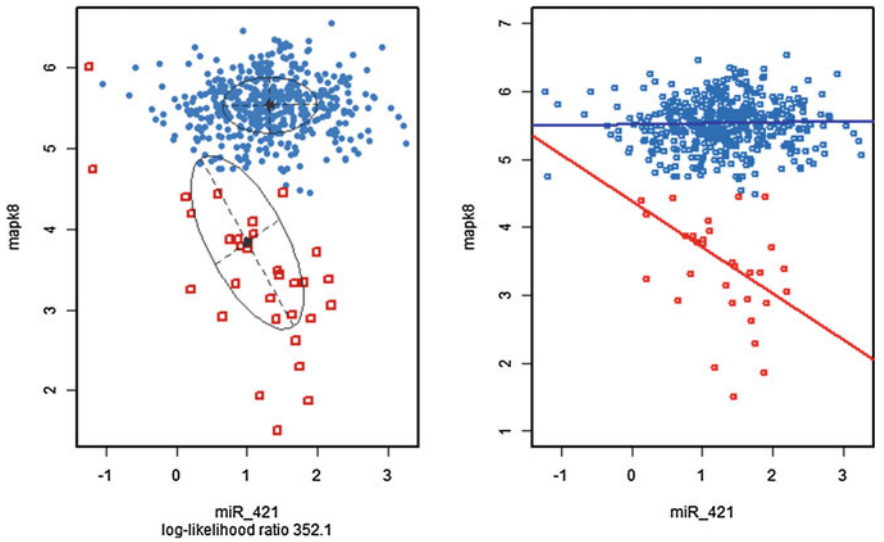


Fig. 1 miRNA-mRNA expression data from "The Cancer Genom Atlas" (TCGA), 495 cases. *Left* Gaussian mixed model. *Right* Regression clustering. $p < 0.0001$ from likelihood ratio bootstrap distribution for both approaches

4 Conclusion

So far, parametric model clustering is a useful method for descriptive and visual detection of heterogenous correlation of continuous variables of molecular biological data. Also, statistically significant clusters can be detected by parametric bootstrapping of the likelihood ratio statistic as proposed by McLachlan [7]. As a next step it remains to assess the significance of the covariance parameters of the bivariate normal distributed clusters or of the slope parameters in the regression clusters. Methods in Jamshidian and Jennrich [5] try to calculate standard errors for the EM estimates and could possibly be adapted for the presented kind of data. One of these methods uses Fisher scores and the information matrix. Publicly available data sets, for e.g., The Cancer Genome Atlas (TCGA) generated by the TCGA Research Network: http://cancergenome.nih.gov, provide a source to investigate the applicability to omics data.

References

1. D.J. Brenner, R. Doll, D.T. Goodhead, E.J. Hall, C.E. Land, J.B. Little, *et al.*, "Cancer risks attributable to low doses of ionizing radiation: assessing what we really know", *Proceedings of the National Academy of Sciences* **100**(24) (2003), 13761–13766.

2. A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the royal statistical society. Series B (Methodological)* **39**(1) (1977), 1–38.
3. Z.D. Feng and C.E. McCulloch, "Using bootstrap likelihood ratios in finite mixture models", *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(3) (1996), 609–617.
4. J.B. Hsu, C.M. Chiu, S.D. Hsu, W.Y. Huang, C.H. Chien, T.Y. Lee, and H.D. Huang, "miRTar: an integrated system for identifying miRNA-target interactions in human", *BMC bioinformatics* **12**(1) (2011), 300.
5. M. Jamshidian and R.I. Jennrich, "Standard errors for EM estimation", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(2) (2000), 257–270.
6. F. Leisch, "FlexMix: a general framework for finite mixture models and latent class regression in R", *Journal of Statistical Software* **11**(8) (2004), 1–18.
7. G.J. McLachlan, "On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture", *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **36**(3) (1987), 318–324.
8. G.J. McLachlan, R.W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data", *Bioinformatics* **18**(3) (2002), 413–422.
9. Y. Ruike, A. Ichimura, S. Tsuchiya, K. Shimizu, R. Kunimoto, Y. Okuno, and G. Tsujimoto, "Global correlation analysis for micro-RNA and mRNA expression profiles in human cell lines", *Journal of human genetics* **53**(6) (2008), 515–523.
10. D.M. Titterington, A.F.M. Smith, and U.E. Makov, "Statistical analysis of finite mixture distributions", New York: Wiley (1985).
11. K. Unger, "Integrative radiation systems biology", *Radiation Oncology* **9**(1) (2014), 21.
12. W.N. Van Wieringen, M.A. Van De Wiel, and A.W. Van Der Vaart, "A test for partial differential expression", *Journal of the American Statistical Association* **103**(483) (2008), 1039–1049.

Overview of Topics Related to Model Selection for Regression

Riccardo De Bin

Abstract We review some strategies proposed in the literature to combine clinical and omics data in a prediction model. We show how these strategies can be performed by using two well-known statistical methods, lasso and boosting, through an application to a biomedical study with a time-to-event outcome.

1 Introduction

The goal of a prediction model is to provide a function useful to predict a specific disease outcome. In recent years, a lot of attention has been devoted to taking advantage of the information contained in high-dimensional data (omics data), for example copy-number alterations or gene-expressions. Nonetheless, in the medical practice several low-dimensional predictors are often available, which generally have a predictive value well validated in the literature. Models which integrate clinical and omics information have been investigated in the recent years and recent comparative studies show that they often outperform those models based only on clinical or only on omics data. The studies Boulesteix–Sauerbrei [4] and De Bin–Sauerbrei–Boulesteix [6] discuss possible strategies to combine these two kinds of data in a prediction model. The main challenge in deriving a good combined prediction model lies in the need of fully exploiting both data sources. Clinical data, in particular, are low-dimensional, and the information contained in them risks getting lost among the large number of high-dimensional omics predictors; see Binder–Schumacher [2]. Here, we review the results of De Bin–Sauerbrei–Boulesteix [6] through an application to a biomedical study with a time-to-event outcome [1].

R. De Bin (✉)
Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany
e-mail: debin@ibe.med.uni-muenchen.de

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_13

2 Strategies for Combining Low and High Dimensional Data

The integration of clinical and omics predictors in the same model may be difficult. The different dimensionalities of the data (we usually have a few clinical predictors, and numerous omics predictors) make this process complicated. Here we review three strategies considered in De Bin–Sauerbrei–Boulesteix [6].

2.1 Naive Strategy

The easiest way to integrate clinical and omics predictors in the same model consists in simply merging the two sources of data. This procedure does not require any particular care, as the clinical and omics predictors are used as they would come from the same source. However, this procedure is quite dangerous, as the clinical predictors may "get lost" among the large number of omics predictors; see Binder–Schumacher [2]. As a result, the final model may not sufficiently exploit the information provided by the clinical data.

2.2 Clinical Offset Strategy

A possible way to force clinical predictors to enter in the model is to first fit a preliminary model on the clinical data only, and then use the omics information to explain the response variability not already explained by the preliminary (clinical) model. In other words, the residuals of the preliminary model are regressed on the omics predictors. From a practical point of view, it is sufficient to add the linear predictor of the preliminary model as an offset in the model fitted on the omics data. This two-step procedure completely exploits the clinical information, while the omics part is only complementary. An important drawback is that clinical and omics predictors are used in two separate steps, making it difficult to treat possible interactions between the two sources of data.

2.3 Favouring Strategy

This approach integrates both previous strategies in a compromise. The idea is to favour the clinical predictors in order to assure (or make more probable) their inclusion in the final model, without letting omics predictors be confined to a secondary role. The difference in dimensionality between the two sources of data is compensated by giving more importance to the clinical information. For example, in a penalized

regression approach the clinical data may be favoured by simply excluding them from the penalty term. This way, the variable selection aspect of a penalized regression involves only the omics part. As all data are used simultaneously, the interactions between clinical and omics predictors can be modelled.

3 Statistical Methods

The three aforementioned strategies can be easily performed using classical statistical methods; see Bin–Sauerbrei–Boulesteix [6]. Here, we consider two well-known statistical/machine learning techniques, namely lasso [11] and boosting [7].

3.1 Lasso

The lasso is a famous penalized regression method which handles high-dimensional data. Lasso applies a L_1 norm based penalty to the usual regression, with the double goal of forcing several regression coefficients to be 0 and shrink the rest toward 0. The former aspect is important to obtain a sparse model, i.e., a model including only few predictors. In the context of high-dimensional data, several predictors have no (or too weak) effect on the outcome, and they should be excluded from the final model. The shrinkage property, instead, improves the model prediction ability, decreasing the variance (at the cost of a small increase of the bias) of the regression coefficients estimators. Lasso relies on one important tuning parameter, usually denoted with λ, which regulates the amount of penalty applied to each predictor. The implementation of the favouring strategy may simply consist of setting $\lambda = 0$ for the clinical predictors.

3.2 Boosting

In addition to lasso, we also use boosting as a statistical method. Boosting is a stepwise procedure that can handle high-dimensional data as well. The idea of boosting is to build a model by improving, at each step, the model's ability to explain the variability of the outcome. The goodness of the model is measured through a loss function, which is problem-specific. Let us consider the regression case. Starting from a model with 0 coefficients, the boosting applies a weak estimator, for example a penalized least square estimator, to provide a slightly (depending on the amount of the penalty) better estimate of the regression coefficients. In a following step, the weak estimator is applied to the residuals of the model, obtaining a small improvement in terms of goodness of fit (for linear regression, usually the sum of the residuals). In the case of high-dimensional data, only one regression coefficient is updated at each step.

The procedure continues until a specific number of iterations (boosting steps) is reached. This tuning parameter is really important and controls both the amount of shrinkage and the model sparsity. For a discussion on boosting in the survival analysis context, see De Bin [5].

3.3 Tuning Parameters

Both lasso and boosting rely on tuning parameters. We compute their value using a 20-time repeated 10-folds cross-validation, which consists of averaging the cross-validation results obtained with 20 separate fold-splits; see Seibold–Bernau–Boulesteix–De Bin [9].

4 Data

In our analysis we considered data from a study by Bauer et al. [1] on patients with head and neck squamous cell carcinoma. The study investigates the patient response to a radiotherapy treatment in terms of time to local relapse (time-to-event). For each patient, the copy number alterations in chromosomes (omics data) and some clinical predictors (age, anaemia status, operating stage, tumour size, grading and lymph node metastasis) have also been collected. Preliminary analyses, including data quality control, lead to a dataset with 108 observations (patients), 6 clinical and 300 genomic predictors. The effective sample size (number of events) is 49. We work under the proportional hazards assumption.

4.1 Split in Training and Test Data

In order to evaluate the three strategies here considered, we need a test set. Since this is not available in the original study, we create it by splitting the available data into two sets. The first, with 2/3 of the observations, is used as the training set, in which the prediction models are fitted; the latter as the test set, in which we evaluate the prediction performance of the models. Please note that training and test sets are totally independent (no overlapping observations). The patients are assigned randomly to one set, but we consider censored and uncensored observations separately, in order to ensure that enough events are present in both training and test sets. In order to obtain a result that does not depend on a specific data split, we repeat our analysis 1000 times, each time using a different partition. We stress the importance of this repetition-based procedure to obtain reliable results.

Table 1 Average integrated Brier score (IBS) computed up to 2 years using different strategy/statistical method combinations

Strategy		Naive		Clinical offset		Favoring	
Statistical method	Null model	Lasso	Boosting	Lasso	Boosting	Lasso	Boosting
Average IBS	0.196	0.185	0.183	0.162	0.161	0.162	0.160

5 Evaluation

To evaluate the performances of the models resulting from the implementation of lasso and boosting within the three strategies, we contrast their prediction errors. Dealing with time-to-event data, we use as a measure of error the integrated Brier score [8], which captures both calibration (similarity between the actual and predicted outcomes) and discrimination (ability to predict the survival times of the observations in the right order), [10]. Calibration and discrimination are the two most important aspects of a good prediction model in the case of survival data.

6 Results

Table 1 reports the average prediction errors computed for the models obtained by implementing lasso and boosting within the naive, clinical offset and favouring strategies. The average is taken over the 1000 replications (i.e., different training/test sets splits) considered in our analysis. Excluding the null model, i.e., the model without predictor, here reported as a reference, in this example we obtain the worst performances for the statistical methods implemented within the naive strategy. The clinical offset and favouring strategies outperform the naive strategy, resulting in similar performances for both lasso and boosting. This behaviour has been noted in several studies (see, for example, [2, 3, 6]) and it is likely caused by the inability of the statistical methods to fully exploit the clinical information when implemented within the naive strategy. In this example, indeed, most of the models obtained by following the naive strategy include only one clinical variable (operating stage). The information provided by age, tumor size and number of affected nodes, for example, is not exploited.

Note that, in this specific example, the best performance is obtained with the favouring strategy/boosting combination. Obviously, this result highly depends on the particular characteristics of the data and cannot be generalized. The results reported in Table 1 must be understood as illustrative. In particular, in this example the validation set contains only 15 events (and the models are trained on a sample with only 30 events), definitely not a sufficient number to draw any general conclusion.

Acknowledgements Thanks go to Anne-Laure Boulesteix, Carine Legrand, Herbert Braselmann, Julia Hess and Kristian Unger. RDB was supported by grant BO3139/2-2 from the German Research Foundation (DFG).

