VII JORNADAS

CIENTÍFICAS DE ESTUDIANTES DE LA SEB

7-9 FEBRERO 2024

Parc de Recerca Biomèdica de Barcelona (PRBB)

(shorturl.at/CETVZ) VII-JSEB

WWW.BIOMETRICSOCIETY.NET

Organizadores

Parque Investigación Biomédica Barcelona

SOCIEDAD ESPAÑOLA DE BIOESTADÍSTICA

Patrocinadores

BIOstatnet

GRBIO

ISGlobal Barcelona Institute for Global Health

Servei d Estadística
Universitat Autònoma de Barcelona

SoE

VII Jornadas Científicas de Estudiantes de la SEB

PRBB Barcelona Biomedical Research Park

C/ Doctor Aiguader, 88,
Barcelona, España
7, 8 y 9 de febrero de 2024
www.biometricsociety.net/2023/06/18/
vii-jornadas-cientificas-de-estudiantes-de-la-seb

Bienvenida

¡Bienvenidas y bienvenidos a todas y todos a la VII edición de las Jornadas Científicas de Estudiantes de la Sociedad Española de Bioestadística!

Tras haber celebrado la última edición en Valencia, ciudad en la que se organizaron por primera vez estas jornadas científicas, os comunicamos con grata ilusión que en la presente edición podremos encontrarnos nuevamente en Barcelona. Como ya sabéis, el objectivo principal de estas jornadas es promover e impulsar el trabajo que los estudiantes y jóvenes investigadores realizamos en el área de la bioestadística. Por esta razón, estas jornadas, de la misma forma que las ediciones anteriores, seguirán siendo un espacio seguro, donde las y los jóvenes estudiantes podremos compartir el trabajo que estamos realizando, sin temor a críticas muy duras o preguntas complejas. Se constituye, por tanto, en un espacio ideal para que podamos reconocer y compartir la pasión por la investigación y aquellos desarrollos en los que estamos trabajando.

Estas jornadas científicas, organizadas por y para los y las estudiantes de la Sociedad Española de Bioestadística (SEB), se han convertido en una de las actividades principales de la sociedad. No hubiera sido posible organizar ninguna de estas ediciones si no fuera por la ayuda y el apoyo de la SEB, que demuestra año tras año y edición tras edición, que sigue creyendo en el potencial de sus jóvenes estudiantes. Este año no ha sido una excepción, y desde el comité organizador, queremos agradecerle a toda la sociedad el respaldo que nos ha dado, por apoyarnos en todas y cada una de las decisiones que hemos tenido que adoptar y también por la financiación económica.

Queremos agradecer también el apoyo que hemos recibido por parte del Parc de Recerca Biomèdica Barcelona (PRBB), quien nos ha abierto sus puertas y nos ha ofrecido todas las facilidades. Junto con el apoyo institucional y económico facilitado por BIOSTATNET, ISGlobal, GRBIO, SEA y SoCE. Por otra parte, queremos agradecer también a Jordi Cortés Martínez por impartir el curso *"Review of machine learning models for survival analysis"*, y a las y los ponentes de la mesa redonda Marta Bofill, Daniel Fernández, Maria Grazia Pennino y Guillermo Villacampa por aceptar nuestra invitación y participar en estas jornadas en las que tratarán *"El papel del profesional de la estadística en diferentes partes del mundo"*.

Para acabar, queremos agradecer a todos los participantes, estudiantes de la SEB, por mostrar interés y confianza en esta séptima edición de las jornadas. ¡Gracias a todas y todos! ¡Esperamos que entre todas y todos, disfrutemos y hagamos que estas jornadas sean especiales!

Comité Científico y Organizador

Comité científico/organizador

- Alba Fuster Alonso (ICM-CSIC)
- Andrea Toloba López-Egea (UPC)
- Armand González Escalante (BBRC)
- Blanca Rodríguez Fernández (BBRC)
- Garazi Retegui Goñi (UPNA, INAMAT2)
- Leire Garmendia Bergés (BCAM)
- Mario Figueira Pereira (UV)
- Patricia Genius Serra (BBRC)
- Pavel Hernández Amaro (UC3M)
- Sofía Aguilar Lacasaña (ISGlobal)

Contenidos

Programa

MIÉRCOLES 7		JUEVES 8		VIERNES 9	
9:00h 9:30h	Recepción	9:00h 10:15h	**Sesión 3: Genética e inferencia causal**		
9:30h 10:10h	Inauguración				
10:10h 10:30h	Presentación Sociedad Catalana de Estadística (SoCE)	10:15h 10:45h	Descanso		
10:30h 12:30h	**Curso: *"Review of machine learning models for survival analyses"* Impartido por Jordi Cortés (UPC)**	10:45h 12:30h	**Sesión 4: BIOSTATNET**	10:30h 12:00h	**Sesión Pósters + Desayuno**
12:30h 14:30h	Comida	12:30h 14:30h	Comida	12.00h 13:15h	**Mesa redonda: *"El papel del profesional de la estadística en diferentes partes del mundo"***
				13:15h 14:15h	Charla + Clausura
14:30h 15:45h	**Sesión 1: Clustering**	14:30h 15:30h	**Sesión 5: Modeling**		
15:45h 16:15h	Coffee break	15:30h 16:30h	**Sesión 6: Survival**		
16:15h 17:15h	**Sesión 2: Medical Studies**	16:30h 17:00h	Coffee break + merienda		
		17:00h 17:45h	**Sesión 7: Machine Learning y Software**		
17:45h	Actividad social				
		21:00h	Cena		

Miércoles, 7 de febrero

9:00-9:30 Recepción

9:30-10:10 Inauguración

10:10-10:30 Presentación Sociedad Catalana de Estadística (SoCE)

10:30-12:30 Curso: *"Review of machine learning models for survival analysis"* Jordi Cortés (UPC)

12:30-14:30 Comida

SESIÓN 1: CLUSTERING

Chair: Patricia Genius Serra

14:30 "The Cortical Asymmetry Index (CAI) for subtyping frontotemporal dementia and Alzheimer's disease patients", *Agnès Pérez-Millan.*

14:45 "COVID-19 reinfected patient profiles: A clustering approach", *Lander Rodríguez Idiazabal.*

15:00 "Flexible calibration curves in multicenter studies with binary outcomes", *Lasai Barreñada.*

15:15 "The choice of OTUs vs. ASVs on Antartic samples of air microbial communities", *Lucía Yubero Fernández.*

15:30 "Herramientas metodológicas para benchmarking en atención sanitaria basada en el valor: arquetipos y clasificación de pacientes diagnosticadas con cáncer de mama", *Maialen Otamendi Garitano.*

15:45 - 16:15 Coffee break

SESIÓN 2: MEDICAL STUDIES

Chair: Blanca Rodríguez Fernández

16:15 "Influencia del médico prescriptor en el tiempo para recibir tratamiento antiviral frente a la Covid-19", *Cristóbal Manuel Rodríguez Leal.*

16:30 "Stress Testing the CL concept: Evaluating Centiloid Stability to Tracer, Effective Image Resolution and Quantification Method", *Mahnaz Shekari.*

16:45 "Wave and ceiling of care impact on COVID-19 in-hospital mortality: An inverse probability weighting analysis", *Natàlia Pallarès Fontanet.*

17:00 "Análisis de la Exposición a Fármacos antidiabéticos Orales en la Población Metropolitana de Barcelona. Estudio sobre el análisis estadístico con sobrerepresentación de "Ceros" ", *Pablo Castillo Jiménez.*

17:45 Actividad Social

Jueves, 8 de febrero

SESIÓN 3: GENÉTICA E INFERENCIA CAUSAL

Chair: Armand González Escalante

9:00 "Analysis of Spatial Gene Expression Data: A Case Study in Neuroscience", *Carlos Javier Peña de los Santos.*

9:15 "Unraveling the molecular mechanisms associated with polycystic ovary syndrome using a multi-omics analysis strategy", *Edmond Géraud Aguilar.*

9:30 "Feature Selection in cell-free RNA-Seq data for the detection of biomarkers with predictive value in pathology", *Esther Tercero.*

9:45 "El impacto del tabaco provoca cambios epigenéticos en el tejido adiposo subcutáneo: Implicaciones en la Progresión de la Enfermedad de Crohn", *Irene Vaño Segarra.*

10:00 "Meta-Analysis of Epigenome wide association study of DNA Methylation and Ultra-Processed Food Consumption in middle-childhood", *Joana Llauradó Pont.*

10:15 - 10:45 Descanso

SESIÓN 4: BIOSTATNET

Chair: David Moriña

10:45 Presentación Biostatnet

11:00 "Short-time cancer incidence projections addressing missing data challenges", *Garazi Retegui Goñi.*

11:15 "The individual causal association for the evaluation of surrogate endpoints based on causal Inference under non-normality", *Gokce Deliorman.*

11:30 "Double data fusion and calibration modeling for zooplankton abundance measured in space and time", *Jorge Castillo-Mateo.*

11:45 "Time-dependent AUC for survival and competing risks models", *Leire Garmendia Bergés.*

12:00 "Survival analysis in genetics", *María del Pilar González Zamorano.*

12:15 "A comparison between the Bayesian MCMC software packages WinBUGS and NIMBLE for modeling spatial ordinal survey-based data", *Miguel Ángel Beltrán Sánchez.*

12:30-14:30 Comida

SESIÓN 5: MODELING

Chair: Alba Fuster Alonso

14:30 "Spatial point processes: from the mathematical basis to its applications", *Arnau Garcia Fernandez.*

14:45 "A survival analysis and multilevel modeling study to approach gender bias in Time Compactation", *Gonzalo Aparicio Rodríguez*

15:00 "Spatial Bayesian Distributed lag non-linear models: a case study of small-area temperature-mortality association", *Marcos Quijal Zamorano.*

15:15 "Statistical approaches to correct for baselines in clinical trials", *Matilde Francisco.*

SESIÓN 6: SURVIVAL

Chair: Andrea Toloba López-Egea

15:30 "MLE-based approach for inference of the clustered-state Markovian arrival process for recurrence-death data in patients with oncological diseases", *Álvaro Díaz Pérez.*

15:45 "Evaluation of discrimination ability of time-dependent variables in a Cox proportional hazards model", *Antia Enriquez Yurrebaso*

16:00 "Identificación de factores de riesgo para la supervivencia de pacientes con cáncer colorrectal mediante un análisis de riesgos competitivos", *María Gascón Pérez.*

16:15 "A joint model for (un)bounded longitudinal markers, competing risks, and recurrent events using patient registry data", *Pedro Miranda Afonso.*