References

1. V.L. Bauer, H. Braselmann, M. Henke, D. Mattern, A. Walch, K. Unger, M. Baudis, S. Lassmann, R. Huber, J. Wienberg, *et al.*, "Chromosomal changes characterize head and neck cancer with poor prognosis", *Journal of Molecular Medicine* **86** (2008), 1353–1365.
2. H. Binder and M. Schumacher, "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models", *BMC Bioinformatics* **9** (2008), 14.
3. H. Bøvelstad, S. Nygård, and Ø. Borgan, "Survival prediction from clinico-genomic models —a comparative study", *BMC Bioinformatics* **10** (2009), 413.
4. A.L. Boulesteix and W. Sauerbrei, "Added predictive value of high-throughput molecular data to clinical data and its validation", *Briefings in Bioinformatics* **12** (2011), 215–229.
5. R. De Bin, "Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost", *Computational Statistics* (2015).
6. R. De Bin, W. Sauerbrei, and A.L. Boulesteix, "Investigating the prediction ability of survival models based on both clinical and omics data: two case studies", *Statistics in Medicine* **33** (2014), 5310–5329.
7. J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent", *Journal of Statistical Software* **33** (2010), 1.
8. E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data", *Statistics in Medicine* **18** (1999), 2529–2545.
9. H. Seibold, C. Bernau, A.L. Boulesteix, and R. De Bin, "On the choice and influence of the number of boosting steps", *Technical Report* **188**, Department of Statistics, University of Munich (2016).
10. E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M.J. Pencina, and M.W. Kattan, "Assessing the performance of prediction models: a framework for some traditional and novel measures", *Epidemiology* **21** (2010), 128.
11. R. Tibshirani, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society. Series B (Methodological)* **58** (1996), 267–288.

Understanding Plaque Overlap Is Essential for Modelling Radiation Induced Atherosclerosis

Fieke Dekkers, Arjan Maud-Briels, Teun van van-Dijk, Astrid Dillen, and Kloosterman

Abstract Until recently, it was assumed that only relatively high doses of ionizing radiation could cause cardiovascular disease. In the last few years, several studies have challenged this assumption. Epidemiological studies remain inconclusive about the relevance in cardiovascular effects of exposure to ionising radiation in day-to-day situations. We present a preliminary version of a mechanistic model for atherogenesis and show that, in order to interpret the role of radiation in atherogenesis, it is essential to understand how plaques may overlap.

1 Radiation Induced Atherosclerosis

Atherosclerosis is a chronic inflammatory disease in which lipid-laden plaques form in arteries. If large, these plaques can influence the blood flow. A second possible health effect is that plaques may rupture, releasing the plaque's content into the artery, with heart attack or stroke as potential consequences. Patients who undergo radiotherapy as treatment for head-and-neck cancer are known to be at an increased risk of developing atherosclerosis in their carotid arteries. These patients are exposed to high doses of radiation and until recently, it was assumed that the much lower doses people are exposed to on a day-to-day basis would not lead to cardiovascular effects. In this respect, cardiovascular disease was thought to differ from cancer, which for radiation protection purposes is assumed to be a possible health effect of exposure to radiation at any dose, no matter how small. Recently, the existence of a high threshold for the induction of cardiovascular effects was challenged by studies such as Shimizu–Kodama–Nishi–Kasagi–Suyama–Soda–Grant–Sugiyama–Sakata–Moriwaki–Hayashi–Konda–Shore [3]. However, epidemiological studies remain inconclusive about the relevance on cardiovascular effects of exposure to ionising radiation at low doses, with risk estimates ranging over orders of magnitude. We present a preliminary version of a mechanistic model for

F. Dekkers (✉) · A. Maud-Briels · T. van van-Dijk · A. Dillen · Kloosterman
National Institute for Public Health and the Environment (RIVM), Utrecht, The Netherlands
e-mail: Fieke.dekkers@rivm.nl

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_14

atherogenesis that can be used to test hypotheses on the role of radiation in plaque formation. Particular focus is on the relevance of plaque overlap.

1.1 Plaque Initiation and Early Plaque Growth

We assume that atherosclerosis is an inflammatory process in which cholesterol and macrophages are key players. It is a multistage disease, starting off with an initiating event that causes the layer of cells that lines the vessel wall to become permeable for low density lipoproteins ("bad cholesterol"). This sets off an immune response, leading to an influx of monocytes. Once in the vessel, low-density lipoproteins oxidize to form oxidized low-density lipoproteins (oxLDL), and monocytes differentiate into macrophages that proceed to phagocyte oxLDL. This leads to the development of foam cells, macrophages laden with lipids. The process then escalates, eventually resulting in the development of plaques.

Mice do not normally develop plaques, but genetically modified ApO E-/-mice lack an enzyme that plays a role in the cholesterol metabolism, making them vulnerable to atherosclerosis. We have developed a model for murine radiation induced atherosclerosis (manuscript in progress), and have validated the model by fitting it to experimental results; see Hoving–Heeneman–Gijbels–te Poele–Russell–Daemen–Stewart [1].

In our model we assume that, in the absence of radiation, a Poisson process describes plaque initiation. We assume that exposure to ionizing radiation leads to a temporary increase in the rate of initiation, persisting for about two weeks. This is consistent with the observation in Tribble–Barcellos-Hoff–Chu–Gong [4] that ApOE-/- mice on a fat free diet do not develop plaques, but they do if exposed to ionizing radiation and switched to a normal diet within eight days after exposure.

For subsequent plaque growth, we follow the ODE model developed in Ougrinovskaia–Thompson–Myerscough [2]. This is a model for spontaneous plaque growth. It cannot be ruled out that chronic ionizing radiation affects plaque growth, but for the acute exposures considered here, this is unlikely to be significant: the duration of the experiment was approximately one year, and even if an effect of plaque growth persists for some time after exposure, plaque growth rates during this time would have to be implausibly high to produce an observable effect.

Figure 1 gives a typical plaque volume growth curve: plaques initially grow slowly, but the rate of growth increases rapidly. Obviously, this model can only describe the early phases of plaque growth.

Spatial effects have not been included in the model. It should be noted that this implies that the optimal values of the model's parameters can differ between arteries: for example, in the mouse's descending thoracic aorta, plaque development is markedly different from the aortic arch, where blood flow is much more turbulent, resulting in more or larger plaques.

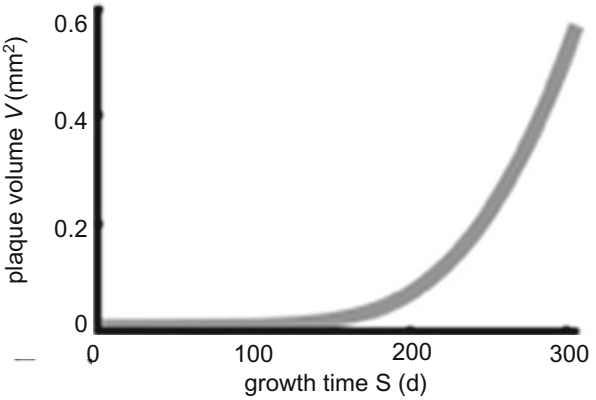


Fig. 1 Typical plaque volume growth

1.2 Plaque Overlap

It follows from the assumptions made in the model that both, with and without radiation, the model predicts a distribution without large gaps. A typical example is given in Fig. 2, right. Results from an experiment in which mice were exposed to low doses of ionizing radiation (data courtesy of Anna Saran, ENEA) are illustrated on the left.

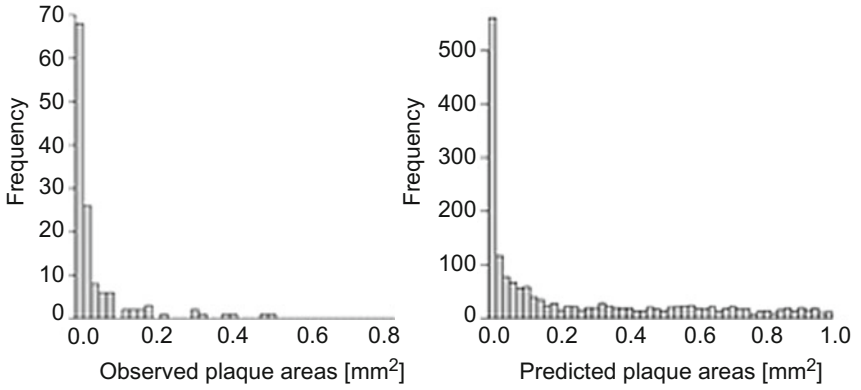


Fig. 2 Observed (*left*) and predicted (*right*, 500 simulations) distribution of plaque sizes. The large number of small plaques in the model prediction reflects the initial slow growth of plaques. Note that in the observed plaque areas, the largest plaque is almost twice the size of the next largest plaque

At first sight, the predictions may appear to be irreconcilably different from the observations. However, the observed distribution of plaque sizes suggests that some of the larger plaques may actually consist of two plaques that have started to overlap. In the experiment, plaques are identified by colouring with a dye that stains fatty material. Thus, overlapping plaques cannot easily be identified.

We have tested the assumption that plaque coalescence influences the observed distribution of plaque sizes with a very simple Monte Carlo simulation, where plaques were created uniformly over a rectangular artery following the distribution over time of initiations found with the original model. Subsequently, these plaques are allowed to grow, following the original growth curve. If two plaques overlap, we describe them as a single plaque.

The model for plaque initiation and growth combined with this simple model for plaque coalescence results in a distribution of plaque sizes that is in qualitative agreement with that observed in experiments (Fig. 2), leading to the conclusion that understanding plaque coalescence is essential for a better understanding of the influence of radiation on atherosclerosis: if plaque overlap is ignored, the number of plaques counted in experiments will be an underestimate for the true number of plaques, and plaque sizes measured may be an overestimate for the actual sizes of individual plaques. Therefore, simply counting plaques in an experiment will lead to an underestimate of the risk associated with exposure to ionizing radiation. It cannot be ruled out that in extreme cases a straightforward counting procedure would lead to the conclusion that ionizing radiation has a beneficial effect, whereas a model taking into account plaque overlap would indicate the opposite. Thus, our simple overlap model illustrates the statement that mechanistic modelling can contribute to improved risk estimates (Fig. 3).

Fig. 3 Plaque overlap in a small section of an artery, cut open at the *top*/*bottom* to produce a rectangular section. Plaques are created with uniform distribution over the section. Subsequently, plaques develop following the growth rate illustrated in Fig. 1. If overlap occurs, plaques are joined to form a single plaque: the plaques at the *top* and *bottom* both consist of plaques that have coalesced

References

1. S. Hoving, S. Heeneman, M.J. Gijbels, J.A. te Poele, N.S. Russell, M.J. Daemen, and F.A. Stewart, "Single-dose and fractionated irradiation promote initiation and progression of atherosclerosis and induce an inflammatory plaque phenotype in ApoE(-/-) mice", *Int. J. Radiat. Oncol. Biol. Phys.* **71**(3) (2008), 848–57.
2. A. Ougrinovskaia, R. Thompson, and M. Myerscough, "An ODE model of early stages of atherosclerosis: mechanisms of the inflammatory response", *Bulletin of Mathematical Biology* **72**(6) (2010), 1534–1561.
3. Y. Shimizu, K. Kodama, N. Nishi, F. Kasagi, A. Suyama, M. Soda, E.J. Grant, H. Sugiyama, R. Sakata, H. Moriwaki, M. Hayashi, M. Konda, and R.E. Shore, "Radiation exposure and circulatory disease risk: Hiroshima and Nagasaki atomic bomb survivor data, 1950–2003", *BMJ* **340** (2010), b5349.
4. D.L. Tribble, M.H. Barcellos-Hoff, B.M. Chu, and E.L. Gong, "Ionizing radiation accelerates aortic lesion formation in fat-fed mice via SOD-inhibitable processes", *Arterioscler Thromb Vasc Biol.* **19**(6) (1999), 1387–1392.

On the Use of Random Effect Models for Radiation Biodosimetry

Jochen Einbeck, Elizabeth Ainsbury, Stephen Barnard, Maria Oliveira, Grainne Manning, Pere Puig, and Christophe Badie

Abstract The application of random effect models to different radiation biomarkers, including cytogenetic, protein-based, and gene-expression based biomarkers, is discussed. Explicit case studies are provided for the latter two scenarios, in which random effect models appear especially attractive as they can cope well with the large inter-individual variation which is typical for these biomarkers.

1 Introduction

After some radiation accident or incident, the triage of individuals requires rapid and reliable procedures to determine the contracted radiation dose. Biomarkers estimate the dose through radiation-induced changes within cells of the human body.

J. Einbeck (✉) · M. Oliveira
Department of Mathematical Sciences, Durham University, Durham DH13LE, UK
e-mail: jochen.einbeck@durham.ac.uk

E. Ainsbury · S. Barnard · G. Manning · C. Badie
Centre for Radiation, Chemical and Environmental Hazards, Public Health England, Chilton, Didcot, Oxon OX11 0RQ, UK
e-mail: Liz.Ainsbury@phe.gov.uk

S. Barnard
e-mail: Stephen.barnard@phe.gov.uk

G. Manning
e-mail: Grainne.Manning@phe.gov.uk

C. Badie
e-mail: Christophe.Badie@phe.gov.uk

M. Oliveira
Department of Statistics and Operations Research, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
e-mail: maria.oliveira@usc.es

P. Puig
Departament de Matemàtiques, Universitat Autònoma de Barcelona, Cerdanyola Del Vallès, 08193 Barcelona, Spain
e-mail: ppuig@mat.uab.cat

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_15

Cytogenetic biomarkers, which count chromosome aberrations in blood lymphocytes, have formed the gold standard for at least three decades, largely due to the fact that they feature very little inter-individual variation. Other biomarkers, based on protein phosphorylation or gene expressions, have been developed more recently. While these can be obtained quicker and at much lower cost, they come with large inter-individual variation, which needs to be properly addressed in the modelling stage, since any variation unaccounted in this stage would lead to an incorrect uncertainty quantification of the succeeding dose estimate. Sections 2 and 3 give two case studies which report the results of random effect modelling for a particular protein biomarker (γ-H2AX) and a biomarker based on gene expressions. Section 4 gives some thoughts on random effect models for cytogenetic biomarkers such as dicentrics.