16:30-17:00 Coffee break

SESIÓN 7: MACHINE LEARNING & SOFTWARE

Chair: Pavel Hernández Amaro

17:00 "Efficient and reproducible table summaries with *gtsummary* in R", *João Pedro Carmezim Correia.*

17:15 "Detección de fraude alimentario en leche: Análisis de especiación de leche y detección de leche de cabra adulterada con leches de menor calidad, empleando aprendizaje automático e implementación en aplicación web", *Miguel Ángel López González.*

17:30 "R-shiny application of the evolution of COVID-19 in Catalonia with Bayesian spatio-temporal analysis", *Pau Satorra Herbera.*

21:00 Cena

Viernes, 9 de febrero

SESIÓN DE PÓSTERES

(10:30-12:00)

1. "Dealing with complex sampling designs on the estimation of the ROC curve and the area under it", *Amaia Iparragirre Letamendi.*
2. "El gran impacto de INLA y SPDE para la modelización estadística espacial desde la perspectiva bayesiana", *Carmen Guarner Giner.*
3. "Analysis of Big Data with the integrated nested Laplace approximation", *Héctor López Gómez.*
4. "Block-wise missing omics data integration using kernel methods", *Ignacio Pérez Blasco.*
5. "Compare variable selection methods using Random Forest and Boruta to have a tool for predicting multi-resistance", *Janire Gallejones Eskubi.*
6. "Assessing the cause of death in patients with heart failure through a Bayesian competing risks survival model", *Jesús Gutiérrez Botella.*
7. "The COVID-19 Pandemic's impact on clinic indicators of health for population in chronic conditions", *Manuel Alejandro Moreno Vasquez.*
8. "Global species distribution models for penguin species", *Marc Moreno Candón.*
9. "Statistical modeling to adjust for time trends in platform trials utilising non-concurrent controls", *Pavla Krotka.*
10. "Análisis estadístico de la supervivencia de pacientes con implante Cardioband", *Sandra Gonzalez Martin.*

12:00-13:15 Mesa redonda: "El papel del profesional de la estadística en diferentes partes del mundo"

13:15-14:15 Charla y Clausura

Curso

Title: Review of machine learning models for survival analysis

Instructor: Jordi Cortés Martínez, Universitat Politècnica de Catalunya (UPC)

Abstract: This course presents an in-depth, succinct review of machine learning techniques utilized in survival analysis, juxtaposing traditional statistical methods with cutting-edge algorithms. It encompasses a thorough exploration of the fundamentals of prevalent machine learning models, including Random Survival Forests, Support Vector Machines, and deep learning approaches, with a specific focus on their application using R. The practical segment of the course offers hands-on implementation of selected methods, underscoring the evaluation and comparison of these models against the Cox regression benchmark. Ultimately, the course is designed to endow participants with the requisite skills to comprehend and leverage machine learning in the analysis and prediction of time-to-event outcomes across a multitude of sectors.

Mesa Redonda

Título: El papel del profesional de la estadística en diferentes partes del mundo

Resumen: Si tuviéramos que señalar una disciplina transversal que esté presente en muchísimos aspectos de nuestro día a día esta es, sin duda, la estadística. La sociedad es cada vez más consciente del papel fundamental que tienen los profesionales de la estadística, aunque todavía queda un largo tramo por recorrer. Pero ¿cuál es el papel del estadístico y, además, se valora de la misma manera en todos los países?

Para abordar estas preguntas, nuestra mesa redonda de hoy contará con la participación de 4 panelistas con un gran recorrido internacional. Compartirán sus experiencias como profesionales de la estadística en los diferentes países en los que han trabajado: Marta Bofill (Medical University of Vienna), Daniel Fernández (UPC), Maria Grazia Pennino (Instituto Español de Oceanografía - CSIC) y Guillermo Villacampa (Instituto de Investigación de Vall d´Hebron).

Sus perfiles profesionales son muy distintos, pero tienen algo en común y es la estadística. Gracias a sus conocimientos podremos profundizar sobre las diferencias que hay entre países en temas como: el rol del estadístico y sus funciones en la empresa y la academia; implicación del estadístico en los proyectos de investigación; y avances en la inclusión y comprensión de la estadística en la comunidad. Estas son algunas de las cuestiones a debatir entre los panelistas, y también contaremos con rondas de preguntas por parte del público, para que no te quedes con ninguna duda.

Marta Bofill: Medical University of Vienna. Marta Bofill Roig es investigadora postdoctoral en el Center for Medical Data Science de la Medical University of Vienna. Su investigación se centra en métodos y software para el diseño y análisis de ensayos de clínicos complejos. Marta se licenció en Matemáticas (Universitat de Barcelona) e hizo un máster en Estadística e Investigación Operativa (Universitat de Barcelona-Universitat Politècnica de Catalunya). Tiene un doctorado en Estadística e Investigación Operativa por la Universitat Politècnica de Catalunya, durante el cual trabajó en el diseño de ensayos clínicos con múltiples variables correlacionadas.

Daniel Fernández: Universitat Politècnica de Catalunya (UPC). Daniel Fernández es profesor lector Serra-Húnter en el Departamento de Estadística e Investigación Operativa de la Universitat Politècnica de Catalunya. Tiene un máster en Probabilidad y Estadística del Centro de Investigaciones Matemáticas, en México. Su doctorado y, posterior investigación postdoctoral, se dieron a cabo en la Victoria University of Wellington, en Nueva Zelanda. En 2016, hizo una postdoc en el Center for Data Science de la New York University en temas de estadística computacional (EEUU) y en 2017 trabajo como profesor en el Departamento de Epidemiología y Bioestadística la State University of New York (SUNY) en Albany (EEUU). Su investigación se centra en desarrollar métodos de clasificación y clustering enfocados en respuestas ordinales.

Maria Grazia Pennino: Instituto Español de Oceanografía. Maria Grazia Pennino es bióloga marina, con una maestría en Bioestadística y un doctorado en Matemáticas y Estadística. Actualmente trabajando en el C.O. Madrid, Instituto Español de Oceanografía (IEO-CSIC) en el Departamento de Pesca. Sus principales campos de investigación son la modelización espacial-temporal y la bioestadística en general para asesorar una gestión pesquera eficaz. Ha estudiado diferentes pesquerías (industrial y de pequeña escala / artesanal) y varios ecosistemas (Mediterráneo, Océano Índico y áreas costeras brasileñas), trabajando en diferentes escalas espaciales (de locales a globales) y temporales. Recientemente, también se ha interesado en vincular factores sociales y económicos en el marco de los modelos de distribución de especies para comprender cómo podrían afectar a la distribución de las especies.

Guillermo Villacampa: Instituto de Investigación de Vall d´Hebron. Guillermo Villacampa trabaja como estadístico en el Instituto de Investigación Vall d'Hebron en Barcelona y en el grupo SOLTI un grupo cooperativo dedicado a la investigación en cáncer de mama. Durante 2021-2023, Guillermo residió en Londres colaborando con el Institute of Cancer Research del Reino Unido. En el ámbito académico, Guillermo es graduado en estadísitca, licenciado en sociología y cursó el máster de estadística e investigación operativa de la Universitat Politècnica de Catalunya.

Sesiones

Sesión 1: Clustering

7 de febrero, 14:30 a 15:45

Chair: Patricia Genius Serra

The Cortical Asymmetry Index (CAI) for subtyping frontotemporal dementia and Alzheimer's disease patients

Agnès Pérez-Millan[1,2,3], Uma Maria Lal-Trehan Estrada[2,3], Neus Falgàs[1], Núria Guillén[1] Sergi Borrego-Écija[1], Beatriz Bosch[1], Jordi Juncà-Parella[1], Adrià Tort-Merino[1], Jordi Sarto[1], Josep Maria Augé[4], Anna Antonell[1], Nuria Bargalló[5], Raquel Ruiz-García[6], Laura Naranjo[6], Mircea Balasa[1], Albert Lladó[1,2], Roser Sala-Llonch[2,3,7], Raquel Sánchez-Valle[1,2,]

[1]Alzheimer's disease and other cognitive disorders Group. Service of Neurology, Hospital Clínic de Barcelona. Fundació Recerca Clínic Barcelona-IDIBAPS; [2]Institut de Neurociències, University of Barcelona; [3]Department of Biomedicine, University of Barcelona, Barcelona, 08036, Spain; [4]Biochemistry and Molecular Genetics Department, Hospital Clínic de Barcelona, Barcelona, Spain; [5]Image Diagnostic Centre, Hospital Clínic de Barcelona, Barcelona, Spain. CIBER de Salud Mental, Instituto de Salud Carlos III. Magnetic Resonance Image Core Facility, IDIBAPS; [6]Immunology Service, Biomedical Diagnostic Center, Hospital Clínic de Barcelona, Barcelona, Spain; [7]Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Barcelona, Spain.

Frontotemporal dementia (FTD) patients usually show more asymmetric brain atrophy patterns than Alzheimer's Disease (AD). Here, we define the individual Cortical Asymmetry Index (CAI) and explore its diagnostic utility. We collected structural T1-MRI scans from 554 participants from the Alzheimer's disease and other cognitive disorders unit of the Hospital Clínic de Barcelona (Spain), including FTD, AD, and healthy controls, and processed them using Freesurfer. We defined the CAI using summary region-derived cortical thickness measures and a metric derived from information theory. A subset of the study participants had cerebrospinal fluid (CSF), plasma measures, or additional follow-up MRIs. We analyzed differences in CAI at cross-sectional and longitudinal levels. We then clustered FTD and AD subjects based on the CAI values and studied the fluid biomarkers characteristics within each cluster. CAI differentiated FTD, AD, and healthy controls. It also distinguished the semantic variant Primary Progressive Aphasia (svPPA) from the other FTD phenotypes. In FTD, the CAI increased over time. The cluster analysis identified two subgroups within FTD, characterized by different CSF and plasma neurofilament-light (NfL) levels, and two subgroups within AD, with different plasma Glial fibrillary acidic protein (GFAP) levels. In AD, CAI correlated with plasma-GFAP and Mini-Mental State Examination (MMSE); FTD was associated with NfL levels (CSF and plasma). The method proposed here for the CAI at the individual level could quantify asymmetries previously described visually. The CAI could define clinically and biologically meaningful disease subgroups.

Keywords: Cortical asymmetry index, Alzheimer's Disease, frontotemporal dementia

COVID-19 reinfected patient profiles: A clustering approach

Lander Rodriguez[1], Irantzu Barrio[2,1], José M. Quintana-López[3]
[1]Applied Statistics, Basque Center for Applied Mathematics; [2]Department of Mathematics, University of the Basque Country; [3] Research Unit of the Galdakao-Usansolo University Hospital, Osakidetza Basque Health Service

The appearance of even more transmissible SARS-CoV-2 variants resulted in secondary or even multiple reinfections, increasing the pressure on health systems and bringing worldwide concern. While reinfections were thoroughly studied, the identification of the worst prognosis reinfected individuals was not investigated in detail. In a pandemic context with a high circulation of the virus, identifying individuals at a greater risk is necessary so that preventive strategies can be specified for their protection. This can be accomplished with clustering techniques, which are able to find the hidden and inherent patterns of electronic health records.

A total of 380,074 adult patients were infected with SARS-CoV-2 from March 1, 2020 to January 9, 2022 in the Basque Country; from those, 10,968 (2.89%) were reinfected in the same period. In this study we implement a three-stage process. First, we identify the COVID-19 reinfected profiles from the Basque Country; second, we assess their association with the adverse outcomes of the disease; and third, we evaluate the clusters' and vaccination association with the time between infections. To identify the profiles of reinfected patients, we apply the novel KAMILA (KAy-means for MIxed LArge data) clustering algorithm [1], which is well-suited to cluster mixed-type data. For the second and third goals, we consider logistic and linear regression models, respectively.

Four patient profiles were identified: the *Very low* cluster was composed of young patients with no comorbidities; the *Low* group was mainly formed of middle-aged individuals with no comorbidities; the *High* profile was composed of older adults with few comorbidities; and the *Very high* cluster included very old patients with various comorbidities. Age distinguished the clusters the most although there were also important differences in the proportion of patients with diabetes, heart failure, arterial hypertension and dyslipidemia. In addition, the proportion of adverse outcomes of COVID-19 increased with the risk level of the clusters. Finally, differences were found in the time between infections among the risk clusters.

This study identified COVID-19 reinfected patient profiles in an unsupervised manner, which were correctly segregated according to their risk for the adverse outcomes of the disease. This indicates that this technique could be useful for the quick identification of higher-risk patients, leading to multiple care intervention strategies. This would improve the medical attention offered to the most vulnerable individuals.

Keywords: COVID-19, reinfection, clustering

References

1. Foss A., Markatou M., Ray B., and Heching A. (2016). A semiparametric method for clustering mixed data. *Machine Learning, 83(16)*, 1-44.

Flexible calibration curves in multicenter studies with binary outcomes

Lasai Barreñada[1], Laure Wynants[1,2], Ben Van Calster[1,3,4]

[1]Department of development and regeneration, KU Leuven; [2]Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University; [3]Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven; [4]Department of Biomedical Data Sciences, Leiden University Medical Centre

Calibration of a prediction model refers to the agreement between observed outcomes and predictions. Calibration plots are a key tool to measure moderate calibration of individual predicted risks. This plot has predictions on the x-axis, and the outcome on the y-axis. When the outcome is binary, observed proportions of the outcome must be calculated. This is usually done fitting a model with the outcome as independent variable and the linear predictor as the only predictor. For obtaining smooth curves LOESS or restricted splines are used. Most of clinical datasets contain clustering (e.g. centers in a multicenter study) which entails challenges for the accurate estimation of observed proportions.

Current procedures do not account for clustering in the sample. In this study we devise three methodological approaches for estimating observed proportions accounting for clustering. First, we present a simple extension to Hosmer-Lemeshow plot where observed proportions are obtained by quantile group in each center and then summarised for the complete dataset. Second, we develop a two stage approach based on random effects meta-analysis. In this methodology we first calculate the observed proportions per center and then using random effects meta-analysis we pool the performance of the estimated observed proportions per center obtaining a smooth curve with confidence and prediction intervals. Third, we implement a one stage approach using an appropriate mixed model that accounts for clustering of participants. This model includes a random intercept and a random slope per center and allows us to calculate confidence interval around the observed proportions. We illustrate the three methodologies using a logistic regression model for ovarian tumour diagnosis in an dataset of 8403 patients and 32 centers. We also make available an R function that calculates the observed proportions based on any of the three presented methodologies and generates the multicenter calibration plot.

Keywords: Calibration, prediction modelling, clustering.

References

1. Riley RD, Ensor J, Hattle M, et al. Two-stage or not two-stage? That is the question for IPD meta-analysis projects. Research Synthesis Methods. 2023;14:903–10.

2. Wynants L, Vergouwe Y, Van Huffel S, et al. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. Stat Methods Med Res. 2018;27:1723–36.

The choice of OTUs vs. ASVs on Antartic samples of air microbial communities

Lucía Yubero Fernández[1], Sofía Galbán[2] and Ana María Justel Eusebio[3]
[1]Student of Master's Degree in Bioinformatics and Computational Biology, UAM; [2]Biology Department, UAM; [3]Department of Mathematics, UAM

High Throughput Sequencing (HTS)-based technology enables identifying the genome of microbial organisms present on an environmental sample. These techniques demand a quality control since errors, contamination and noise may be introduced in the PCR and sequencing process. Here three clustering algorithms used for the above purpose will be analysed and compared: CROP, DADA2 and UPARSE. They infer a (sub)set of true biological sequences named OTUs and ASVs, respectively. By employing diversity indexes and various statistical tools to analyze diversity in metagenomics, the differences in results for the three procedures will be assessed for Antartic Plateau data collected by MICROAIRPOLAR research group.

Keywords: Errors correction, OTU, ASV

References

1. Callahan, B.J., McCurdie and Holmes, S.P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal 11*, 263–264.

2. Callahan, B.J., McCurdie, P. J., Rosen, M.J., Han, A.W., Johnson, A.J. and Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods 13*, 581–583

3. Chen, W., Zhang, C.K., Cheng, Y., Zhang, S. and Zhao, H. (2013). A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. *PLoS One*, vol.8

4. Chiarello, M, McCauley, M., Villéger, S. and Jackson, C. (2022). Ranking the biases: The choice of OTUs vs. AVSs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. *Austral Ecology, 26*, 32-46

5. Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods 10*, 996–998

6. Hao, X., Jiang, R. and Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics*, vol. 27 no. 5, 611–618

7. Rosen, M.J., Callahan. B.J., Fisher, D.S. and Holmes, S.P. (2012). Denoising PCR-amplified metagenome data. *BMC Bioinformatics*, 13:283

Herramientas metodológicas para benchmarking en atención sanitaria basada en el valor: arquetipos y clasificación de pacientes diagnosticadas con cáncer de mama

Otamendi M.[1], Gorostiza A.[1], Alayo I.[1], Fullaondo A. [1] y García-Lorenzo B. [1]
[1]Biosistemak Institute for Health Systems Research, Torre del Bilbao Exhibition Centre, Barakaldo, Spain

El cáncer de mama es el cáncer más común y de mayor mortalidad entre las mujeres. A pesar de los avances en investigación sanitaria, sigue teniendo un gran impacto clínico y económico sobre sus pacientes y el sistema de salud. En este contexto, aparece como nuevo paradigma en los servicios sanitarios la Atención Sanitaria Basada en el Valor (ASBV) poniendo al paciente en el centro y vinculando los resultados en salud a los costes. La ASBV propone observar no solo los resultados tradicionales clínicos y de procesos (*Clinical-Related Outcomes*, CROs y *Care-Process Related Outcomes, CAPROs*), sino también los resultados reportados directamente por el paciente (*Patient-Reported Outcomes, PROs*), así como sus usos de recursos sanitarios y los costes asociados a su proceso asistencial, para determinar el valor de la atención sanitaria.

No existe en la literatura publicada una identificación de arquetipos de pacientes diagnosticadas con cáncer de mama, ni una clasificación basada en su trayectoria asistencial que pueda ser relacionada con los PROs. Esta identificación es fundamental para llevar a cabo un benchmarking riguroso entre pacientes y centros desde la perspectiva de la ASBV. En la comunidad VOICE, -un consorcio europeo de 8 hospitales que centra su investigación en ASBV en cáncer de mama y pulmón-, se identificaron arquetipos de pacientes diagnosticadas con cáncer de mama mediante el *Clustering* Jerárquico en Componentes Principales (CJCP). Sin embargo, por un lado, este enfoque presenta algunas debilidades metodológicas y por otro, el equipo clínico de referencia de la comunidad VOICE planteó la necesidad de contar también con una herramienta automatizada que permitiera clasificar nuevas pacientes. Por lo tanto, este trabajo tiene un doble objetivo: i) superar las limitaciones del CJCP explorando alternativas para la identificación de arquetipos basada en su trayectoria asistencial, y ii) diseñar un clasificador de nuevas pacientes en los arquetipos anteriormente definidos. Se realizó una revisión de la literatura sobre metodologías para la identificación de arquetipos de pacientes y su clasificación. El *Clustering* tradicional y el Análisis de Clases Latentes (ACL) fueron los métodos más utilizados para la identificación de arquetipos. El ACL, los Bosques Aleatorios (BA) y la Regresión Multinomial (RM), fueron los métodos mayoritariamente aplicados para la clasificación. En este estudio para la identificación de arquetipos de pacientes y su clasificación se utilizó el ACL. La identificación de arquetipos se validó con CJCP, mientras que la clasificación se validó con los BA y la RM. El uso del ACL como única metodología para alcanzar ambos objetivos ofrece evidencia científica para su uso como herramienta eficiente y robusta para el benchmarking en el ámbito de la ASBV.

Este estudio no solo proporciona mejoras metodológicas en la identificación de arquetipos de pacientes diagnosticadas con cáncer de mama en el contexto de la ASBV, sino que además, es el primer estudio que aborda la clasificación de las pacientes en el mismo ámbito como herramienta para benchmarking.

Keywords: Atención Sanitaria Basada en el Valor; Clasificación; Análisis de Clases Latentes

Sesión 2: Medical studies

7 de febrero, 16:15 a 17:15

Chair: Blanca Rodríguez Fernández

INFLUENCIA DEL MÉDICO PRESCRIPTOR EN EL TIEMPO PARA RECIBIR TRATAMIENTO ANTIVIRAL FRENTE A LA COVID-19

Rodríguez-Leal CM[1], Susi-García R[2], Amador-Pacheco J[2], Pérez-Pérez T[2]
[1]Servicio de Urgencias, Hospital del Henares, Coslada (Madrid); [2]Departamento de Estadística y Ciencia de los Datos, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid

Introducción. La COVID-19 es una enfermedad producida por el virus SARS-CoV-2. El espectro de la enfermedad es amplio, pudiéndose producir cuadros muy graves en pacientes vulnerables. Se han desarrollado fármacos antivirales que han demostrado reducir la morbimortalidad en pacientes vulnerables cuando se administran precozmente. Su efectividad es menor cuanto más tarde se administran. **Objetivo**. Determinar si el tiempo entre la llegada de un paciente con COVID-19 a un Servicio de Urgencias Hospitalario (SUH) y la prescripción de un fármaco antiviral (tiempo de latencia) está influido por la especialidad del médico prescriptor. **Material y métodos**. Se utilizarán los datos provenientes de la cohorte COVID-CODE- SPAIN correspondiente a un análisis intermedio basado en los datos recogidos hasta mayo de 2023. Dicha cohorte está integrada por los datos de pacientes vulnerables mayores de 18 años atendidos en 16 SUH españoles con COVID-19 leve-moderado en los 7 primeros días de la enfermedad. Para explorar la asociación entre el tiempo de latencia en días y la especialidad del médico prescriptor (urgenciólogo vs no urgenciólogo) se utilizarán técnicas paramétricas, modelo ANOVA o tasa de fallo acelerado; como no paramétricas, test de Mann Whitney. Dado que los participantes no han sido aleatorizados en uno u otro grupo, es necesario controlar por potenciales factores de confusión(1) como son: inmunosupresión, edad, sexo, índice de comorbilidad de Charlson y hospitalización en los últimos 6 meses, aplicando técnicas de inferencia causal(2). Se compararán los resultados obtenidos utilizando distintos métodos de aplicación del índice de propensión, ponderación por el inverso de la varianza, un modelo multivariable, g-fórmula (estandarización) y modelo estructural anidado (g-estimación)(3).