2 Protein Biomarkers

Protein-based biomarkers such as γ-H2AX make use of relatively new technology, compared to cytogenetic biomarkers. Radiation-induced double strand breaks (DSBs) lead to a 'phosphorylation' of the H2AX protein. This cellular response is manifested for analysis as fluorescent dots (γ-H2AX foci), which need to be counted within cells, typically using immunofluorescence microscopy or flow cytometers. While this technique is quicker, cheaper, and less invasive than the dicentric approach, the phosphorylation is only visible up to approximately 24 h after radiation exposure. Anonymised data from 36 donors are available for analysis, with measurements at irradiation levels of 0, 0.5, and 4 Gy. The data, which comprise 57 observations in total, feature several peculiarities requiring a careful modelling approach:

(i) The 0.5 Gy data were collected 24 h and the 4 Gy data 30 min after exposure, which requires 'dose' to be modelled as a factor.
(ii) For some observations, the γ-H2AX foci were counted manually but an automated scoring method was used for others. Hence, we have considered the inclusion of an indicator variable to capture the type of scoring methodology used, which, however, turned out to be insignificant.
(iii) The design is far from balanced: there are 15 measurements for dose zero, 26 measurements for dose 0.5 Gy, and 16 measurements for 4 Gy. For 17 individuals, there is only one measurement, for another 17 there are two measurements, and for two individuals there are three measurements.

We fitted to these data a Poisson regression model with log-link, and a linear predictor consisting of an intercept and the factor for dose (entering in form of two indicator variables for the 0.5 and 4 Gy categories). Additionally, a donor-specific random effect was added to the linear predictor, for which we considered a Gaussian distribution (in this case the model was fitted using function `glmer` of R package **lme4**), and an unspecified distribution (which can be estimated using Nonparametric

Table 1 Summary results of models fitted to H2AX data. The value p gives the number of parameters used; df $= 57 - p$ is the degrees of freedom for error. The deviance (Dev) is obtained from the disparity $(-2 \log L)$ after subtracting the disparity corresponding to the saturated log-likelihood. The abbreviation r.e. denotes 'random effect' and k gives the number of mass points used for the nonparametric maximum likelihood estimation

Model	Linear predictor	p	df $-2 \log L$	Dev	
(1)	Intercept only	1	56	8360.5	7984.9
(2)	(1) + dose	3	54	565.9	190.3
(3)	(2) + r.e. (Gauss)	4	53	517.1	141.5
(4)	(2) + r.e. ($k = 3$)	7	50	508.0	132.4
(5)	(2) + r.e. ($k = 4$)	9	48	506.3	130.7

Maximum Likelihood, using function `allvc` of R package **npmlreg**). Summarized results of a detailed analysis are provided in Table 1. One clearly sees the improvements at each model refinement stage. Obviously, dose acts as a very useful predictor (and the dose indicators are in fact highly significant, with p-values $< 10^{-10}$; not displayed here). Model (2) indicates a relatively strong overdispersion, with a rule-of-thumb estimation of the dispersion via Dev/df = $190.3/54 = 3.52 > 1$. Part of this overdispersion is explained by the inter-individual variation, with the deviance dropping by ca 50 points after the inclusion of the random effect. There is some weaker evidence that a nonparametric random effect distribution should be employed, with a further improvement of up to 10 deviance points, investing 5 additional degrees of freedom.

3 Biomarkers Based on Gene Expressions

Currently, there is no established technique which would allow fast (< 24 h) dose assessment with samples that have been taken at least 24 h after the radiation incident. Following radiation exposure, many genes see their expression modified at the transcriptional level and modifications of the copy number of mRNAs can be measured in a very sensitive way by qPCR (quantitative polymerase-chain-reaction) technology. Here, two genes up-regulated through the ATM/CHEK2/P53 pathway, CCNG1 and CDKN1A (p21), were studied. Gene expression measurements (positive continuous data) were obtained from peripheral blood samples from 32 healthy human donors, at three dose levels (0, 2, and 4 Gy), with three replicates for each donor and dose level. After irradiation the samples were placed in an incubator at 37° for 24 h.

For this tentative analysis, we use the measured expressions without standardization by housekeeping genes. Also here, the design is not complete, with the 2 Gy

Table 2 Summary results of models fitted to qPCR data. The number p denotes the number of parameters estimated (including the Gamma shape parameter), $-2\log L$ denotes the model disparity, and BIC the Bayesian Information Criterion $-2\log L + p\log 456$ (low values of either indicate well-fitting models)

Model	Linear predictor	p	$-2\log L$	BIC
(1)	Intercept only	2	3132.1	3144.3
(2)	(1) + gene type	3	3009.3	3027.7
(3)	(2) + dose	4	2651.1	2675.6
(4)	(3) + r.e. (Gauss)	5	2617.0	2647.6
(5)	(3) + dose^2	5	2480.0	2510.6
(6)	(5) + r.e. (Gauss)	6	2443.4	2480.1

measurements missing for 20 of the individuals. Hence, the total number of observations is $20 \times 2 \times 2 \times 3 + 12 \times 2 \times 3 \times 3 = 456$. Given the nature of the expressions (positive, continuous) we decide for a Gamma regression model with log-link. There are two possible routes to proceed with the modelling. Firstly, one could consider a model with the 456 expression values forming the response vector, but including an indicator for the gene type. Since there is no need to model dose as a factor for these data, it can be included as a continuous predictor, in linear or quadratic (or possibly more sophisticated) form. The results of a similar analysis as in Sect. 2 are provided in Table 2. We do not report deviances (since there is no need to assess 'overdispersion' for a Gamma model), but instead the BIC model selection criterion.

It is again seen that the dose variable is highly informative. Given the inclusion of dose, there is justification for the inclusion of both a random effect term (on the level of the donors) and a quadratic dose term, with the evidence for the latter being stronger. Both the linear and the quadratic dose term are again highly significant with p-values of less than 10^{-10}. Proceeding towards a nonparametric random effect distribution did not give further improvement so, the corresponding results are not reported.

There is a second, possibly more appealing view on the data, in which one may consider the expressions for the two gene types as a bivariate response. Under such a model, each of the two model equations would use the same parametric form but with separately estimated parameters. The two equations would be linked through the response variance specification (equivalently, the person-specific random effect) which then would lead to reduced parameter standard errors. This could be highly beneficial for dose estimation problems, as the reduced standard errors will lead to reduced model-based uncertainty, and, hence, more precise dose estimates. Unfortunately, even though its foundations have been laid conceptually, this model is quite poorly supported by statistical software so far, apart from a SAS procedure which allows for multivariate Gaussian response models with Gaussian random effects. Further practically focused research activity in this direction appears hence desirable.

4 Cytogenetic Biomarkers

Cytogenetic biomarkers—such as dicentrics, centric rings or micronuclei—feature little inter-individual variation, suggesting that there is no strong need to address individual-specific variation through random effect terms. However, this simple conclusion has recently been disputed; see Mano–Suto [1]. The authors of this reference provided a fully Bayesian approach to dose estimation from multi-individual data (using Poisson response with log-link, a quadratic dose model, and random effects for all three regression parameters), and demonstrated that the credible intervals obtained using their method are more appropriate in terms of coverage probability than the standard method. It would be interesting to see an application of (an extension/adaptation of) their method to protein or expression-based biomarkers, where the benefits of this approach could possibly be much larger.

Random effect models may also appear attractive from another perspective: usually, several blood samples are irradiated with different doses, and then the number of cytogenetic aberrations within a number of cells are counted for each sample. Hence, one may reasonably assume that there does exist within-blood sample correlation, which could be modelled by a random effect term operating on the blood sample level. Such models have been discussed in detail in the supplementary material of reference [2], available at http://onlinelibrary.wiley.com/doi/10.1002/bimj.201400233/suppinfo.

5 Conclusion

For both H2AX and qPCR biomarkers, our model fitting exercise has confirmed (i) the existence of very strong dose effects, and (ii) the need for the inclusion of a donor-specific random effect. A well fitting model is essential for the succeeding dose estimation, which involves an inverse regression step which is non-trivial under the presence of a random effect, since it requires integration over the latter. This step can be done quite naturally within a fully Bayesian framework (though possibly computationally burdensome) or via an empirical Bayes-like strategy under a frequentist approach. In either case, the integration step will naturally incur a loss of precision of the dose estimate; or in other words, inclusion of the random effects increase the size of credible/confidence intervals; a fact which has also been noted by Mano–Suto [1]. Hence, it still would feel desirable to reduce the inter-individual variation as far as possible before any modelling takes place, for instance, by calibration via appropriate housekeeping genes which mirror the inter-individual variation but are not radiation-responsive. Increasing the accuracy of models to assess dose and uncertainty will form a valuable contribution to the field of radiation dosimetry. Such work has the potential to lead to further advances in the wider field of biomarker development, for instance, panels of transcriptional and other biomarkers of radiation exposure and/or effect which would be applicable to specific individuals or groups of

individuals in different exposure scenarios. A great deal of work remains to be done, but EU and worldwide radiation emergency preparedness and health protection will greatly benefit as a result of such developments.

Acknowledgements Part of this research has been carried out while the first author was a guest of the Centre de Recerca Matemàtica (CRM) in Barcelona. This report is also based on independent research supported by the National Institute for Health Research (NIHR), "Random effects modelling for radiation biodosimetry" (NIHR-RMOFS-2013-03-4). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Dept. of Health.

References

1. S. Mano and Y. Suto, "A bayesian hierarchical method to account for random effects in cytogenetic dosimetry based on calibration curves", *Radiat Environ Biophys.* **53** (2014), 775–780.
2. M. Oliveira, J. Einbeck, M. Higueras, E. Ainsbury, P. Puig, and K. Rothkamm, "Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study", *Biometrical Journal* **58**(2) (2015), 259–279.

Modelling of the Radiation Carcinogenesis: The Analytic and Stochastic Approaches

Krzysztof W. Fornalski, Ludwik Dobrzyński, and Joanna Reszczyńska

Abstract The paper summarizes the analytic and stochastic approaches to radiation induced carcinogenesis among generic cells population. Many important effects were taken into consideration, like chromosomal aberrations induction, bystander effect, adaptive response effect, etc. The results can be simulated in analytical or Monte Carlo forms that show, e.g., a general probability function for a single cell's cancer transformation.

1 Introduction

The radiation carcinogenesis process of single cell or cell colony can be described by many deterministic and stochastic models. All such models can be easily implemented as computational algorithms for Monte Carlo or analytical simulations. The main scope of the presented paper is to demonstrate two approaches: analytic (deterministic) and stochastic ones, in the generic modelling/description of radiation carcinogenesis process at cells' level.

2 Stochastic Approach

The stochastic modelling based on the Monte Carlo method with probability tree was introduced in previous papers; see [7, 9]. The tree contains several input parameters, such as the probability distributions of many biophysical effects. It can be easily modified and new branches can be added according to either newly acquired knowledge or focusing on one detailed mechanism in a cell. In the following, elementary biophysics used in the model will be described.

K.W. Fornalski (✉)
PGE EJ 1, Warszawa, Poland
e-mail: krzysztof.fornalski@gmail.com

L. Dobrzyński · J. Reszczyńska
National Centre for Nuclear Research (NCBJ), Warsaw, Poland

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_16

The model contains several input parameters which create probability functions (PFs) used in the Monte Carlo modelling in the probability tree. The potential users of the model can use their own values of parameters and forms of PFs, according to their best knowledge of cell's biophysics. This is what makes the model rather flexible.

In general, every PF is a differentiable continuous function which saturates to a constant value. Thus, the simplest example is a quasi-linear function which can be used, e.g., instead of the classical linear relationship [12, 18]:

$$P(\xi) = 1 - e^{-\text{const}\cdot\xi}. \tag{1}$$

For low values of ξ (e.g., the dose) Eq. (1) is quasi-linear and tends asymptotically to 1 at large doses.

Another example is a sigmoidal relationship (stretched-exponential):

$$P(\xi) = 1 - e^{-a\xi^n}, \tag{2}$$

which can be assumed as a basis for carcinogenic mutations accumulation process, e.g., in Knudson hypothesis [11, 13], where it has been shown that, four to seven rate-limiting stochastic events are required for neoplastic transformation of the cell. Based on this assumption, one can model the probability of such process using the sigmoidal formula as presented, e.g., in Eq. (2) (many other forms of sigmoidal curve do exist).

The process of creation of a cancer cell can be described in many ways, including physical models. They can stem from theories of nucleation and growth [3, 4], of catastrophe [21], or self-organised criticality [19]. The common idea underlying these theories is the cumulative impact of some environmental stressors (here: radiation) on complex adaptive systems that may result in a rapid non-linear response when the stress exceeds some critical value. In the context of the theory of nucleation and growth [3, 4], the cancer creation can be treated as a rapid non-linear growth appearing around the nucleus, see Eq. (2). However, a more general form of Eq. (2) is a scaled sigmoid function with additional scaling factor τ:

$$P(\xi) = (1 - \tau)(1 - e^{-a\xi^n}) + \tau. \tag{3}$$

This form of sigmoid can well describe functions which start from a non-zero point, e.g., $P(0) = \tau$. It may also be possible that, if a threshold for an effect exists at a certain ξ_0, the ξ in the exponent could be replaced by $\xi - \xi_0$. Some processes, taking place in a cell, like the possibility of repair, can be described by inverted sigmoid function, describing the rapid decrease of repair process efficiency after reaching some age (or another value of interest):

$$P(\xi) = \delta e^{-a\xi^n}. \tag{4}$$

2.1 Adaptive Response Effect

Many effects (including repair and non-targeted ones), can be inserted into the model. In order to model a positive (beneficial) effect of irradiation, one has to describe first the PF of the adaptive response effect, see [12]. Such a probability should be given by a PF exhibiting a maximum value at low doses with the strongest effect appearing after some period of time, [5]. The probability distribution of adaptive response should thus be dependent both on the dose (D) received by single cell, and time (t).

The dose and time ingredients of adaptive response can be approximated by the following distributions, respectively:

$$p(D) = \alpha_1 D^{\nu} e^{-\alpha_2 D}, \tag{5}$$

$$p(t) = \alpha_4 t^{\delta} e^{-\alpha_3 t}. \tag{6}$$

The time t is measured after the irradiation (with the exposure D) took place. Both leading multipliers, α_1 and α_4 are just proportionality constants. The choice of exponents ν and δ is also somewhat arbitrary, but they should not be less than 1 to create the adaptive-hunchbacked shape of both curves.

Assuming, for simplicity, that the time evolution of the adaptive response is dose-independent, compilation of Eqs. (5) and (6) gives the time and dose dependent distribution [7]:

$$p(D, t) = c D^{\nu} t^{\delta} e^{-\alpha_2 D - \alpha_3 t}, \tag{7}$$

where c is a normalization constant. In the following it will be assumed, again for simplicity reasons, that $\nu = \delta = 2$. This choice is absolutely arbitrary and accepted here just for qualitative description.