Keywords: Inferencia causal. Infecciones por Coronavirus. Tratamiento Farmacológico de COVID-19.

References

1. Pearl J. Causal Diagrams for Empirical Research. Biometrika [Internet]. 1995;82(4):669-88. Disponible en: http://www.jstor.org/stable/2337329

2. Urdinez F, Cruz A. R for Political Data Science. R for Political Data Science. Chapman and Hall/CRC; 2020.

3. Hernán MA, Robins JM. Causal Inference. What if. Boca Raton: Chapman and Hall/CRC; 2020.

Stress Testing the CL concept: Evaluating Centiloid Stability to Tracer, Effective Image Resolution and Quantification Method

Mahnaz Shekari[1,2,3], David Vállez García, Lyduine E. Collij, Daniele Altomare, Fiona Heeman, Hugh Pemberton, Núria Roé Vellvé, Santiago Bullich, Christopher Buckley, Andrew Stephens, Gill Farrar, Giovanni Frisoni, William E. Klunk, Frederik Barkhof, Juan Domingo Gispert[1,2,3], On behalf of ADNI and the AMYPAD consortium

[1]Barcelonaβeta Brain Research Center (BBRC), FPM, Barcelona, Spain; [2]IMIM, Hospital del Mar Research Institute, Barcelona, Spain; [3]Universitat Pompeu Fabra, Barcelona, Spain.

The Centiloid (CL) scale is a well-established standard metric of amyloid load and several cut-off values have been proposed for research and clinical use. The aim of this study was to evaluate the stability of CL values to technical factors such as pipeline design options, tracer, and effective image resolution.

A total of 533 participants from AMYPAD DPMS and ADNI were included who were cognitively unimpaired (CU) or had subjective memory complaints (SMC), mild cognitive impairment (MCI), or dementia. Using SPM12, 32 pipelines were created, calibrated, and validated, based on combinations of four reference regions (whole cerebellum[WCB], cerebellar gray[CG], whole cerebellum+brainstem[WCB+BSTM], pons), two target volume-of-interest (VOI) types (standard GAAIN vs subject-based), two reference region types (standard GAAIN vs subject-based) and two analysis spaces (native vs montreal neurological institute [MNI]). Generalized Estimating Equations (GEE) were used to evaluate the impact of the different factors. All analyses were stratified into amyloid positives (A+) and negatives (A-) using a threshold of 24 CL. First, a base model was defined, including tracer and diagnosis as between-subject factors and the pipeline design factors as within-subject ones. Then, secondary GEE models were used to additionally evaluate the impact of grey matter atrophy and harmonization status. For all comparisons, changes>3 CL were considered relevant and p-value<0.05 was deemed statistically significant.

The tracer effect was not statistically significant for either the A- or A+ groups; the clinical diagnosis was only significant for the A+ group. Reference region (RR) selection and RR type presented the highest impact on CL with pons rendering the lowest CL compared to WCB for the A- group. The A+ group showed similar behavior with smaller effects. When using WCB as RR, comparable CL values were obtained for all tracers across both A- and A+ groups. Subject-based RR resulted in lower CL values than GAAIN RR VOIs. Using subject-based cortical target CL values were slightly higher. Quantification space slightly impacted CL for both groups. The subject-based cortical target was insensitive to the degree of gray matter atrophy, unlike the GAAIN VOI. CL were minimally affected by harmonization when using WCB and WCB+BSTM as reference regions, but CG and pons were sensitive to harmonization.

Using the WCB as RR yields comparable CL values across different tracers which are robust against differences in image resolution. Subject-based cortical VOI provides more accurate measurements of amyloid load in the presence of atrophy.

Keywords: Centiloid, Amyloid PET, Quantification, Alzheimer disease

Wave and ceiling of care impact on COVID-19 in-hospital mortality: An inverse probability weighting analysis

Pallarès N[1], Videla S[2], Carratalà J[3], Tebé C[1]
[1]Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Badalona, Spain;
[2]Department of Clinical Pharmacology, Bellvitge University Hospital, Barcelona, Spain;
[3]Department of Infectious Diseases, Bellvitge University Hospital, Barcelona, Spain;

Background and objective: Since March 2020, 6 waves of the COVID-19 pandemic have been registered in Spain. There are several studies comparing different COVID-19 waves but, as far as we know, none of them uses a matching procedure to make patients comparable or accounts for ceiling of care. Our aim is to analyze whether there are differences between waves regarding in-hospital mortality and ceiling of care.

Methods: Data come from an observational study conducted during four waves of COVID-19 (March 2020-August 2021) in 5 centers in Catalonia [1]. Three methods were used to study the effect of wave in in-hospital mortality in patients with and without ceiling of care: 1) a raw logistic model with only wave as a covariate; 2) a multivariant logistic model with wave and patient baseline information as covariates and 3) a logistic model with weights obtained from an inverse probability of weighting procedure to account for differences in baseline profile between waves. Coefficients and robust standard errors were adjusted for the variability between imputations according to Rubin's rules. All analyses were conducted using R software version 4.3.0.

Results: A total of 3982 patients with ceiling of care and 1831 patients without ceiling of care were analyzed. Patients with a ceiling of care were, in median, 20 years older than patients without a ceiling of care and in-hospital mortality ranged from 5% to 45%. For patients without ceiling of care, adjusted and IPTW models showed that 2nd, 3rd and 4th wave patients were less likely to die in hospital than 1st wave patients (odds ratio [OR] ranged from 0.5 to 0.3). For patients with ceiling of care, no differences in mortality were found between patients in waves 1, 2 and 3. Wave 4 patients were less likely to die in hospital than wave 1 patients (OR=0.4).

Discussion: The likely impact of the wave in the in-hospital mortality differs in patients with and without ceiling of care. Patients without a ceiling of care had greater odds of dying in hospital in the 1st wave than in the other waves, showing that the burden of care in hospitals could have a large impact on COVID-19 outcomes. For patients with ceiling of care, there were no differences between the first three waves, only patients from the wave 4 had greater odds of surviving. The three methods used gave similar results, showing that the analysis is consistent.

Keywords: Inverse probability weighting, Therapeuthic ceiling of care, COVID-19

References

1. Pallarès N et al (2023). Characteristics and Outcomes by Ceiling of Care of Subjects Hospitalized with COVID-19 During Four Waves of the Pandemic in a Metropolitan Area: A Multicenter Cohort Study. *Infect Dis Ther*, 12(1):273–89.

Análisis de la Exposición a Fármacos antidiabéticos Orales en la Población Metropolitana de Barcelona. Estudio sobre el análisis estadístico con sobre-representación de "Ceros"

Castillo P[1,2], Torres F[3], Tebé C[1]

[1]Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Badalona; [2] Universitat Autònoma de Barcelona, Bellaterra; [3]Departament of Pediatrics, Gynecology and Obstetrics, and Preventive Medicine, Universitat Autònoma de Barcelona, Bellaterra.

Introducción: Existe una importante variabilidad en la prescripción de fármacos antidiabéticos orales en la población adulta[1]. Dada la diversidad de medicamentos, se espera una sobre-representación de valores cero en la cuantificación de sus prescripciones. El estudio del impacto de factores demográficos, sociales y económicos en la prescripción requerirá trabajar con modelos específicos capaces de manejar tal sobre-representación de valores cero.

Métodos. Se dispone de una base de datos anónima sobre la dispensación en atención primaria de antidiabéticos orales en la población adulta del área metropolitana de Barcelona para el período 2013-2019. Para tener en cuenta la sobre-representación de valores cero, se aplicarán modelos estadísticos como *Zero-Inflated Poisson*, *Zero-Inflated Negative Binomial*, *Modified Gamma* y *Zero-Inflated Generalized Poisson*[2]. Se analizarán las diferencias y patrones en las prescripciones de medicamentos antidiabéticos en función del nivel socioeconómico, el cuartil de privación del área básica de salud (ABS), la edad y el sexo.

Resultados Esperados. Los modelos mencionados se espera que permitan discernir comportamientos asociados a la sobre-representación de ceros en los datos. Se anticipa que los pacientes con un estatus socioeconómico más alto podrían mostrar una mayor prescripción de fármacos en comparación con aquellos de estatus más bajo, esto podría sugerir una posible desigualdad en el uso de medicamentos antidiabéticos.

Keywords: Modelos *Zero-Inflated*, Antidiabéticos orales

References

1. Pautes per a l'harmonització del tractament farmacològic de la diabetis mellitus tipus 2. Barcelona: Servei Català de la Salut. Departament de Salut. Generalitat de Catalunya; 2017.(Programa d'harmonització farmacoterapèutica de medicaments en l'àmbit de l'atenció primària i comunitària del Servei Català de la Salut; 01/2017).

2. Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: beyond zero-inflated Poisson regression. British Journal of Mathematical and Statistical Psychology, 65(1),163–180. https://doi.org/10.1111/j.2044-8317.2011.02031.x.

Sesión 3: Genética e inferencia causal

8 de febrero, 9:00 a 10:15

Chair: Armand González Escalante

Analysis of Spatial Gene Expression Data: A Case Study in Neuroscience

Carlos Javier Peña[1]
[1]Biostatistics Unit - *IIS INCLIVA*

Introduction: Methods for dimensionality reduction are useful when modeling high-dimensional biological data. These methods typically assume independence of the observed samples, which means that they do not take into account the spatial and/or temporal dependencies between observations that arise from such designs. Dimensionality reduction with spatial information could provide more accurate and interpretable retrieval of underlying patterns by exploiting known spatial dependencies rather than only relying on feature correlations [1].
As an illustration, in the field of neuroscience it is known that the brain spatial organization is closely associated with its function. This link between structure and function is seen with special relevance in the layered arrangement of the human cerebral cortex. Cells dwelling within distinct cortical layers exhibit differential gene expression patterns. For instance, the human dorsolateral prefrontal cortex (DLPFC) is a brain area that has been implicated in several neuropsychiatric disorders [2].