It should be noted that Eq. (7) describes the adaptive response from a single irradiation only. In a real situation, one can find different values of D during cell's lifetime, T, where $t \in [0, T]$. After some modifications, one can find more general and continuous form of adaptive response PF [7]:

$$P_{AR} = c \int_{t=0}^{T} \dot{D}^2(t)(T-t)^2 e^{-\alpha_2 \dot{D}(t) - \alpha_3 (T-t)} dt, \tag{8}$$

where $\dot{D}(t)$ corresponds to the dose rate distribution function. For constant dose-rate, $\dot{D}(t) = \text{const}$, the solution is simpler and only age (T) dependent. It is of interest to see how Eq. (8) can disclose a spectacular effect of a priming dose. This is clearly demonstrated in Fig. 1, which displays the Monte Carlo simulation of the effect: the initial low dose may create the adaptive response and protect the cell from an adverse effect of a higher dose.

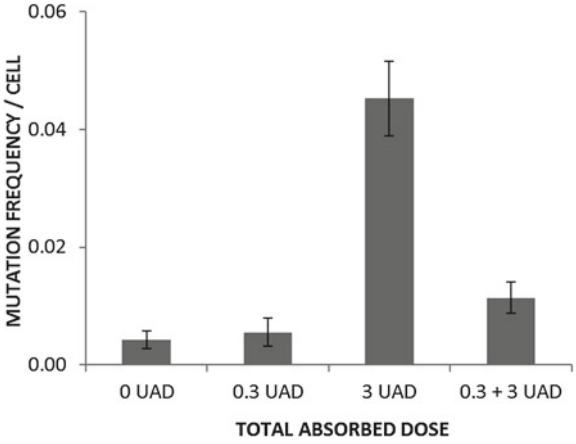


Fig. 1 The exemplary solution of Eq. (8) and simulation of the priming dose of 0.3 UAD (unit of absorbed dose) [7]

2.2 *Bystander Effect*

The bystander effect [12] can appear with the PF dependent on dose, $\xi = D$. A suggestion for its functional shape, similar to Eq. (1) but with an additional scaling parameter, was given by [14]:

$$P(\xi) = \beta_1(1 - e^{-\beta_2 \xi}). \tag{9}$$

However, in a realistic situation, the cells are spatially arranged in a two or three dimensional matrix, which—in the simulations—requires careful accounting for a distance (r) from the irradiated cell. It seems natural that the probability of bystander effect must decrease with the distance, which can be well described by the Poisson distribution connected with a single hit [10, 16]:

$$P_B(r) = c_1 \frac{e^{-\lambda}\lambda^r}{r!} \equiv \frac{1}{r!}, \tag{10}$$

where the last equality simplifies the previous expression by assuming $\lambda = 1$ and normalization to $P(r = 0) = 1$. Using similar reasoning like in the case of Eq. (7), one can postulate the distance and dose (D) dependent probability distribution of bystander effect as [7]:

$$P_B(D, r) = \frac{\beta_1}{r!}(1 - e^{-\beta_2 D}). \tag{11}$$

The approach presented in this section may provide clues for modeling of the impact of low doses on hypothetical cells with the use of the Monte Carlo techniques. The main conclusion is that the colony of cells can be treated as a physical complex system when even linear inputs will give a non-linear response, e.g., the shifted sigmoidal shape with a threshold for cancer cells induction; see [7, 9]. A prospective

user of the model can add new branches to the probability tree and/or can select a completely different set of the input probabilities.

3 Analytic Approach

The first step in a proper analytic approach to radiation carcinogenesis process is to use popular PF for post-irradiation chromosomal aberration frequency [2, 20]:

$$p(D) \propto \sum_{i=1}^{R} \beta_i D^i, \tag{12}$$

where D denotes the dose, β_i's are parameters of the curve, and R type of radiation. At high doses of several grays and more, neither linearity ($R = 1$) nor parabolicity ($R = 2$) can be preserved [15], since Eq. (12) yields diverging values of probability at high doses. In more general versions the probability does not tend to infinity, but the results of low and medium dose calculations are qualitatively the same. Therefore, one can use a more general form of Eq. (12) as presented in Eq. (1):

$$p(D) \propto 1 - e^{-\sum_{i=1}^{R} \beta_i D^i}. \tag{13}$$

One has to note, however, that the frequency of post-irradiation chromosomal aberrations is not constant but rather decreases with time. The decrease can be described as:

$$p(t) \propto e^{-\lambda t}, \tag{14}$$

where t denotes the time and λ is a parameter correlated with natural chromosomal aberrations removal mechanisms; see [2].

Finally, one can find the joint form of chromosomal aberration creation in a single cell, as a time and dose dependent probability function:

$$p(D,t) \propto \left(1 - e^{-\sum_{i=1}^{R} \beta_i D^i}\right)\left(1 + be^{-\lambda t}\right), \tag{15}$$

where b is a constant connected with the number of aberrations created at $t = 0$.

To have a complete view, one needs to add some beneficial response formula, which can take the form of the formerly proposed adaptive response PF after a single hit, see Eq. (7). Finally, the joint form of both types of responses, detrimental Eq. (15) and beneficial Eq. (7), can be described as

$$p(D,t)P_{CT} \propto \left[1 - e^{-\sum_{i=1}^{R} \beta_i D^i}\right](1 + be^{-\lambda t}) - cD^2t^2e^{-\alpha_2 D - \alpha_3 t}, \tag{16}$$

which can result in a hormetic-like saturated shape of the irradiation effects of a single cell [8]. This approach is fully consistent with Dual Response Model [6].

The symbol P_{CT} in Eq. (16) can be treated as a PF of a radiation-induced single cell cancer transformation.

4 Discussion

The process of radiation carcinogenesis over time and dose can be described in many ways, both stochastic and deterministic ones. By reviewing various mathematical approaches to the problem, one can appreciate that the detailed description of the cancer incidence due to dose and time is very complicated, described by many parameters, and certainly far from linear.

The authors are fully aware of the fact that the detailed modeling of dynamics of the cancer cell formation and development of such cells in a colony is quite impossible because of the variety of cancer types and their specific properties [1, 17]. Nevertheless, at least the main characteristics of the cancer development as dependent on dose and time should be caught in relatively simple calculations. In particular, the role of adaptive response and bystander effects is well elucidated in the presented calculations carried out so far.

Acknowledgements The authors wish to thank Dr. Yehoshua Socol and Prof. Marek K. Janiak for stimulating discussions.

References

1. "BEIR VII Report (Biological Effects of Ionizing Radiation committee). Health risks from exposure to low levels of ionizing radiation", Board on Radiation Effects Research, Commision on Life Sciences, National Research Council, National Academy Press, Washington (2006).
2. "Cytogenetic dosimetry: applications in preparedness for and response to radiation emergencies", IAEA (International Atomic Energy Agency), Safety Standards, Vienna (2011).
3. M. Avrami, "Kinetics of phase change. II. Transformation-time relations for random distribution of nuclei", *Journal of Chemical Physics* **8**(2) (1940), 212–224.
4. M. Avrami, "Kinetics of phase change. III. Granulation, phase change, and microstructure", *Journal of Chemical Physics* **9**(2) (1941), 177–184.
5. L.E. Feinendegen, "Low doses of ionizing radiation: relationship between biological benefit and damage induction. A synapsis", *World Journal of Nuclear Medicine* **4** (2005).
6. L.E. Feinendegen, V.P. Bond, and C.A. Sondhaus, "The dual response to low-dose irradiation: induction versus prevention of DNA damage", in "Biological Effects of Low Dose Radiation", Elsevier (2000), 3–17.
7. K.W. Fornalski, "Mechanistic model of the cells irradiation using the stochastic biophysical input", *International Journal of Low Radiation* **9**(5/6) (2014), 370–395.
8. K.W. Fornalski, "Biophysical Monte Carlo modelling of irradiated cells", presentation during LD-RadStats – DoReMi Workshop, CRM, Barcelona, Spain (2015), available at http://www.doremi-noe.net/pdf/events/radstats15/DoReMiRadstats2015_Fornalski.pdf.

9. K.W. Fornalski, L. Dobrzyński, and M.K. Janiak, "A stochastic Markov model of cellular response to radiation", *Dose-Response* **9**(4) (2011), 477–496.
10. S. Gaillard, D. Pusset, S.M. de Toledo, M. Fromm, and E.I. Azzam, "Propagation distance of the α-particle-induced bystander effect: the role of nuclear traversal and gap junction communication", *Radiat Res.* **171**(5) (2009), 513–520.
11. A. Knudson, "Mutation and cancer: statistical study of retinoblastoma", *Proc. Natl. Acad. Sci. USA* **68**(4) (1971), 820–823.
12. B.E. Leonard, "A review: development of a microdose model for analysis of adaptive response and bystander dose response behavior", *Dose-Response* **6** (2008), 113–183.
13. C. Nordling, "A new theory on cancer-inducing mechanism", *British Journal of Cancer* **7**(1) (1953), 68–72.
14. K.M. Prise, M. Folkard, and B.D. Michael, "A review of the bystander effect and its implications for low-dose exposure", *Radiation Protection Dosimetry* **104**(4) (2003), 347–355.
15. M.S. Sasaki, "Chromosomal biodosimetry by unfolding a mixed Poisson distribution: a generalized model", *Int. J. Radiat. Biol.* **79**(2) (2003), 83–97.
16. K. Sasaki, K. Waku, K. Tsutsumi, A. Itoh, and H. Date, "A simulation study of the radiation-induced bystander effect: modeling with stochastically defined signal reemission", *Computational and Mathematical Methods in Medicine* **2012** (2012), 389095.
17. R.D. Schreiber, J.O. Lloyd, and M.J. Smyth, "Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion", *Science* **331** (2011), 1565–1570.
18. B.R. Scott, J. Hutt, Y. Lin, M.T. Padilla, K.M. Gott, and C.A. Potter, "Biological microdosimetry based on radiation cytotoxicity data", *Radiation Protection Dosimetry* **153**(4) (2013), 417–424.
19. M. Stark, "The sandpile model: optimal stress and hormesis", *Dose-Response* **10**(1) (2012), 66–74.
20. M. Szłuińska, A.A. Edwards, and D.C. Lloyd, "Statistical methods for biological dosimetry", Health Protection Agency/Public Health England report no. HPA-RPD-011 (2005).
21. E.C. Zeeman, "Catastrophe Theory", Selected Papers 1972–1977. Addison-Wesley Publ. Co. (1977).

Bayesian Solutions to Biodosimetry Count Data Problems and Supporting Software

Manuel Higueras and Elizabeth A. Ainsbury

Abstract Recently developed Bayesian models for the estimation of the absorbed ionising radiation dose in whole and partial body homogenous exposures are introduced. The whole body exposure models can manage with overdispersed frequencies of chromosomal aberrations. Poisson models have been implemented in the `R` package `radir`.

1 Introduction

Ionising radiation induced damage at human cellular level (e.g., dicentrics and micronuclei) is typically modelled by Poisson regression. The absorbed dose estimation is an inverse problem, the number of chromosomal aberrations per cell is modelled as a function of the absorbed dose but not *vice versa*. The current frequentist methodology does not provide an accurate measure of the uncertainty of the absorbed dose estimation.

2 Bayesian Approach

A Bayesian approach is highly applicable to ionising radiation dosimetry data. It allows cytogenetic experts to consider prior knowledge surrounding an overexposure scenario. This approach implies an accurate measure of the uncertainty of dose

Collaborators: Pere Puig (Universitat AUtònoma de Barcelona), David Moriña (Centre for Research in Environmental Epidemiology), Volodymyr A. Vinnikov (Institute for Medical Radiology), and Kai Rothkamm (University Medical Center Hamburg-Eppendorf).

M. Higueras (✉) · E.A. Ainsbury
Public Health England Centre for Radiation, Chemical and Environmental Hazards, Didcot, UK
e-mail: Manuel.Higueras-Hernaez@phe.gov.uk

E.A. Ainsbury
e-mail: Liz.Ainbsbury@phe.gov.uk

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_17

estimates. The calibrative density is the solution of the Bayesian inverse regression problem,

$$\mathrm{P}(D|y) \propto \mathrm{P}(D) \int \mathrm{L}(y|\Theta)\mathrm{P}(\Theta|D)\mathrm{d}\Theta. \tag{1}$$

The calibrative dose density is the product of the prior dose density by the posterior predictive distribution as function of the absorbed dose D. More details about the calibrative dose density and a review of Bayesian methods in biodosimetry can be found in Ainsbury–Vinnikov–Puig–Higueras–Maznyk–Lloyd–Rothkamm [1].

3 Whole Body Homogeneous Exposure

In Higueras–Puig–Ainsbury–Rothkamm [3], new Bayesian models which manage with equidispersed and overdispersed frequencies of chromosomal aberrations are introduced.

3.1 Poisson Responses

Assuming Poisson responses, the posterior distribution of the population mean of the yield of chromosomal aberrations is approximated to a normal, by asymptotic normality of the posterior distribution for large samples and the delta-method, i.e.,

$$\mu|D \sim \mathrm{N}\left(\mathrm{f}(D, \hat{\beta}), \nabla \cdot \hat{\Sigma} \cdot \nabla^{\mathrm{T}}\right), \tag{2}$$

where $\mathrm{f}(D, \beta)$ is the dose-response curve, $\hat{\Sigma}$ is the variance-covariance matrix of $\hat{\beta}$, and ∇ is the gradient of $\mathrm{f}(D, \beta)$ with respect to β. The calibrating density results in

$$\mathrm{P}(D|Y) \propto \mathrm{P}(D)\mathrm{P}(X_D = s), \tag{3}$$

where X_D is Hermite distributed. If $\mu|D$ is approximated by a gamma, X_D is Negative Binomial distributed.

This methodology provides a closed solution of the calibrative dose density independently of the dose-response curve: linear, linear-quadratic, Gompertz-type, among others.

3.2 Software: `radir`

A new `R` package called `radir` is available in the CRAN repository, to perform dose estimations for the Poisson in Sect. 3.1. It calculates the calibrative dose density for given: expression of the dose-response curve, hyper-parameters set, estimate of the parameter set, variance-covariance matrix of the estimation, total number of cells of the irradiated sample, number of chromosomal aberrations, prior distribution of the chromosomal aberration mean, prior distribution of the absorbed dose, and parameters of the distribution of the dose prior. For more details, see Moriña–Higueras–Puig–Ainsbury–Rothkamm [5].

3.3 Compound-Poisson Responses

Assuming that the yield of chromosomal aberrations follows a compound-Poisson distribution, e.g., Negative Binomial, Neyman A, or Hermite; the joint posterior of the population mean and the dispersion index is defined as follows,

$$\mu, \delta | D \sim \mathrm{N}_2 \left((\mathrm{f}(D, \hat{\beta}), \hat{\delta}), \nabla \cdot \hat{\Sigma} \cdot \nabla^{\mathrm{T}} \right). \tag{4}$$

The calibrative density can be defined directly and calculated by numerical integration. This is not computationally intensive, because the integral is always bivariate, over the absorbed dose D and the dispersion index δ.

Fixing δ by its maximum likelihood estimation, $\hat{\delta}$, the model is reduced and the mean prior is applied like in the Poisson models in Sect. 3.1. The resulting calibrative density is in terms of a compound-Hermite probability mass function, in case of a normal approximation; or a compound-Negative Binomial probability mass function, in case of a gamma approximation.

4 Homogeneous Partial Body Exposure

The zero-inflated Poisson models in Higueras–Puig–Ainsbury–Vinnikov–Rothkamm [4] are the Bayesian alternative for partial body exposure irradiation.

To decide if an irradiated sample of blood cells comes from a partial body exposure, the Bayarri–Berger–Datta [2] Bayes factor is proposed to contrast the zero-inflated Poisson against the Poisson assumptions:

$$BF = \frac{n_0!}{(n+1)!} \sum_{j=0}^{n_0} \frac{(n-j)!}{(n_0-j)!} (1 - j/n)^{-(s+1/2)}, \tag{5}$$

where n, n_0 and s are, respectively, the sample size and frequency of zeros, and the sum of the total number of chromosomal aberrations.

Once the ZIP assumption is supported, the frequency of aberrations per cell is zero-inflated distributed,

$$Z \sim \mathrm{ZIP}\left(\mu,\ \frac{1-F}{F\mathrm{e}^{-D/d_0}-F+1}\right), \tag{6}$$

where F is the fraction of the body irradiated, and d_0 is the lethal dose. Analogously to Higueras–Puig–Ainsbury–Vinnikov–Rothkamm [3], the mean prior is approximated by a gamma distribution,

$$\mu|D \sim \mathrm{Gamma}\left(\frac{\mathrm{f}(D,\hat{\beta})^2}{\nabla\cdot\hat{\Sigma}\cdot\nabla^{\mathrm{T}}},\ \frac{\mathrm{f}(D,\hat{\beta})}{\nabla\cdot\hat{\Sigma}\cdot\nabla^{\mathrm{T}}}\right). \tag{7}$$

An application of Bayes' theorem shows the expression of the likelihood of D, F, and d_0 for the given test data,

$$\mathrm{L}(y|D,F,d_0) \propto (F\mathrm{e}^{-D/d_0}-F+1)^{-n}\sum_{j=1}^{n_0}\binom{n_0}{j}\frac{F^{n-j}(1-F)^j}{(n-j)^s}\mathrm{P}(X_j=s|D), \tag{8}$$

where X_j is a random variable Negative Binomial distribution with mean and variance depending on j and D.

Considering D, F, and d_0 as independent random variables, their prior distributions are defined

$$D \sim \mathrm{Gamma}\left(\frac{\hat{D}^2}{\hat{\sigma}^2_{\hat{D}}},\frac{\hat{D}}{\hat{\sigma}^2_{\hat{D}}}\right),\quad F\sim\mathcal{U}(0,1),\quad d_0\sim\mathcal{U}(2.7,3.5). \tag{9}$$

And, applying again the Bayes' theorem, the joint posterior density

$$\mathrm{P}(D,F,d_0|y) = \frac{\mathrm{L}(y|D,F,d_0)\mathrm{P}(D,F,d_0)}{\displaystyle\int \mathrm{L}(y|D,F,d_0)\mathrm{P}(D,F,d_0)\mathrm{d}D\mathrm{d}F\mathrm{d}d_0} \tag{10}$$

has a non-tractable form. The acceptance-rejection sampling is used to simulate the posterior distribution.

5 Conclusions

Novel solutions for statistical analysis of cytogenetic biological dosimetry data have been developed. These new models are in the Bayesian framework, have been applied in practical cytogenetic dose estimation (see Higueras–Puig–Ainsbury–

Rothkamm [3], Higueras–Puig–Ainsbury–Vinnikov–Rothkamm [4], and Moriña–Higueras–Puig–Ainsbury–Rothkamm [5]), and some of them have been implemented in the `R` statistical software for biodosimetry laboratory researchers. These new solutions lead to more accurate quantification of statistical uncertainty associated with cytogenetic dose estimates. The techniques described have now been implemented into the UK's commercial biodosimetry service. This work provides a framework for further improvements in retrospective dose estimation to support EU emergency preparedness and response.

References

1. E.A. Ainsbury, V.A. Vinnikov, P. Puig, M. Higueras, N.A. Maznyk, D.C. Lloyd, and K. Rothkamm, "Review of Bayesian statistical analysis methods for cytogenetic radiation biodosimetry, with a practical example", *Radiation Protection Dosimetry* **162**(3) (2014), 185–196.
2. M.J. Bayarri, J.O. Berger, and G.S. Datta, "Objective bayes testing of Poisson versus inflated Poisson models", in "Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh", Institute of Mathematical Statistics (2008), 105–121, available at http://projecteuclid.org/euclid.imsc/1209398464.
3. M. Higueras, P. Puig, E.A. Ainsbury, and K. Rothkamm, "A new inverse regression model applied to radiation biodosimetry", *Proceedings Mathematical, Physical, and Engineering Sciences* **471**(2174) (2015), 20140588.
4. M. Higueras, P. Puig, E.A. Ainsbury, V.A. Vinnikov, and K. Rothkamm, "A new Bayesian model applied to cytogenetic partial body irradiation estimation", *Radiation Protection Dosimetry* **168**(3) (2016), 330–336.
5. D. Moriña, M. Higueras, P. Puig, E.A. Ainsbury, and K. Rothkamm, "`radir` package: an `R` implementation for cytogenetic biodosimetry dose estimation", *Journal of Radiological Protection* **35**(3) (2015), 557–569.

Empirical Assessment of Gene Expression Biomarkers for Radiation Exposure

Adetayo Kasim, Nolen Joy Perualila, and Ziv Shkedy

Abstract This paper discusses the relevance of surrogate marker validation method for empirical assessment of gene expression biomarkers for radiation exposure.

1 Introduction

Several studies have investigated the use of gene expression biomarkers for biodosimetry, in order to overcome the limitations of established ionizing radiation biomarkers such as dicentric scores; see, for example, Lu–Hsu–Lai–Tsai–Chuang [2]. Although gene expression biomarkers may be used in a timely and less tedious manner, it is difficult to establish direct links between these biomarkers and health. This is partly due to the multiple layers of systems between stimuli and changes in *mRNA* transcriptions. This paper examines how a well-established method for surrogate marker validation in drug development can be used to evaluate gene expression biomarkers as surrogates for established radiation biomarkers. Our aim is to identify genes that are correlated with dicentric scores and that can be used in a predictive model to classify samples according to their radiation dose. Assuming an in-vitro setting where both dicentric scores and gene expression data are available, the goal is to evaluate whether gene expression biomarker (X) is a good surrogate of dicentric scores (Y) given radiation dose (Z). The modelling framework is described in Fig. 1.

Depending on the nature of the three-way associations in Fig. 1, gene expression biomarkers can be grouped into four types described in Table 1. Gene expression biomarkers in groups (a) and (b) are correlated with dicentric scores, and can be used to classify samples according to their radiation dose. More importantly, the

A. Kasim (✉)
Durham University, Durham, UK
e-mail: a.s.kasim@durham.ac.uk

N.J. Perualila · Z. Shkedy
Hasselt University, Hasselt, Belgium
e-mail: nolenjoy.perualila@uhasselt.be

Z. Shkedy
e-mail: ziv.shkedy@uhasselt.be

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_18

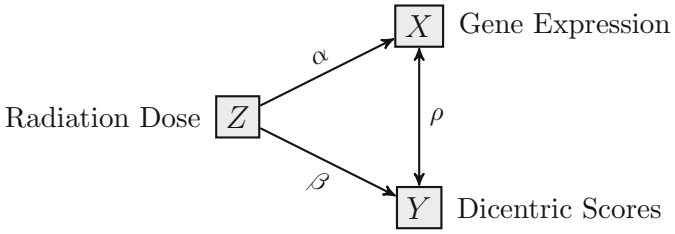
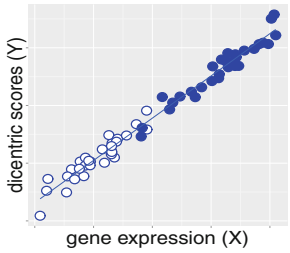
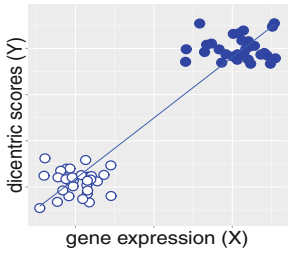
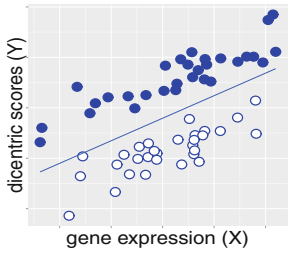
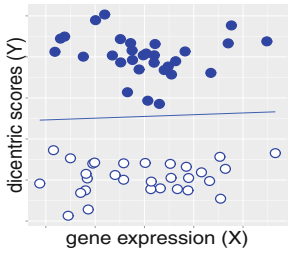


Fig. 1 Three-way associations between gene expression biomarker (X), dicentric scores (Y) and radiation dose (Z)

Table 1 A hypothetical setting showing types of gene expression biomakers based on association with dicentric scores and radiation dose

	$\rho_j \neq 0$	$\rho_j = 0$
$\beta \neq 0$ & $\alpha_j \neq 0$	 **(a)** X and Y are correlated and the gene expression can be used to classify samples according to radiation dose (Z)	 **(b)** X and Y are conditionally independent given radiation dose (Z) and the gene expression can also be used to classify samples according to radiation dose (Z)
$\beta \neq 0$ & $\alpha_j = 0$	 **(c)** X and Y are correlated, but the gene expression is not predictive of radiation dose (Z)	 **(d)** X and Y are uncorrelated and the gene expression is not predictive of radiation dose (Z)

correlation with dicentric scores in group (b) is induced solely by radiation exposure. A good gene expression biomarker with link to chromosomal anomalies due to radiation exposure will be of either type (a) or (b). In both scenarios, gene expression can be used to classify samples according to their radiation dose. Gene expression

in group (c) cannot predict radiation dose and those in group (d) are not correlated with dicentric scores.

2 Method

Assume that both the gene expression data X and Dicentric scores Y are normally distributed. For simplicity, let Z denote high and low radiation dose. The simultaneous joint modelling of gene expression biomarker (for simplicity we dropped indices for both samples and genes), dicentric scores and radiation exposure can be formulated as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X + \alpha Z \\ \mu_Y + \beta Z \end{pmatrix}, \Sigma\right], \tag{1}$$

where the error terms have a joint zero-mean normal distribution with a structured covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}. \tag{2}$$

Parameters α and β represent the mean differences between high and low radiation dose for gene expression biomarker and dicentric score, respectively. The intercepts are μ_X and μ_Y. Thus, the association between gene expression biomarker and dicentric scores can be obtained using adjusted association (see Buyse–Molenberghs [1]), a coefficient that is derived from the covariance matrix Σ:

$$\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}. \tag{3}$$

This quantifies the correlation between gene expression biomarker and dicentric scores after accounting for radiation exposure.

3 Case Study from Drug Discovery

This case study (EGFR project) focuses on inhibition of the epidermal growth factor receptor, which has been identified in many human epithelial cancers, colorectal, breast, pancreatic, non-small cell lung and brain cancer. Thirty-five compounds with a macrocycle structure were profiled in order to identify compounds with similar biological effects as the current EGFR inhibitors, gene expression profiles (X) are available for 3595 genes. Moreover, a total of 138 unique profiles of chemical substructures (Z) were identified for this compound set. EGFR inhibi-

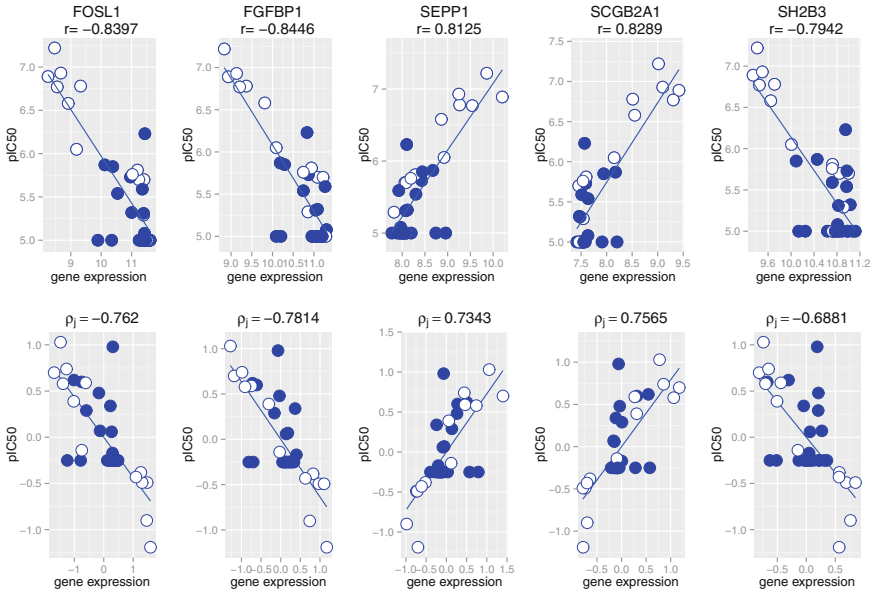


Fig. 2 Top 5 genes with strong association with EGFR inhibition through pIC50 and that discriminates between chemical substructures. The *first row* shows the three-way associations between gene expression, pIC50 and a chemical substructure. The *second row* represents adjusted association between pIC50 and gene expression after removing the effect of the chemical substructure

tion was quantified based on pIC50 (Y). Note that in this specific case study we have multiple genes and multiple chemical structures, but the analysis was done per gene per chemical structure. Figure 2 shows the expression data of the top 5 genes that were associated with EFGR inhibition through pIC50 and that were also discriminatory between the chemical substructure. Annotating chemical substructures (fingerprints) in terms of toxicity and safety has the potential to reduce failure rates in drug trials; see Verbist–Klambauer–Vervoort–Talloen–QSTAR-Consortium–Shkedy–Thas–Bender–Goehlmann–Hochreiter [3].