Objective: The main purpose of this study was to identify spatially variable genes in order to gain further insight into gene expression in the setting of the spatial distribution of the human cortex.
Material: In this work, a public dataset of human dorsolateral prefrontal cortex (DLPFC) generated with the Visium technology from 10x Genomics was analyzed. Spatial gene expression was characterized in postmortem human tissue sections from one adult donor.
Methods: A factor analysis model incorporating a continuous covariate was used to take into account the spatial relationships between observations. In addition, a pathway enrichment analysis was performed to identify the biological functions that are overrepresented in the set of genes that make up the derived latent factors.
Results: Spatially differentiated latent factors were detected. It was also possible to determine that some of the genes, which constitute these latent factors, are involved in various neuropsychiatric and neurodevelopmental disorders.

Keywords: Spatial transcriptomics, Dimensionality reduction

References

1. Velten, B., Braunger, J.M., Argelaguet, R. *et al*. (2022). Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nature Methods*, **19**, 179-186.
https://doi.org/10.1038/s41592-021-01343-9

2. Maynard, K.R., Collado-Torres, L., Weber, L.M. *et al*. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neurosciense*, **24**, 425-436.
https://doi.org/10.1038/s41593-020-00787-0

Unraveling the molecular mechanisms associated with polycystic ovary syndrome using a multi-omics analysis strategy

E Géraud-Aguilar [1], M Insener [2], MA Martínez-García [2], F García García [3], S Barceló Cerdá [1]

[1]Departamento de Estadística e Investigación Operativa aplicadas y Calidad. Universitat Politècnica de València, Valencia, España; [2]Diabetes Obesity and Human Reproduction research Group, Department of Endocrinology and Nutrition, Hospital Universitario Ramón y Cajal & Universidad de Alcalá, Centro de Investigación Biomédica en Red Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain; [3]entro de Investigación Príncipe Felipe, Valencia, España

Polycystic ovary syndrome (PCOS) is a common disorder that affects women of reproductive age. Symptoms characterize it, including irregular or absent menstrual periods, excess androgens, and polycystic ovaries. The exact molecular cause of PCOS has yet to be fully understood. For this reason, a multi-omics approach is crucial. The main question of interest is whether sex hormones play a role in the interaction between the proteome, metabolome, and microbiome. Several multi-omic methods exist based on Machine Learning approaches. However, MOFA2 is a Bayesian framework that tries to explain the joint covariance of the omics by inferring latent variables, which are the key to understanding the molecular mechanism of PCOS. The method allows us to reconstruct the signal to perform downstream analysis. The univariate results are already published; for this reason, it is necessary not only to corroborate the published results but also to draw more information out of the data with the reconstructed signal to observe the benefits of the multi-omic approach using MOFA2.

Keywords: PCOS, multi-omics, latent variables

References

1. El Hayek, S., Bitar, L., Hamdar, L. H., Mirza, F. G., & Daoud, G. (2016). Poly Cystic Ovarian Syndrome: An Updated Overview. Frontiers in Physiology, 7, 124. https://doi.org/10.3389/fphys.2016.00124.

2. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology, 21(1), 111. https://doi.org/10.1186/s13059-020-02015-1

3. ut Microbiota and the Polycystic Ovary Syndrome: Influence of Sex, Sex Hormones, and Obesity. María Insenser, Mora Murri, Rosa Del Campo, M Ángeles Martínez-García, Elena Fernández-Durán, Héctor F Escobar-Morreale. J Clin Endocrinol Metab .2018 Jul 1;103(7):2552-2562. doi: 10.1210/jc.2017-02799. PMID: 29897462 DOI: 10.1210/jc.2017-02799

Feature Selection in *cell-free RNA-Seq* data for the detection of biomarkers with predictive value in pathology

E. Tercero-Atencia[1], B. Rosón[1*], A. Forte[2*]

[1]Department of of Computational Biology, Bioinformatics and Data Science, Carlos Simon Foundation-INCLIVA Health Research Institute, Valencia, Spain; [2]Deptartment of Statistics and OR, Universitat de Valencia, Doctor Moliner, 50, 46100-Burjassot, Spain; [*]Corresponding authors.

The identification of biomarkers capable of discerning between healthy and diseased individuals, classifying disease subtypes, and predicting early onset has become a critical focus in health sciences and biomedical studies. As the demand for minimally invasive diagnostic techniques rises, leveraging the power of transcriptomics becomes essential. The use of cell-free RNA-sequencing (cfRNA-seq) has provided a valuable tool for studying messenger RNA (mRNA) in peripheral samples, presenting a promising avenue for disease marker discovery. Nevertheless, the complexity of cfRNA-seq datasets and the challenge of detecting relevant mRNAs within the vast data volume require advanced statistical approaches that can reduce dimensionality while preserving biological significance, yielding results with significant medical utility and enhanced predictive capacity. These approaches are commonly accounted as feature selection techniques.

Given this challenge, we analyze two feature selection methods in cfRNA-seq data: Elastic Net and Bayesian Adaptive Sampling, from both frequentist and Bayesian perspectives. These methods will be applied to both simulated and real multiple myeloma data to ultimately determine which one performs better. Among other aspects, it has been observed that Bayesian Adaptive Sampling shows better performance in variable selection, identifying variables highly involved in the development and progression of multiple myeloma. Additionally, it demonstrates superior disease prediction performance, with a preliminary model achieving AUC and precision values of 0.90 and 0.83, respectively, while maintaining high sensitivity and specificity. However, the results have also revealed limitations of both methods concerning biological data, highlighting the need for further analysis to draw more conclusive outcomes.

Keywords: biomarker, cell free RNA-seq, feature selection

El impacto del tabaco provoca cambios epigenéticos en el tejido adiposo subcutáneo: Implicaciones en la Progresión de la Enfermedad de Crohn

Irene Vañó-Segarra[1], Diandra Monfort-Ferré[1], Albert Boronat-Toscano[1], Menacho M, Valldosera G[2], Caro A[2], Alfonso Saera-Vila[3], Laura Clua-Ferré[4], Josep Manyé[5], Carolina Serena[1]

[1]Hospital Universitari Joan XXIII de Tarargona. Institut de Investigació Sanitària Pere i Virgi,

[2]Sequentia Biotech, Carrer Comte d'Urgell 240, 08036 Barcelona, Spain,

[3]Institut d'Investigació Germans Trias i Pujol,

[4]Institut d'Investigació Germans Trias i Pujol, CIBERehd

Introducción: Fumar tiene un fuerte impacto negativo en la enfermedad de Crohn (EC), aumentando el riesgo de inicio temprano y recurrencia postoperatoria. Este estudio se centró en la epigenética, especialmente en la metilación del ADN, para investigar posibles cambios en la metilación del tejido adiposo subcutáneo (SAT) entre fumadores y no fumadores con EC. El objetivo era identificar genes clave que influyen en la enfermedad. Nuestra hipótesis planteaba que el tabaco podría alterar la metilación de genes implicados en la EC, lo que podría explicar el empeoramiento del pronóstico en pacientes fumadores.

Métodos: Utilizamos un análisis de amplia cobertura del epigenoma utilizando un array de metilación (Illumina EPIC/450k array) para explorar todo el tejido adiposo subcutáneo en pacientes con EC, fumadores (n=3) y no fumadores (n=2). Utilizando las herramientas dmpfinder y Bumphunter (minfi), identificamos las posiciones (DMPs) y las regiones (DMRs) con metilación diferencial. Además, realizamos un análisis de la expresión génica de los genes con el mayor número de DMRs para establecer correlaciones entre su metilación y la posterior expresión génica. Luego, realizamos un estudio funcional de las vías más afectadas.

Resultados: Identificamos un patrón de metilación del ADN distintivo en el SAT de fumadores con EC en comparación con no fumadores (Fig.1A). Encontramos más de 27,270 DMPs que están involucrados en importantes procesos biológicos relacionados con el daño del ADN, la inflamación o la inmunidad (Fig.1B). Adicionalmente, encontramos más de 13,000 DMRs ubicadas en el promotor (Fig.1B). Entre los genes con un mayor número de DMRs, encontramos correlaciones entre los niveles de metilación y la expresión génica posterior, como en los casos de PPARG y FOXP1, o con genes fundamentales en la EC, como NOD2 y TNF. Específicamente, observamos una mayor metilación de NOD2 en los fumadores en comparación con los no fumadores (valor de p=0.0356), junto con una tendencia a una disminución en los niveles de expresión (Fig.1D).

Conclusiones: El tabaco induce una serie de cambios epigenéticos en pacientes con EC, que podrían empeorar el pronóstico en los fumadores. Además, hemos observado correlaciones entre la metilación y los niveles de expresión, lo que proporciona una mayor comprensión de la complejidad de la enfermedad.

Keywords: Crohn, Metilación del DNA, Tabaco

Meta-Analysis of Epigenome wide association study of DNA Methylation and Ultra-Processed Food Consumption in middle-childhood

Joana Llauradó-Pont[1], Nikos Stratakis[2,3], Giovanni Fiorito[4,] Evangelos Chandakas[5], Mariona Bustamante[6], Camille Lasale[1,7,8,9]

[1]Barcelona Institute of Global Health (ISGlobal), Barcelona, Spain. [2]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA. [3]Barcelona Institute of Global Health (ISGlobal). [4]Clinical Bioinformatics Unit, IRCCS Istituto Giannina Gaslini. [5]Medical Research Council Centre for Environment and Health, School of Public Health, Imperial College London, London, United Kingdom. [6]ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, [7]SpainHospital del Mar Research Institute (IMIM), Dr Aiguader, 88, 08003, Barcelona, Spain. [8]Consortium for Biomedical Research - Pathophysiology of Obesity and Nutrition (CIBEROBN), Instituto de Salud Carlos III, Madrid, Spain.[9] Universitat Pompeu Fabra (UPF), Barcelona, Spain

The worldwide increase in ultra-processed food (UPF) consumption, resulting from extensive industrial processing and the incorporation of various additives, signifies a compromise in dietary health. Nutrition can modulate the epigenome with both positive and negative implications for human health. While prior studies have focused on the influence of diet on the epigenome of adults, there is an increasing interest in the effects of ultra-processed foods (UPFs) on the human epigenome, not only in adults but also in other life stages. This study contributes to filling this gap by investigating the correlation between UPF consumption and DNA methylation in children aged 7 to 11 years from three European birth cohorts (ALSPAC, Generation XXI, and HELIX consortium).