4 Discussion

Validation of transcriptional and translational biomarkers of exposure needs more than just an empirical assessment based on statistical methodology. However, the methods discussed in this paper can help to focus on gene expression biomarkers with high chances of success in external validation studies by relying on the strength of cytogenetic biomarkers. The method has been discussed in an in-vitro setting, but there are other approaches of data integration that can be used to integrate dif-

ferent sources of data in radiation research including demographic, socio-economic, clinical and epidemiological data. The method can be used on any type of data after minor adaptation to reflect distributional assumptions. We hope that this paper will contribute to on-going scientific discussion on identifying and validating transcriptional and translational biomarkers for radiation exposure.

Acknowledgements We thank Janssen Pharmaceutica NV, Beerse, Belgium and the QSTAR Consortium for the case study in drug development.

References

1. M.G. Buyse and G. Molenberghs, "The validation of surrogate endpoints in randomized experiments", *Biometrics* **54** (1998), 186–201.
2. T.P. Lu, Y.Y. Hsu, L.C. Lai, M.H. Tsai, and E.Y. Chuang, "Identification of gene expression biomarkers for predicting radiation exposure", *Scientific Reports* **4** (2014), 6293.
3. B. Verbist, G. Klambauer, L. Vervoort, W. Talloen, QSTAR Consortium, Z. Shkedy, O. Thas, A. Bender, H. Goehlmann, and S. Hochreiter, "Using transcriptomics to guide lead optimization in drug discovery projects: lessons learned from the qstar project", *Drug Discovery Today* **20** (2015), 505–513.

Poisson-Weighted Estimation by Discrete Kernel with Application to Radiation Biodosimetry

Célestin C. Kokonendji, Nabil Zougab, and Tristan Senga-Kiessé

Abstract Reminding the framework of discrete smoothing using discrete associated kernel methods, binomial kernel with local Bayesian bandwidth selection is presented, for estimating a probability mass function under a Poisson-weighted assumption (Senga-Kiessé et al. Comput Stat 31:189–206, 2016, [11]). Model diagnostics are evoked between three approaches: parametric, nonparametric and semiparametric. Finally, some applications are done on real count datasets of low and high radiation doses in biodosimetry, as alternatives to the parametric approaches in Pujol et al. (PLoS ONE 9(12):e114137, 2014, [9]).

1 Introduction and Motivations

It is known that low-dose radiation may cause heart disease and stroke. However, the mechanisms for such effects are unclear. Note that the activity administered must be such that the resulting radiation dose is as low as reasonably achievable bearing in mind the need to obtain the intended diagnostic result; e.g., Magnetic Resonance Imaging.

Thus, for modelling, the authors in [9] used a parametric Poisson-weighted distribution which could be a good representation of any count data distribution (e.g., [3, 4]). But the real problem is to choose the Poisson-weighted function. Here is how it is written

C.C. Kokonendji (✉)
Laboratoire de Mathématiques de Besançon - UMR 6623 CNRS-UFC,
University of Franche-Comté, 16 Route de Gray, 25030 Besançon Cedex, France
e-mail: celestin.kokonendji@univ-fcomte.fr

N. Zougab
University of Tizi-Ouzou, LAMOS, Route de Targa-Ouzemmour, 06000 Béjaïa, Algeria
e-mail: nabilzougab@yahoo.fr

T. Senga-Kiessé
INRA Rennes, UMR 1069 SAS, 65 Rue de Saint Brieuc, 35042 Rennes, France
e-mail: Tristan.SengaKiesse@rennes.inra.fr

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_19

$$f(x) := \mathbb{P}(X = x) = p(x; \lambda)\omega(x; \theta) =: f_{\lambda,\omega}(x), \quad x \in \{0, 1, 2, \ldots\} =: \mathbb{N}, \quad (1)$$

where $x \mapsto p(x; \lambda) := e^{-\lambda}\lambda^x/x!$ is the probability mass function (pmf) of the Poisson distribution with mean parameter $\lambda > 0$, and $\omega := \omega(\cdot\,; \theta)$ describes the parametric mechanism which is considered to be unknown (for the moment) with $\theta \in \mathbb{R}$. This ω could represent the repairing process of the chromosomes in the cell after a dose of radiation. Since parametric model means that its form is known but not the parameter values, the authors of [9] proposed the following quadratic polynomial function

$$x \mapsto \omega(x; \theta) = 1 + \theta x^2, \quad \forall x \in \mathbb{N}. \quad (2)$$

In principle, a choice of such parametric model (1) and (2) needs a good knowledge of the subject to be treated. The expected forms of any parametric model are always regular.

An alternative approach is to consider a nonparametric method for estimating the pmf $x \mapsto f(x)$ of (1). It requires any particular form; that is it is more free and more flexible for data. Hence, we can say "let talk the data". For instance, let us consider the discrete kernel method which is an evolution of the histogram for discrete data (e.g., [5, 7, 8] for pmf) defined as follows. Let $X_1, \ldots, X_n$ be an n-sample of iid unknown pmf f on $\mathbb{T} \subseteq \mathbb{N}$, then a discrete associated kernel estimator of f is

$$\widetilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i), \quad x \in \mathbb{T} \subseteq \mathbb{N} \ \ (\subset \mathbb{R}^d); \quad (3)$$

$h > 0$ is the smoothing parameter (or bandwidth) such that $h = h(n) \to 0$ as $n \to \infty$; $K_{x,h}(\cdot)$ is the discrete associated kernel function (or discrete smoother), intrinsically connected to the target x and h, and which is the pmf of a discrete random variable $\mathcal{Z}_{x,h}$ on $\mathbb{S}_x$ such that, for all $x \in \mathbb{T}$: $\mathbb{S}_x \supseteq \{x\}$, $\lim_{h\to 0}\mathbb{E}(\mathcal{Z}_{x,h}) = x$, $\lim_{h\to 0} Var(\mathcal{Z}_{x,h}) = 0$ or in $[0, 1)$. This definition is extended to multivariate and also to the classical continuous (symmetric) cases. Depending on the support $\mathbb{T}$ ($\subseteq \mathbb{Z}$) of f to be estimated, some examples of discrete (i.e. count or categorial) are given in Table 1.

Table 1 Some discrete associated kernels; see [5]

Discrete kernel	$K_{x,h}(u)$	$\mathbb{S}_x$
Dirac	$\delta_x(u), \quad h := 0$	$\mathbb{N}$
DiracDU	$(1-h)\delta_x(u) + h/(c-1)(1-\delta_x(u))$	$\{0, 1, \ldots, c-1\}$
DiracDUextZ	$(1-h)\delta_x(u) + h^{\lvert u-x\rvert}(1-h)/2(1-\delta_x(u))$	$\mathbb{Z}$
Discrete triangular	$[(a+1)^h - \lvert u-x\rvert^h]/P(a,h),\ a \in \mathbb{N}$	$\{x, x \pm 1, \ldots, x \pm a\}$
Binomial	$\mathcal{B}in(x+1; (x+h)/(x+1))(u)$	$\{0, 1, \ldots, x+1\}$

Note that the DiracDU kernel is appropriated for categorial data and the binomial one is for count data having small and moderate sample sizes (as for all discrete kernels, except the Dirac one). See [5, 8] for further details and properties such as the criterion of "mean integrated squared error (MISE)" and normalization of estimated pmf by binomial and discrete triangular kernels. The discrete associated kernel $K_{x,h}(\cdot)$ works as a stochastic distance around the target x with a dispersion (or scale) parameter h. The choice of h given a dataset is crucial to avoid over- and under-smoothing. Several techniques of selection of h can be found in the literature: the well-known one is the cross-validation, but we have recently developed some Bayesian approaches; see [11–13].

Finally, a compromise between parametric (1) and nonparametric models is to consider the semiparametric model having the Poisson-weighted function (2) without any specific form; see, e.g., [6] using cross-validation method for selecting the bandwidth. In practice, the semiparametric model eliminates some noises in the data analysis. For improving the parametric model (1) with (2) of [9], the semiparametric model seems to be the most interesting and appropriated approach by using our recent development in [11] with Bayesian estimation of bandwidth (instead of the cross-validation in [10]). Thus, the next section is devoted to the methodology with applications to the datasets of [9]. The last section gives some concluding remarks.

2 Methodology and Results

Let us assume that the pmf f of (1) is here written as

$$f(\cdot) = p(\cdot;\lambda)\omega(\cdot) =: f_{\lambda,\omega}(\cdot) \quad \text{on } \mathbb{T} \subseteq \mathbb{N},$$

where $p(\cdot;\lambda)$ is the parametric Poisson model and $\omega(\cdot)$ the nonparametric Poisson-weighted function. From [6], the corresponding semiparametric Poisson-weighted estimation of f by discrete associated kernel is given by

$$\widehat{f}_{n,h}(x) = p(x;\widehat{\lambda}_n) \times \frac{1}{n}\sum_{i=1}^{n} \frac{K_{x,h}(X_i)}{p(X_i;\widehat{\lambda}_n)} = \frac{1}{n}\sum_{i=1}^{n} \frac{p(x;\widehat{\lambda}_n)}{p(X_i;\widehat{\lambda}_n)} K_{x,h}(X_i), \quad x \in \mathbb{T}. \tag{4}$$

Indeed, the estimated parameter $\widehat{\lambda}_n$ of λ is obtained by the maximum likelihood method as $\widehat{\lambda}_n = (X_1 + \cdots + X_n)/n$; it converges to the true value λ_0 of the Poisson part satisfying $\lambda_0 = \text{argmin}_\lambda \sum_x f(x) \log[f(x)/p(x;\lambda)]$ from Kullback–Leibler distance. The discrete associated kernel $K_{x,h}(\cdot)$ is the binomial kernel function:

$$B_{x,h}(X_i) = \frac{(x+1)!}{X_i!(x+1-X_i)!}\left(\frac{x+h}{x+1}\right)^{X_i}\left(\frac{1-h}{x+1}\right)^{x+1-X_i},$$

with $X_i \leq x+1$. From [13] we will select the bandwidth $h \in [0,1]$ by the Bayesian local approach defined as follows:

$$\widetilde{h}_{B\ell}(x) \propto \int_{h>0} h\,\pi(h)\,\widetilde{f}_{n,h}(x)\,dh, \quad x \in \mathbb{T},$$

where $\widetilde{f}_{n,h}(x)$ is given in (3). Consider the (conjugate) beta prior density of h:

$$\pi(h) = (\mathcal{B}(\alpha,\beta))^{-1} h^{\alpha-1}(1-h)^{\beta-1}, \quad h \in [0,1],$$

where $\alpha, \beta > 0$ and $\mathcal{B}(\cdot,\cdot)$ is the Beta function. Then, for all $x \in \mathbb{N}$, one has exactly

$$\widetilde{h}_{B\ell}(x) = \frac{\sum_{i=1}^{n}\sum_{k=0}^{X_i} \frac{x^k}{(x+1-X_i)!k!(X_i-k)!}\mathcal{B}(X_i+\alpha-k+1, x+\beta+1-X_i)}{\sum_{i=1}^{n}\sum_{k=0}^{X_i} \frac{x^k}{(x+1-X_i)!k!(X_i-k)!}\mathcal{B}(X_i+\alpha-k, x+\beta+1-X_i)},$$

with $X_i \leq x+1$. Now, it is necessary to normalize (globally or adaptively) the estimated pmf $\widehat{f}_{n,h=\widetilde{h}_{B\ell}(x)}$ before the sequel. For instance, the practical performance of the estimator (4) is measured through $\mathrm{ISE}_0(\widehat{f}_{n,h}) = \sum_{x\in\mathbb{T}}\{\widehat{f}_{n,h}(x) - f_0(x)\}^2$, where f_0 is the empirical (or naive) estimator of f. Finally, from (4) we investigate the model diagnostics for checking the adequacy of the model by examining a plot of $x \mapsto \widehat{\omega}(x)$ or $Z(x) := \log\widehat{\omega}(x) = \log[\widehat{f}_{n,h}(x)/p(x;\widehat{\lambda}_n)]$ with a pointwise confidence band of ± 1.96; that is to see whether or not $\omega(x) = 1$ is reasonable (e.g., $< 5\%$ for pure nonparametric, in $[5\%, 95\%]$ for semiparametric, and $>95\%$ for full parametric $p(\cdot;\widehat{\lambda}_n)$ models).

Table 2 presents a summary of the numerical results of the semiparametric Poisson-weighted estimation of datasets in [9] through binomial kernel, and using Bayesian local to select bandwidth. The dispersion index [2] which is the ratio of sample variance by the sample mean (<1) shows the underdispersion for all the datasets. This departure with respect to the classical Poisson distribution points out the use of other models like (1) and (4). The corresponding model diagnostics confirm that the semiparametric Poisson-weighted models are more appropriated than the parametric ones. This fact can be seen through the ISE_0 values which could be unappropriated for parametric model.