Ultra-processed food (UPF) consumption was evaluated through the NOVA classification, and the measurement of DNA methylation was done in whole blood using Illumina arrays. Associations were estimated within each cohort using robust linear regression models, adjusting for potential confounders such as age, sex, education, among others.

Meta-analysis across cohorts reveals suggestive associations between UPF consumption and differential methylation at the following CpG sites: cg00163372, cg14665028, cg18968409, cg26295786, cg00339913 and cg09709951. Functional and enrichment analyses linked some of these CpGs with MYC, PHYHIP, and NHEJ1 genes and highlighted potential age-related effects and associations with immune and proliferation-related functions. Despite limitations, this study provides valuable insights into the epigenetic implications of UPF consumption in children, emphasizing the need for further research to elucidate these associations.

Keywords: Epigenome Wide Association Study (EWAS), Ultra Processed Food (UPF), DNA methylation

Sesión 4: BIOSTATNET

8 de febrero, 10:45 a 12:30

Chair: David Moriña Soler

Short-time cancer incidence projections addressing missing data challenges

Garazi Retegui[1,2], Jaione Etxeberria[1,2], María Dolores Ugarte[1,2]
[1]Department of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA), Arrosadia Campus, 31006, Pamplona, Spain; [2]Institute for Advanced Materials and Mathematics (INAMAT2), Public University of Navarre (UPNA), Arrosadia Campus, 31006, Pamplona, Spain.

Cancer control aims to reduce cancer mortality through the systematic implementation of evidence-based interventions in prevention, early detection, treatment and palliative care. Thus, cancer data, including information on cancer incidence and mortality, are essential to understand cancer burden, to set targets for cancer control and to evaluate the evolution of the implementation of cancer control policies. In this context, regional Population-Based Cancer Registries (PBCRs) are responsible for assessing the current magnitude of the cancer burden occurring in areas within a country (states or provinces for example). However, these figures are usually available with a delay and therefore, national or regional PBCRs have a strong interest in using methods to forecast them. Furthermore, in large countries, regional Population-Based Cancer Registries (PBCRs) are often established in different years, resulting in incomplete data series on cancer incidence among various regions. This result in non-harmonised data series within a country, with lack of information mainly at the beginning of the data series. In this work, we focus on deriving short-time cancer incidence predictions in the presence of missing data. In particular, we propose flexible shared component spatio-temporal models that include interactions with a time-varying scaling parameter for the joint analysis of mortality and available cancer incidence data. The new models will be implemented in INLA. The performance of our proposal will be analyzed using cancer incidence and mortality data reported in England.

Keywords: cancer burden, predictions, shared component models

The individual causal association for the evaluation of surrogate endpoints based on causal Inference under non-normality

Gokce Deliorman[1], Florian Stijven[2], Wim Van der Elst[3], Maria del Carmen Pardo[4], Ariel Alonso Abad[5]

[1-4]Department of Statistics and OR, Complutense University of Madrid;
[3]Janssen Pharmaceutica, Companies of Johnson & Johnson;
[2-5]I-BioStat, KU Leuven Statistics;
[4]Interdisciplinary Mathematics Institute (IMI), Complutense University of Madrid

The use of surrogate endpoints in clinical trials is becoming increasingly common as a way to develop new treatments. Surrogate endpoints may be used instead of the true endpoint which is the most clinically relevant outcome and measured directly and they can be measured earlier, allowing clinical benefit to be predicted earlier. Proving the effect of the treatment on the surrogate and showing that it can accurately predict the clinical benefit, however, is not always straightforward. For this purpose, Alonso *et. al* (2015) introduced a new measure of surrogacy, the so-called individual causal association (ICA) to assess the validity of the surrogate in the causal inference framework, when both endpoints are continuous and normally distributed. ICA works properly for the normal causal model. However, if the model misspecifications, then assessments of surrogacy may lead to a wrong conclusion. In this work, ICA's performance is evaluated when the model has other distributions that are not normal.

Keywords: Individual causal association, non-normality, surrogate endpoint

References

1. Alonso, A., Van der Elst, W., Molenberghs, G., Buyse, M., & Burzykowski, T. (2015). On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics, 71(1)*, 15-24.

Double data fusion and calibration modeling for zooplankton abundance measured in space and time

Jorge Castillo-Mateo[1], Alan E. Gelfand[2], Robert S. Schick[3]
[1]Department of Statistical Methods, University of Zaragoza;
[2]Department of Statistical Science, Duke University;
[3]Nicholas School of the Environment, Duke University

To better understanding the distribution and abundance of marine predator species it is a critical yet difficult challenge to understand the spatiotemporal dynamics of their prey. These prey species are small, difficult to sample, and respond to a variety of biotic and abiotic features. The data are often collected at points, yet it is imperative to understand the broader dynamics to better predict how predators may respond. In addition, the data are often collected with a variety of sampling mechanisms that are usually analyzed separately. In Castillo-Mateo *et al.* [1], we propose a space-time hierarchical model to understand the distribution and abundance of zooplankton species in Cape Cod Bay (CCB), MA, USA during the period 2003–2019. The model is comprised of a double data fusion and calibration of different sampling mechanisms.

At a given location in space and time, we have two sources of data collection for zooplankton to inform about sea surface zooplankton abundance. We offer a fusion of these two response sources which requires calibration. An important driver of sea surface zooplankton abundance is sea surface temperature. We have two data sources to supply sea surface temperature which we fuse to develop a sea surface temperature regressor. These sources also require calibration, hence our terminology of *double* fusion and *double* calibration.

The data are collected at varying sites for varying days within the first half of the year over the 17 year period. We use the fusion model to develop prediction of daily spatial zooplankton abundance surfaces over CCB. We present suitable integration of the abundance surface over CCB to infer about average abundance on a given day within a given year in CCB. We extend the inference to consider abundance averaged to monthly or annual scale as well as to make annual comparison of abundance.

Keywords: Bayesian melding, Gaussian process, Markov chain Monte Carlo

References

1. Castillo-Mateo J., Gelfand A.E., Hudak C.A., Mayo C.A., and Schick R.S. (*in press*). Space-time multi-level modeling for zooplankton abundance employing double data fusion and calibration. *Environmental and Ecological Statistics*.

Time-dependent AUC for survival and competing risks models

Leire Garmendia Bergés[1,2], Irantzu Barrio[1,2], Guadalupe Gómez Melis[3]
[1]BCAM - Basque Center for Applied Mathematics; [2]Department of Mathematics, University of the Basque Country UPV/EHU; [3] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya

Competing risks situations appear often in survival analysis when the endpoint of interest (i.e., recovery) is precluded by another event (i.e., death). The probability of failure in the presence of competing risks can be modeled through cause-specific hazards or the incidence function. In addition, these models are commonly used to predict the course of future individuals and in this case, their predictive capacity has to be evaluated.

There are several methods to measure the predictive capacity of a model. In survival analysis, the area under the time-dependent ROC curve (AUC(t)) is commonly used to quantify the ability of a survival model to correctly predict future events (non-events) at a time t. The estimation of the AUC(t) is not straightforward under censoring situations, so different estimators of the AUC(t) have been proposed in the literature in order to work with these censored times [1].

In the presence of competing events, the time-dependent AUC can be extended. One event type is considered the main event, and the AUC(t) or partial AUC(t) is estimated for that transition. For that, different definitions and estimators of the partial AUC(t) have been proposed in the literature [2].

The objective of this work is to propose a global AUC(t) to quantify the predictive capacity of a competing risks model by means of the partial AUC(t)s obtained in each transition. With that aim, different simulation studies have been done to compare the different AUC(t) estimators for survival and competing risks models and choose the best estimator.

Keywords: Survival analysis, competing risks models, time-dependent AUC

References

1. Blanche P., Dartigues J.F., and Jacqmin-Gadda H. (2013). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55, 687-704.

2. Blanche P., Dartigues J.F., and Jacqmin-Gadda H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine, 32,* 5381-5397.

Survival analysis in genetics

María del Pilar González Barquero[1], Rosa Elvira Lillo Rodríguez[2], Álvaro Méndez Civieta[3]

[1]Department of Statistics, Universidad Carlos III de Madrid; [2]Department of Statistics, Universidad Carlos III de Madrid; [3]Department of Statistics, Universidad Carlos III de Madrid

The motivation behind this study stems from the need to address a critical problem within the framework of survival analysis, which is a statistical technique employed to investigate time-to-event data. This problem consists of determining the variables that influence the survival of patients with triple-negative breast cancer (TNBC), which is a type breast cancer with low survival rates due to its aggressive nature. For this case study, a real high-dimensional dataset was provided by the Gregorio Marañón Health Research Institute containing clinical and genetic information from patients with TNBC that are or have been treated with a specific type of chemotherapy and their survival time measured from the beginning of the treatment until death.

This work assesses the effectiveness of Cox regression models in the context of high-dimensional data and high proportion of censure, where dimensionality reduction techniques are crucial for model interpretability and predictive accuracy. Two regularization techniques are evaluated: the lasso penalty and the adaptive lasso penalty. One of the key contributions of this project lies in the investigation of various weight calculation procedures for the adaptive lasso. These proposed weights are based on Principal Component Analysis, ridge regression, univariable Cox regressions, and the Random Survival Forest (RSF) algorithm. The paper compiles the evidence obtained via a very exhaustive simulation study and the findings related to the real database.

Keywords: Survival analysis, Adaptive Lasso, TNBC.

A comparison between the Bayesian MCMC software packages WinBUGS and NIMBLE for modeling spatial ordinal survey-based data

Miguel Ángel Beltrán Sánchez[1], Miguel Ángel Martínez Beneito[1], Ana Corberán Vallet[1]
[1]Departamento de Estadística e Investigación Operativa, Universitat de València

Health surveys allow exploring health indicators such as health habits, mental health issues, physical limitation problems, social support needs…that are of great value from a public health point of view. These indicators, typically coded as ordinal variables, may depend on covariates associated with individuals. We introduce a Bayesian individual-level model for small-area estimation of survey-based health indicators. A categorical likelihood is used at the first level of the model hierarchy to describe the ordinal data, and spatial dependence among small areas is accounted for using the Leroux conditional autoregressive (CAR) distribution [1]. Post-stratification of the results of the proposed individual-level model allows extrapolating the results to any administrative areal division, even for small areas.

We apply this methodology to the analysis of the Health Survey of the Region of Valencia (Spain) to describe the geographical distribution of different health indicators of interest in this region. Specifically, we present the implementation of this model in both WinBUGS [2] and NIMBLE [3], providing a comparison between the two Bayesian MCMC software packages. Overall, NIMBLE emerges as one of the most robust modern proposals in Bayesian MCMC software for running models like the one presented.

Keywords: Bayesian inference, spatial statistics, survey-based studies

References

1. Leroux B.G., Lei X., Breslow N. (2000) Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*.

2. Lunn D.J., Thomas A., Best N., Spiegelhalter D. (2000). WinBUGS – A Bayesian Modelling Framework: Concepts, Structure, and Extensibility. *Statistics and Computing*, 10(4): 325–337.

3. Valpine dP., Turek D., Paciorek C.J., Anderson-Bergman C., Temple Lang D., Bodik R. (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2): 403–413.

Sesión 5: Modeling

8 de febrero, 14:30 a 15:30

Chair: Alba Fuster Alonso

Spatial point processes: from the mathematical basis to its applications

Arnau García[1]

[1]Department of Mathematics, University of Barcelona.

This work is a study about the spatial point processes. We study the mathematical basis of this object, we expose statistical tools which are used in the analysis of spatial point patterns and, finally, we apply all the exposed theory in a real case study with real data.

In the first and second chapter we present the mathematical theory behind the spatial point processes. In this chapter we have used Stoyan et al. 2013 $[1]$. In the second chapter, using as a reference Diggle 2013 $[2]$, we explain the mathematical theory of the point processes in the plane.

In the third chapter, based mainly in Baddeley et al. (2015) [3], we present, giving examples, the statistical tools used in the analysis of point processes in the plane and how to use it in R.

Finally, in the last chapter, we put into practice all the knowledge we have acquired in a real case study. Using the database employed in Jorge Mateu, P. Diggle and I. Tamayo-Uria (2014) [4], shared by Jorge Mateu, we perform a study about the rat and cockroach sightings in Madrid city.

Keywords: Spatial point processes.

References

1. Dietrich Stoyan, Wilfrid S Kendall, Sung Nok Chiu, and Joseph Mecke. Stochastic geometry and its applications. John Wiley and Sons, 2013.

2. Peter J Diggle. Statistical analysis of spatial and spatio-temporal point patterns. CRC press, 2013.

3. Adrian Baddeley, Ege Rubak, and Rolf Turner. Spatial point patterns: method-ology and applications with R. CRC press, 2015.

4. Ibon Tamayo-Uria, Jorge Mateu, and Peter J Diggle. Modelling of the spatio-temporal distribution of rat sightings in an urban environment. Spatial Statistics, 9:192–206, 201

A survival analysis and multilevel modeling study to approach gender bias in Time Compactation

Gonzalo Aparicio Rodríguez[1], José Antonio Villacorta Atienza[2], Abel Sánchez Jiménez[3]

[1]Department of Biomathematics, Complutense University of Madrid; [2] Department of Biomathematics, Complutense University of Madrid; [3] Department of Biomathematics, Complutense University of Madrid

Time Compaction is a salient cognitive mechanism used by humans [1] to deal with complex dynamic situations, i.e., environments in which spatial relations between subject and elements change rapidly over time. This mechanism deletes temporal dimension from moving elements by estimating only the locations of future interactions [2], i.e., where subject and elements will coincide. The brain then generates an internal map spatially arranging these future interactions, named compact internal representation or CIR. While this cognitive mechanism is salient in men, women use a wider set of cognitive strategies [1].

This gender bias may result from cognitive sexual dimorphism or from cultural aspects. We focused on sport practice as a paradigm of cultural dimorphism. Thus, we have analyzed through multilevel modeling and survival analysis techniques the results of visual discrimination tests performed by athletes and non-athletes, in which the recurrence to CIR's would facilitate a better performance and higher learning speed in the first attempts (evaluated by mixed models) and a shorter time to find the hidden key in the tests (evaluated by survival analysis).

During tests athletes have a lower recurrence to CIR than non-athletes, demonstrating for the first time the effect of environmental factors on time compaction. Furthermore, among athletes, women recur to a greater extent to CIR than men. Multilevel modelling and survival analysis have proven to be accurate and complementary tools to approach cognitive experiments.

Keywords: Spatiotemporal cognition, Mixed models, Survival analysis

References

1. Villacorta-Atienza J.A., Calvo Tapia C., Díez-Hermano S., Sánchez-Jiménez A., Lobov S., Krilova N., et al. (2021). Static internal representation of dynamic situations reveals time compaction in human cognition. *Journal of advanced research, 28*, 111–25.

2. Villacorta-Atienza J.A., Velarde M.G., and Makarov V.A. (2010). Compact internal representation of dynamic situations: neural network implementing the causality principle. *Biological Cybernetics*, *103*, 285–97.

Spatial Bayesian Distributed lag non-linear models: a case study of small-area temperature-mortality association

Marcos Quijal-Zamorano[1,2]; Miguel A. Martinez-Beneito[3,4]; Joan Ballester[1]; Marc Marí-Dell'Olmo[5,6,7]

Affiliations:

1 ISGlobal, Barcelona, Spain

2 Universitat Pompeu Fabra (UPF), Barcelona, Spain

3 Departament d'Estadística i Investigaciò Operativa, Dr. Moliner 50, 46100, Burjassot, Valencia, Spain

4 Unitat Mixta de recerca en mètodes estadístics per a dades biomédiques i sanitàries, UV-FISABIO, Dr. Moliner 50, 46100 Burjassot, Valencia, Spain

5 Agència de Salut Pública de Barcelona (ASPB), Pl. Lesseps 1, 08023 Barcelona, Spain

6 Institut d'Investigació Biomèdica Sant Pau (IIB SANT PAU), Sant Quintí 77-79, 08041 Barcelona, Spain

7 Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Av. Monforte de Lemos, 3-5, 28029 Madrid, Spain

In the context of climate change and increasing temperatures, the interest in the health effects of environmental exposures has remarkably increased. The development of the distributed lag non-linear models (DLNM) become rapidly the referent framework when studying temperature-mortality short-term associations. DLNMs facilitates the modelling of the non-linear and lagged effect of temperatures on mortality. However, the small-area analysis of temperature-mortality is scarce. In that sense, here we present four models. The first two models generalize standard DLNMs to a Bayesian framework, using a case-crossover design (model 1) and the common DLNM time-series configuration, where time trends and seasonality are modelled by using splines (model 2). Models 3 and 4 are extensions of model 1 and model 2 respectively, where we propose Leroux models to spatially-smooth in one-stage approaches the coefficients of the exposure-response relationships. We propose these last two models specifically for dealing with the noise from small numbers in small-area analysis. We apply all proposed models to a case-study for assessing the temperature-mortality relationships in the 73 neighborhoods of Barcelona. 39.569 deaths were considered in the period 2007-2016, 19 of them corresponding to the neighborhood with lower number of deaths and 1.454 deaths to the one with higher number. Curves defining the relative risks of mortality were noisy and uncertain in the independent models, with regions with extremely high and others with extremely low risks distributed all over the city. Spatial models benefit from adjacent regions to smooth the association and reveal hidden spatial-patterns of risk. In addition, the flexibility of these Bayesian models allowed us to explore the results of these epidemiological models in new-intuitive ways. This novel multidimensional approach brings the opportunity to estimate ecological temperature-mortality models in a smaller spatial scale to better understand the socioeconomic factors driving the effect of temperature on human health.

Keywords: temperature, mortality, Bayesian, distributed lag non-linear models, spatial models.

Statistical aproaches to correct for baselines in clinical trials

Matilde Francisco[1], Klaus Langohr[2]
[1]Department of Statistics and Operational Research, Polytechnic University of Catalonia;
[2]Department of Statistics and Operational Research, Polytechnic University of Catalonia;

Longitudinal studies allow the repeated monitoring of health outcomes or risk factors, and the identification of differences in outcomes. Baseline measurements are demographic characteristics or measurements taken at the beginning of the study of the response variable or variables correlated with it. Whether to consider baseline as a covariate or a dependent variable is a frequently asked question. Not accounting for baseline, can not only affect the magnitude of differences detected in a study, but also the direction of these differences, which can result in different clinical conclusions. The lack of consistency in the literature around this topic contributes to the difficulty to establish a standard statistical approach, so studies' specific characteristics influence the decision on what statistical approach should be used.
When adjusting a model, it is possible to adopt different strategies regarding the use of baseline. It can be included in the model as a covariate, and the post-baseline values are the response variable or, assuming that the randomization of the subjects involved was efficient, baseline and post-randomization values can be treated as dependent variables.
In this work, two different methods, constrained longitudinal data analysis (cLDA) and analysis of covariance (ANCOVA) were applied to a real data set from a clinical trial and simulated data, in order to study the behaviour of the methods under different conditions and try to figure out what would be the best approach. The obtained results indicate that cLDA can be appropriate in the cases where data follows a normal distribution, and its application can bring advantages especially in the presence of missing data. However, when there is a deviation from normality, ANCOVA showed to be a better approach regardless of the other conditions.

Keywords: Baseline-value adjustment, linear mixed models, longitudinal studies.

References

1. Liu, Guanghan F and Lu, Kaifeng and Mogg, Robin and Mallick, Madhuja and Mehrotra, Devan V (2009). Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? *Statistics in Medicine*

2. Senn, Stephen (2014). Baseline adjustment in longitudinal studies *Wiley StatsRef: Statistics Reference Online*

3. Liang, Kung-Yee and Zeger, Scott L (2000). Longitudinal data analysis of continuous and discrete responses for pre-post designs *The Indian Journal of Statistics, Series B*

Sesión 6: Survival

8 de febrero, 15:30 a 16:30

Chair: Andrea Toloba López-Egea

MLE-based approach for inference of the clustered-state Markovian arrival process for recurrence-death data in patients with oncological diseases

Álvaro Díaz Pérez[1], Rosa Elvira Lillo Rodríguez[1 2],Pepa Ramírez Cobo[3]
[1]uc3m-Santander Big Data Institute (IBiDat); [2]Department of Statistics, Universidad Carlos III de Madrid; [3]Departamento de Estadística e Investigación Operativa, Universidad de Cádiz

A real dataset related to the recurrence and death of patients with oncological diseases motivates the extension of the Markovian arrival process to the clustered-state version. Novel closed-forms concerning this new model are presented, including explicit expressions for the moments and correlations and for the marginal and joint densities of the inter-event times. Additionally, we derive an explicit expression for the likelihood function and propose a simplified version to improve computational efficiency. Furthermore, we present an MLE-based approach for inference of this new model specifically for recurrence-death data with censoring. This approach is illustrated with simulated data and applied to a real dataset of bladder cancer.

Keywords: Markovian arrival process, recurrence-death, survival analysis

References

1. L. J. Wei, D. Y. Lin, and L. Weissfeld (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*, Vol. 84, No. 408.

2. Ramírez-Cobo, P, Carrizosa, E., Lillo, R. E., (2021). Analysis of an aggregate loss model in a Markov renewal regime. *Applied Mathematics and Computation*, Vol. 396.