Table 2 Semiparametric analysis of datasets from [9]

Dose (in Gy)	25	20	15	10	7	5	3	1	0.5
Dispersion index	0.94	0.83	0.81	0.41	0.67	0.62	0.82	0.98	0.96
Semiparam. $\mathrm{ISE}_0 10^{-3}$	4.59	2.27	9.59	3.02	2.60	7.91	1.37	0.0096	0.12
Parametric $\mathrm{ISE}_0 10^{-3}$	12.65	5.30	11.98	25.15	7.84	26.24	2.79	0.0093	0.10
Diagnostic (%)	94.4	94.1	93.8	81.8	81.8	90.0	71.4	80.0	75.0

Now, if we consider the corresponding nine graphical representations of the estimated Poisson-weighted functions $x \mapsto \widehat{\omega}(x) = \widehat{f}_{n,h}(x)/p(x;\widehat{\lambda}_n)$ then it is clear that there exist some differences with the parametric Poisson-weighted function (2) of [9]. In fact, the forms of $x \mapsto \widehat{\omega}(x)$ are not regular. One can observe some mixtures of three, two or only one quadratic polynomial functions with high negative coefficient. This means that the parametric Poisson-weighted function (2) is not completely unsuitable; however, the new approach is more adequate, taking into account the reality of datasets.

3 Concluding Remarks

Directed by the Poisson distribution, we have estimated the count weighted functions for modelling mechanisms of high-low-dose radiation in biodosimetry. The semiparametric approach using the discrete kernel with Bayesian local estimation for bandwidth improves suitably the regular parametric model (1) and (2) of [9], and eliminates some noises if we have used the purely nonparametric estimator (3). However, an interpretation of the estimated Poisson-weighted functions needs to be found in the way of a birth-death process or process in queuing theory because of the underdispersion of datasets. More generally, the discrete nonparametric and semiparametric approaches can be used for real datasets in regression models; see, for example, [1] for cross-validation technique and also [11] with the Bayesian selection of bandwidth.

Acknowledgements The first author would like to thank Pere Puig and Amanda Fernández-Fontelo for numerous discussions on this subject based on the two papers [9, 11].

References

1. B. Abdous, C.C. Kokonendji, and T. Senga-Kiessé, "On semiparametric regression for count explanatory variables", *J. Statist. Plann. Inference* **142** (2012), 1537–1548.
2. C.C. Kokonendji, "Over- and underdispersion models", in "Methods and Applications of Statistics in Clinical Trials", Vol. **2**, Chap. 30 (2014), 506–526.
3. C.C. Kokonendji, D. Mizère, and N. Balakrishnan, "Connections of the Poisson weight function to overdispersion and underdispersion", *J. Statist. Plann. Inference* **138** (2008), 1287–1296.
4. C.C. Kokonendji and M. Pérez-Casany, "A note on weighted count distributions", *J. Statist. Th. Appl.* **11** (2012), 337–352.
5. C.C. Kokonendji and T. Senga-Kiessé, "Discrete associated kernels method and extensions", *Statistical Methodology* **8** (2011), 497–516.
6. C.C. Kokonendji, T. Senga-Kiessé, and N. Balakrishnan, "Semiparametric estimation for count data through weighted distributions", *J. Statist. Plann. Inference* **139** (2009), 3625–3638.
7. C.C. Kokonendji, T. Senga-Kiessé, and C.G.B. Demétrio, "Appropriate kernel regression on a count explanatory variable and applications", *Adv. Appl. Statist.* **12** (2009), 99–125.
8. C.C. Kokonendji and S.S. Zocchi, "Extensions of discrete triangular distribution and boundary bias in kernel estimation for discrete functions", *Statist. Probab. Lett.* **80** (2010), 1655–1662.

9. M. Pujol, J.F. Barquinero, P. Puig, R. Puig, M.R. Caballin, and L. Barrios, "A new model of biodosimetry to integrate low and high doses", *PLoS ONE* **9**(12) (2014), e114137.
10. T. Senga-Kiessé and D. Mizère, "Weighted Poisson and semiparametric kernel models applied for a parasite growth", *Aust. New Zealand J. Statist.* **55** (2012), 1–13.
11. T. Senga-Kiessé, N. Zougab, and C.C. Kokonendji, "Bayesian estimation of bandwidth in semiparametric kernel estimation of unknown probability mass and regression functions of count data", *Comput. Statist.* **31** (2016), 189–206.
12. N. Zougab, S. Adjabi, and C.C. Kokonendji, "Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation", *Comput. Statist. Data Anal.* **75** (2014), 28–38.
13. N. Zougab, S. Adjabi, and C.C. Kokonendji, "Comparison study to bandwidth selection in binomial kernel estimation using Bayesian approaches", *J. Statist. Theor. Pract.* **10** (2016), 133–153.

R Implementation of the Excess Relative Rate Model: Applications to Radiation Epidemiology

David Moriña and Elisabeth Cardis

Abstract In many radiation environmental and occupational contexts, the excess relative rate (ERR) of disease is modelled related to the exposure in an additive fashion, in contrast to the usual exponential rate model. Most of the available software packages are restricted to models of the log-linear form, with the exception of *Epicure*, a commercial software package commonly used for fitting linear relative rate models. Recently, some macros allowing the fitting of the ERR have been developed for *SAS*. However, to the best of our knowledge, no methods have been developed yet for the widely used *R* software. In this paper, we introduce the package *linERR*, which aims to allow the user to fit linear rate models within the framework of the *R* programming language.

1 Introduction

Usual approaches to the analysis of cohort and case control data often follow from risk-set sampling designs, where at each failure time a new risk set is defined, including the index case and all the controls that were at risk at that time. That kind of sampling designs are usually related to the Cox proportional hazards model, available in most standard statistical packages but limited to log-linear models, except *Epicure* (see [2]) which is of the form

$$\log(\phi(z, \beta)) = \beta_1 z_1 + \cdots + \beta_k z_k, \tag{1}$$

D. Moriña (✉) · E. Cardis
ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
e-mail: dmorina@creal.cat

E. Cardis
e-mail: ecardis@creal.cat

D. Moriña · E. Cardis
Universitat Pompeu Fabra (UPF), Barcelona, Spain

D. Moriña · E. Cardis
CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_20

where z is a vector of explanatory variables and ϕ is the rate ratio. This implies exponential dose-response trends and multiplicative interactions, which may not be the best exposure-response representation in some cases, such as radiation exposures. One model of particular interest, especially in radiation environmental and occupational epidemiology, is the ERR model

$$\phi(z, \alpha, \beta, \text{dose}) = g(z, \beta)(1 + \alpha f(\text{dose})). \tag{2}$$

The ERR model represents the excess relative rate per unit of exposure and $z = z_1, \ldots, z_k$ are covariates. Estimation of a dose-response trend under a linear relative rate model implies that, for every 1-unit increase in the exposure metric, the rate of disease increases (or decreases) in an additive fashion. The modification of the effect of exposure in linear relative rate models by a study covariate m can be assessed by including a log-linear subterm for the linear exposure effect [3, 7], implying a model of the form

$$\phi(z, \alpha, \beta, \text{dose}) = e^{\beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k}(1 + \alpha f(\text{dose})). \tag{3}$$

In general, the likelihood contribution of each risk set can be written as

$$L(\beta_1, \ldots, \beta_k, \alpha) = \frac{\phi_{\text{case}}(z_1, \ldots, z_k, \beta_1, \ldots, \beta_k, \alpha, \text{dose})}{\phi_{\text{whole risk set}}(z_1, \ldots, z_k, \beta_1, \ldots, \beta_k, \alpha, \text{dose})}. \tag{4}$$

Recently, some papers appeared describing how to fit general relative risk models using *SAS* in the context of logistic and Poisson regression [5, 6] and also in the Cox regression framework [1]. Despite its biological interpretation advantages, the ERR model has worse statistical properties than the standard Cox model. In particular, parameter estimates may be inaccurate if the number of events is not large, but a large number of events will involve large computational time. It is possible that parameter estimates fall under the lower boundary of the feasible region $-1/z_{max}$, where z_{max} is the maximum cumulated value for variable z, and in such cases no estimate of the standard error can be obtained.

2 The Package *linERR*

The package has been designed to integrate the ERR model within a more complete and flexible freely available statistical package like *R*; see [4]. Results are presented in a very similar way to that of *Epicure*. The likelihood function (4) has been written in **C** to improve computing times. It can handle a broad class of relative risk models mixing linear and log-linear terms as introduced in (3). It also handles lagged times, when it is convenient to exclude the exposure in a time period before registration of the outcome of interest. The main function is *fit.linERR()*, which needs the following input parameters:

(i) `formula`: An object of class *formula* (or one that can be coerced to that class), i.e., a symbolic description of the model to be fitted. The response must be a survival object as returned by the *Surv()* function, and the log-linear and linear terms are separated by the character "|". Strata are defined using the *strata()* function. See Sect. 3 for more details.
(ii) `beta`: Starting values for parameter estimates. Its default value is *NULL*.
(iii) `data`: Data frame containing the cohort.
(iv) `ages`: Age at each exposure.
(v) `lag`: Lag to be applied (its default value is zero).

Profile likelihood based confidence intervals can be computed by means of function *ERRci()*, with the following parameters:

(i) `object`: An object of class *fit.linERR*.
(ii) `prob`: Level of confidence (its default value is 0.95).

The profile log-likelihood function can be plotted with the usual method *plot* applied to an object returned by *fit.linERR*, with an option to highlight the profile likelihood confidence intervals. The function *plot* uses the following arguments:

(i) `object`: An object of class *fit.linERR*.
(ii) `lower`: Lower value of the parameter.
(iii) `upper`: Upper value of the parameter.
(iv) `ci`: Highlight the profile likelihood confidence interval (its default value is *NULL*).

3 Examples

Three cohorts with 10000 subjects and a different number of cases, and an ERR model was fitted in *R* using the code

```
> fit.linERR(Surv(entryage, exitage, leu)~1|dose1+dose2+dose3+dose4+dose5+dose6+
+         dose7+dose8+dose9+dose10+dose11+dose12+dose13+dose14+dose15+dose16+
+         dose17+dose18+dose19+dose20+dose21+dose22+dose23+dose24+dose25+dose26+
+         dose27+dose28+dose29+dose30+dose31+dose32, beta=NULL, data=ex1,
+         ages=ex1[, 7:38])
```

The same model was fitted in *Epicure* and the obtained estimates, standard errors and computing times are shown in Table 1.

A cohort consisting of 150000 subjects was generated, and a linear excess relative risk model including three covariates and different strata was fitted using the code

```
> fit.linERR(Surv(entryage, exitage, leu)~ds+ds2+ds3 | dose1+dose2+dose3+dose4+
+         dose5+dose6+dose7+dose8+dose9+dose10+dose11+dose12+dose13+dose14+
+         dose15+dose16+dose17+dose18+dose19+dose20+dose21+dose22+dose23+
+         dose24+dose25+dose26+dose27+dose28+dose29+dose30+dose31+dose32+
+         strata(sid), beta=NULL, data=ex2, ages=ex2[, 9:40])
```

Table 1 Results from Epicure and R

	Cases	β	$\hat{\beta}$ (Epicure)	$\hat{\beta}$ (R)	Computing time (Epicure)	Computing time (R)
Cohort 1	13	0.04	0.07213 (0.1714)	0.07140 (0.1374)	11.64	2.26
Cohort 2	19	0.1	0.3745 (0.8677)	0.3795 (0.8165)	27.34	3.97
Cohort 3	14	0	−0.0242 (NA)	−0.0242 (NA)	10.47	0.4

Table 2 Results from Epicure and R

	Cases	β	$\hat{\beta}$ (Epicure)	$\hat{\beta}$ (R)	Computing time (Epicure)	Computing time (R)
Cohort 1	84	0.1	0.1084 (0.09263)	0.1077 (0.0866)	176.8	123.74

Again, the same model was fitted in *Epicure* and the obtained estimates, standard errors and computing times are shown in Table 2.

We can see that, in all the considered cases, computing time using the *linERR* package is less than in *Epicure*. The standard errors are slightly different due to different maximization algorithms. The maximum likelihood estimates can be obtained by

```
> summary(fit.ex2)
...
Parameter Summary Table:
            Estimate  Std. Error Test Stat.      p-value
dose     0.107639724 0.086605628  1.2428722 2.139149e-01
ex2$ds   4.351506257 0.281627035 15.4513087 7.392707e-54
ex2$ds2  0.048913212 0.245572769  0.1991801 8.421219e-01
ex2$ds3 -0.002126584 0.008645213 -0.2459840 8.056946e-01


AIC:  1291.527
Deviance:  1283.527
Informative risk sets:  84
```

The 95% profile likelihood-based confidence interval can be computed with the command

```
> ERRci(fit.ex2)
 lower 2.5% upper 97.5%
 0.01352891  0.72524683
```

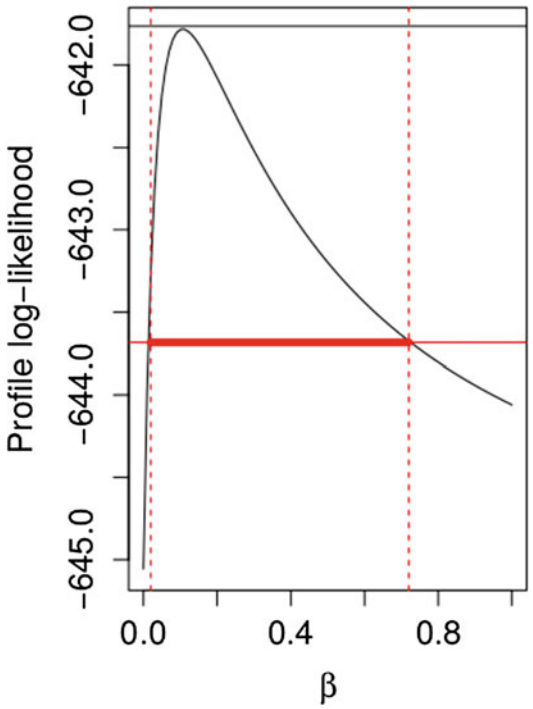


Fig. 1 Profile log-likelihood function

while the command

```
> plot(fit.ex2, 0, 1, 0.95)
```

produces the plot represented in Fig. 1.

Acknowledgements We would like to thank Patrycja Gradowska and Michael Hauptmann from Netherlands Cancer Institute (NKI) for their contribution to the generation of the cohorts used in the examples.

References

1. B. Langholz and D.B. Richardson, "Fitting general relative risk models for survival time and matched case-control analysis", *American Journal Epidemiology* **171**(3) (2010), 377–383.
2. D.L. Preston, J.H. Lubin, D.A. Pierce, and M.E. McConney, "Epicure: user's guide", HiroSoft International Corporation, Seattle, WA (1993).
3. D.L. Preston, Y. Shimizu, D.A. Pierce, A. Suyama, and K. Mabuchi, "Studies of mortality of atomic bomb survivors. Report 13: solid cancer and noncancer disease mortality: 1950–1997", *Radiation Research* **160**(4) (2003), 381–407.

4. R Core Team, “R: a language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria (2015), http://www.R-project.org/.
5. D.B. Richardson, “A simple approach for fitting linear relative rate models in SAS”, *American Journal of Epidemiology* **168**(11) (2008), 1333–1338.
6. D.B. Richardson and J.S. Kaufman, “Estimation of the relative excess risk due to interaction and associated confidence bounds”, *American Journal Epidemiology* **169**(6) (2009), 756–760.
7. E. Ron, J.H. Lubin, R.E. Shore, K. Mabuchi, B. Modan, L.M. Pottern, A.B. Schneider, M.A. Tucker, and J.D. Boice, “Thyroid cancer after exposure to external radiation: a pooled analysis of seven studies”, *Radiation Research* **141**(3) (1995), 259–277.

Uncertainty Considerations Following a Mechanistic Analysis of Lung Cancer Mortality

Ignacio Zaballa and Markus Eidemüller

Abstract Lung cancer mortality after radon exposure in the Wismut cohort was analysed with the two-stage clonal expansion (TSCE) model. Careful examination of the results of this study suggests that model misspecification and individual error in radon exposure estimates may be leading to large uncertainties in the estimation of risk.

1 Introduction

Exposure to radon and its progeny has been recognized as a cause of lung cancer for many decades; see [1]. The workers of the Wismut company were exposed to high concentrations of radon and its progeny chiefly working underground or in uranium ore processing facilities. Radon-222 results from the natural decay of uranium-238, and emanates from the soil, water and building materials, becoming trapped in homes. Understanding how radon acts on the development of lung cancer is therefore a general population health concern.

In the present work, lung cancer mortality among the Wismut workers has been analysed using the TSCE model of carcinogenesis. We have considered 58695 males and a total of 2996 lung cancer deaths. The average exposure is about 280 WLM ($\sim$1.7 Sv for 5.9 mSv/WLM; see [11]) and considerably larger than typical residential levels, but extrapolations of the risk to the low dose regime have been found useful before; see [1].

I. Zaballa (✉) · M. Eidemüller
ISS, Helmholtz Zentrum München, Oberschleißheim, Germany
e-mail: ignacio.zaballa@helmholtz-muenchen.de

M. Eidemüller
e-mail: markus.eidemueller@helmholtz-muenchen.de

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_21

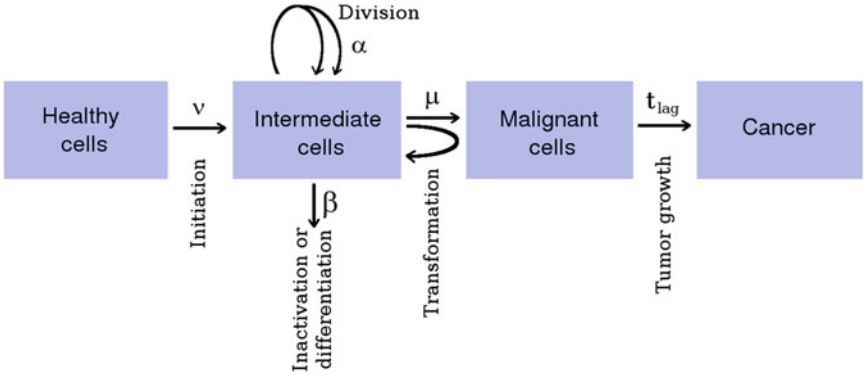


Fig. 1 Sketch of the processes leading into a malignant tumor in the TSCE model

2 The TSCE Model

In the TSCE model a healthy stem cell undergoes two irreversible steps and clonal expansion during the intermediate stage before it develops a tumor; see [5, 6]. Thus, the model distinguishes three distinctive processes on a cell's pathway to cancer: initiation with rate ν, clonal expansion with rate γ, and transformation with rate μ; see Fig. 1. The effective clonal expansion is given by $\gamma = \alpha - \beta - \mu$, where α and β represent the division and inactivation rate of initiated cells. To analyse epidemiological data the parameter estimates σ_j in the hazard function $h(t; \sigma_j)$ for a definite model are obtained by maximizing the total likelihood. The exposure response of radon on the different steps is carefully evaluated.

A varying clonal expansion γ with radon and silica exposure describes best the data in the TSCE model. The combined effect of both exposures is additive, $\gamma = \gamma_b(1 + f_r + f_s)$, with γ_b being the baseline rate. The radon response is a non-linear function of the exposure rate d_r,

$$f_r\,(d_r) = r_1\left(1 - \mathrm{e}^{-r_2 d_r / r_1}\right), \tag{1}$$

where $\gamma_b r_1$ gives the saturation level for large exposure rates, and $\gamma_b r_2$ the linear slope for small exposure rates. The resulting silica response is linear above the exposure rate $d_c \simeq 1\,\mathrm{mg/m^3} \cdot \mathrm{yr}$, which corresponds to an exposure of $0.02\,\mathrm{mg/m^3}$ during a 40 h working week. This is of the same order of magnitude as the NIOSH recommended limit for respirable silica of $0.05\,\mathrm{mg/m^3}$; see [7].

The model parameters resulting from this analysis, which are thoroughly examined in [13], are consistent with other occupational alpha-particle studies. The saturation of the clonal expansion may be related to the inverse exposure-rate effect; see Fig. 2. According to Brenner [2], this effect could be the consequence of wasted dose due to multiple track traversals.

3 Sources of Uncertainty in the Current Study

Exposure measurements of radon and its decay products were performed regularly after 1966 for the Wismut cohort. Before that date exposure rates were highest, and only retrospective and partial measurements were carried out. It is possible then that either systematic as well as random errors are present in the individual exposure estimates. These errors could have an effect on the risk estimation and in its temporal variation with age, exposure rate, and other time variables; see [1].

Wismut Miners were also exposed to other carcinogenic agents: external γ radiation, long-lived radionuclides, arsenic, fine dust and silica. Previous studies of this cohort have found that only exposure to silica dust confounds the lung cancer risk; see [9]. Silica dust exposure has been taken into account in our analysis for a more precise estimation of the risk by radon exposure.

Another important source of uncertainty is the lack of complete smoking information. Nevertheless, comparison to other studies of uranium miners with available smoking information suggests that the baseline values obtained in our analysis for the clonal expansion, $\gamma_b = 0.16\ \mathrm{yr}^{-1}$, already contain the smoking contribution; see [13]. These results agree very well with case control studies of European miners cohorts, which found that adjustment to smoking does not vary appreciably the radon exposure risk; see [4, 8]. The high percentage of workers that smoked has been proposed to explain this observation.

Below, we consider the model specification as a source of uncertainty in more detail. Other possible sources of uncertainty that we do not consider here are: errors assigning the underlying cause of death, uncertainty in the background exposures, and demographic factors.

4 Conclusions and Open Questions

We believe that analyses using mechanistic models have a number of advantages over the standard descriptive models more often used in epidemiological studies. Standard descriptive methods require several parameters to describe the complex temporal risk patterns. In the case of the Wismut cohort, at least 3 covariates modifying the radon exposure have been found; see [12]. The effect of these covariates depends on the precision of the exposure measurements, which are themselves subject to possible large errors. Contrary to the descriptive methods, effect modification is built-in in the dynamics of the TSCE model avoiding this pitfall. Besides the radon response is described by only two parameters in the model [13], in contrast with the at least four needed in descriptive methods. Having a smaller number of parameters describing the exposure response results in tighter error margins in the parameters and risk estimates.

As shown in Fig. 2, the ERR and TSCE models risk predictions differ more pronouncedly at young attained ages and low exposure rates. Away from the mean cancer

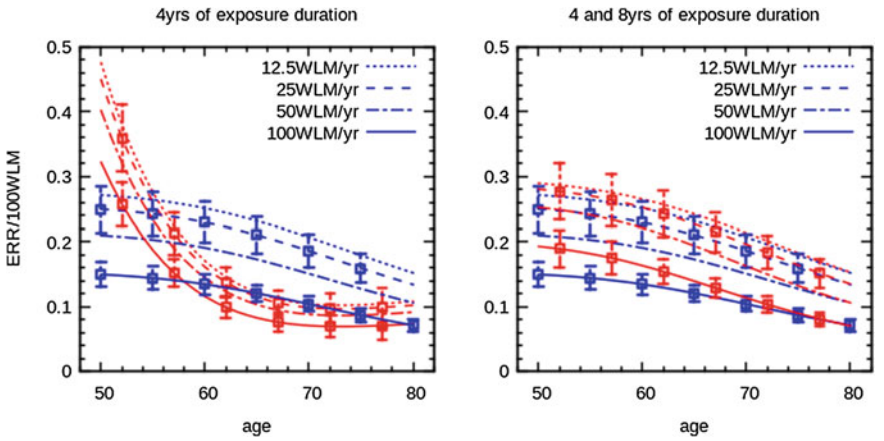


Fig. 2 Radon ERR versus attained age for different exposure scenarios. In the *left panel* the *red curves* correspond to the ERR model, while *blue* ones to the TSCE model. The *blue* and *red lines* in the *right panel* represent the 4 and 8 years duration for the TSCE model. The descriptive ERR model used here, which is similar to the one used in [12], can be found in [13]. Error bars are 1σ and are based on the uncertainty of the parameters and their correlations.

age, about 64 yrs, one is inclined to assume that age extrapolation in a model based on a dynamical process is more reliable than in a purely descriptive model; see [1]. The exposure rate is defined in the ERR model as an average over the total exposure duration; see [12]. In the TSCE model, annual exposure rates are incorporated in a natural way in the time dependent covariates. This difference may account, at least partially, for the larger discrepancy in the risk estimates in the low exposure regime, although the correlation between the effect modifiers in the ERR model and the precision of the radon exposure measurements could play a part too.

The TSCE model is one of the simplest mechanistic models describing the multistage nature of cancer development in terms of stochastic processes. Although it is not clear whether it may be capturing all the relevant aspects in tumor pathogenesis, it is apparent that it reproduces the inverse exposure rate effect and the risk decay with age observed in different uranium miners cohorts; see Fig. 2. More information on lung cancer sub-types as well as on likely values of typical mutation and clonal growth rates from biological measurements would allow to establish and develop more specific models.

Regardless of the model used in the analysis, the relevance of the individual exposure estimation errors in the accurate evaluation of the exposure response and risk estimation cannot be overlooked. In fact, subsets of miners cohorts corresponding to lower exposure rates and uncertainty levels in the exposure estimates have been analysed, and yield substantially larger radon risks; see [3, 10]. The price to pay in this approach is the loss of statistical power and therefore larger confidence level

intervals in the estimations. To have a more accurate exposure response on the whole range of exposure regimes, it may be necessary to have a consistent method to gauge more precisely the effect of error in the individual exposure estimates.

Acknowledgements Work supported by the European Commission under FP7 EpiRadBio no. 269553.

References

1. BEIR VI, "Health effects of exposure to radon. Committee on the Biological Effects of Ionizing Radiation, National Research Council", Washington, DC: National Academy Press (1999).
2. D.J. Brenner, "The significance of dose rate in assessing the hazards of domestic radon exposure", *Health Phys.* **67** (2015), 76–79.
3. M. Kreuzer *et al.*, "Lung cancer risk at low radon exposure rates in German uranium miners", *Br. J. Cancer* **113** (2015), 1367–1369.
4. K. Leuraud *et al.*, "Radon, smoking and lung cancer risk: results of a joint analysis of three European case-control studies among uranium miners", *Radiat. Res.* **176** (2011) 375–387.
5. S.H. Moolgavkar, "Model of human carcinogenesis: action of environmental agents", *Environ. Health Perspect.* **50** (1983), 285–291.
6. S.H. Moolgavkar and E.G. Luebeck, "Two-event model for carcinogenesis: biological, mathematical and statistical considerations", *Risk Anal.* **10** (1990), 323–341.
7. National Institute for Occupational Safety and Health (NIOSH), "Criteria for a recommended standard: occupational exposure to crystalline silica", *DHEW (NIOSH) Publication* (1974), 75–120.
8. M. Schnelzer *et al.*, "Accounting for smoking in the radon-related lung cancer risk among german uranium miners: results of a nested case-control study", *Health Phys.* **98** (2010), 20–28.
9. M. Sogl *et al.*, "Quantitative relationship between silica exposure and lung cancer mortality in German uranium miners, 1946–2003", *Br. J Cancer* **107** (2012), 1188–1194.
10. L. Tomasek *et al.*, "Lung cancer in French and Czech uranium workers: radon-associated risk at low exposure rates and modifying effects of time since exposure and age at exposure", *Radiat. Res.* **169** (2008), 125–137.
11. UNSCEAR 2000, "Sources and effects of ionizing radiation, Vol. II: effects", United Nations, New York, (2000).
12. L. Walsh *et al.*, "The influence of radon exposures on lung cancer mortality in German uranium miners, 1946–2003", *Radiat Res* **173** (2010), 79–90.
13. I. Zaballa and M. Eidemüller, "Mechanistic study on lung cancer mortality after radon exposure in the Wismut cohort supports important role of clonal expansion in lung carcinogenesis", *Radiat Environ Biophys.* **55**(3) (2016), 299–315.