3. Ramírez-Cobo, P y Lillo, R. E. (2012). New results about weakly equivalent MAP2 and MAP3 processes. *Methodology and Computing in Applied Probability*, 14 (3), 421-444.

Evaluation of discrimination ability of time-dependent variables in a Cox proportional hazards model

Antia Enriquez Yurrebaso[1,2], Irantzu Barrio Beraza[1,2]
[1]Department of Mathematics, University of the Basque Country UPV/EHU; [2]BCAM - Basque Center for Applied Mathematics

The use of prognostic models is becoming more and more widespread in daily practice, where they have become relevant as a support for decision-making. When the interest is focused on predicting the time until an event occurs (mortality, machine failure, etc.), survival analysis is used. One of the most widely used models in survival analysis is the Cox Proportional Hazards (CPH) model. Thus, the goal is to predict the survival time based on some variables, which may or may not vary over the study time.

In order to obtain adequate models for the prediction of results, a prognostic model must have a high predictive or discriminatory capacity, being able to distinguish between those individuals with and without the event at each instant of time. Therefore, a crucial aspect is the availability of appropriate measures to evaluate the predictive capacity of the model.

Over time, different measures and estimators have been proposed to calculate the discriminatory capacity of CPH models. These include Harrell's concordance index (C-index), which is an extension of the area under the ROC curve (AUC) to the case of censored survival data and different definitions of time-dependent AUC, i.e., AUC$^{C/D}$ (cumulative/dynamic) or AUC$^{I/D}$ (incident/dynamic).

Some studies have been carried out comparing different measures and estimators when time independent variables are used, observing differences between the estimators proposed $[1, 2]$. However, to the best of our knowledge, no studies that take into account time-dependent variables have been conducted. Therefore, our objective has been to study the advantages and disadvantages of the different measures and estimators proposed in the literature in the presence of time-dependent variables.

To do this, a review of the measures and estimators proposed in the literature has been carried out to point out their differences and similarities. Afterward, a simulation study has been done, with the aim of observing the behaviour of each measure and finally, we present an application to a COPD (chronic obstructive pulmonary disease) cohort. The results suggest that in presence of time-dependent variables the discriminatory capacity increases.

Keywords: Survival Analysis, time-dependent, discriminatory capacity

References

1. Kamarudin A.N., Cox T. and Kolamunnage-Dona R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, 17, 53.

2. Blanche P., Dartigues J.F. and Jacqmin-Gadda H. (2013). Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5), 687–704.

Identificación de factores de riesgo para la supervivencia de pacientes con cáncer colorrectal mediante un análisis de riesgos competitivos

María Gascón[1,2,4], Juan Eloy Ruiz[5], María José Legarreta[1,2,4], María Amparo Valverde[1,4], José María Quintana[1,2,3,4], Urko Aguirre[1,2,3,4]

[1]Osakidetza Servicio Vasco de Salud, Hospital Galdakao-Usansolo, Vizcaya; [2]Red de Investigación en Cronicidad, Atención Primaria y Promoción de la Salud (RICAPPS), España; [3]Red de Investigación de Servicios de Salud en Enfermedades Crónicas (REDISSEC), España; [4]Instituto Biosistemak de Investigación en Servicios Sanitarios, Vizcaya; [5]Departamento de Estadística e Investigación Operativa Universidad de Granada (UGR), Granada;

Introducción y objetivos: El cáncer colorrectal (CCR) se encuentra actualmente entre los cánceres más frecuentes tanto en mujeres como en hombres y es importante desarrollar reglas de predicción clínica para predecir su evolución. El análisis de supervivencia se centra en la modelización del tiempo hasta que ocurre un evento concreto como puede ser el fallecimiento. El enfoque de los modelos de riesgos competitivos complementa el análisis de supervivencia convencional al evaluar el riesgo de un evento especifico en presencia de otros eventos competitivos. El objetivo de este estudio es identificar los factores de riesgo que predicen la mortalidad específica a 5 años por CCR y la mortalidad por otras causas.

Métodos y resultados: Se reclutaron 898 pacientes de hospitales públicos pertenecientes al Servicio Vasco de Salud con diagnóstico de CCR y que fueron intervenidos quirúrgicamente entre junio de 2010 y diciembre de 2012. Se les realizó un seguimiento en diferentes puntos temporales hasta los cinco años después de la intervención. Se llevó a cabo un análisis de riesgos competitivos para predecir la mortalidad específica por cáncer colorrectal y por otras causas, evaluando un modelo de supervivencia para cada causa. Los resultados mostraron que el 64 % de los participantes eran hombres con una edad media de 68 años. Hubo un total de 254 fallecimientos, de los cuales 173 fueron debido a CCR. Al analizar la supervivencia específica, se identificaron los siguientes factores de riesgo: alcoholismo, edad, riesgo de ASA, resultado de la cirugía, invasión de órganos adyacentes, escala pTNM, nivel de leucocitos y presencia de anemia. En cuanto al modelo de mortalidad por otras causas, se mantuvieron como predictores el alcoholismo, la edad, el resultado de la cirugía y la anemia, añadiendo las variables del índice de comorbilidad de Charlson, el tratamiento con quimioterapia y radioterapia y el PLR (ratio de plaquetas-linfocitos). Ambos modelos obtuvieron una buena capacidad predictiva con un c-index de 0,785 y 0,82 respectivamente.

Conclusiones: Existen diferencias en los predictores de riesgo de mortalidad específica por cáncer colorrectal y por otras causas.

Keywords: Análisis de riesgos competitivos, Cáncer colorrectal, Supervivencia

References

1. Beyersmann J., Allignol A., Schumacher M. (2012). *Competing Risks and Multistate Models with R.* Springer.

A joint model for (un)bounded longitudinal markers, competing risks, and recurrent events using patient registry data

Pedro Miranda Afonso[1], Dimitris Rizopoulos[2], Eleni-Rosalina Andrinopoulou[3]
[1]Department of Biostatistics, Erasmus Medical Center

Cystic fibrosis (CF) is a severe genetic disorder that primarily affects the lungs and digestive system, leading to respiratory impairment and malnutrition. Patients with CF often experience recurrent lung infections, known as pulmonary exacerbations (PEx), which can cause permanent lung damage and increase the risks of lung transplantation and death. The BMI and the percentage of predicted forced expiratory volume in one second (ppFEV$_1$) are routinely measured to monitor disease progression. CF care teams are interested in understanding the associations between ppFEV$_1$ decline, BMI changes, recurrent PEx, and the competing risks of death and lung transplantation using the US Cystic Fibrosis Foundation Patient Registry. Previous studies that aimed to investigate such associations using joint models were hampered by the lack of an appropriate framework.

The joint modeling framework has previously been extended to incorporate complex survival data structures, such as recurrent and competing event time data. However, the integration of both recurrent events and competing risks within a unified model remains a challenge, leading researchers to omit important information. An additional limitation of existing frameworks is their tendency to rely on exclusively on the Gaussian distribution to model continuous markers. An important aspect of joint modeling is the appropriate parameterization of longitudinal submodels to ensure accurate extrapolation of unobserved biomarker evolution up to the event time. A Gaussian parameterization can be problematic for a bounded biomarker with many observations close to the boundaries, such as ppFEV$_1$, as it can cause the model to yield biologically implausible values, resulting in biased estimates of the marker evolution and its associations.

We address these limitations by introducing a comprehensive joint modeling framework that can (i) effectively accommodate competing risk and recurrent event processes together with multiple longitudinal outcomes, and (ii) appropriately model bounded longitudinal markers with constrained distributions. Our model captures the complex dynamics of CF by simultaneously considering recurrent PEx and the competing risks of death and lung transplantation, and by appropriately parameterizing the longitudinal markers ppFEV$_1$ and BMI using beta and Gaussian distributions, respectively. The model allows for the use of various functional forms to link time-to-event and longitudinal processes, and accommodates discontinuous risk intervals and both gap and calendar timescales. The model has been made available in the user-friendly CRAN R package for joint models, `JMbayes2`.

Keywords: multivariate longitudinal data, competing risks, recurrent events

Sesión 7: Machine Learning y software

8 de febrero, 17:00 a 17:45

Chair: Pavel Hernández Amaro

Efficient and reproducible table summaries with gtsummary in R

Carmezim J[1], Satorra P[1]
[1]Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Badalona, Spain;

A fundamental aspect of data analysis involves generating and exporting refined and easily reproducible tables. The gtsummary package [1] serves as a comprehensive R tool designed to streamline the process of creating informative and visually appealing summary tables for data analysis. It provides a flexible structure to meet the needs of analysts and researchers who want to confidently communicate their results. One of the key strengths of gtsummary is its emphasis on reproducibility. Analysts can easily produce polished, publication-ready tables that can be replicated across different projects and datasets. This package integrates easily with popular R packages such as dplyr and ggplot2, handles diverse data types such as statistical models or survival data, and simplifies the process of comparing and sharing tables. For descriptive tables, it automatically detects continuous, categorical and dichotomous variables and calculates the appropriate descriptive statistics including the amount of missing values in each variable. For regression models, it will automatically recognize the model and fill the table with relevant statistics like coefficients, confidence intervals, p-values. The package offers a wide range of customisation options, for example add p-value to summary tables or add model fit statistics to regression model tables. This allow users to tailor the appearance and content of tables to suit specific reporting requirements. Additionally, the ability to merge tables side-by-side or stack them on top of each other facilitates efficient synthesis of output from multiple sources. By adopting gtsummary, researchers can significantly enhance the transparency and impact of their data presentations, making it an indispensable tool across a wide range of academic disciplines.

Keywords: Statistical reporting, Reproducibility, R software

References

1. Sjoberg D, Whiting K, Curry M, Lavery J, Larmarange J (2021). Reproducible summary tables with the gtsummary package. *The R Journal*, 13, 570–580. doi:10.32614/RJ-2021-053, https://doi.org/10.32614/RJ-2021-053.

Detección de fraude alimentario en leche: Análisis de especiación de leche y detección de leche de cabra adulterada con leches de menor calidad, empleando aprendizaje automático e implementación en aplicación web

López González, Miguel Ángel[1]

[1]Estudiante del máster en Bioinformática i Bioestadística, Universitat Oberta de Catalunya

El fraude alimentario es un riesgo que compromete la calidad y seguridad alimentaria e implica un agravio económico. El presente trabajo realiza un estudio de especiación y de adulteración de leches a partir de datos de espectrometría de masas. Tiene la finalidad de encontrar los modelos más eficientes e implementarlos en una herramienta que haga accesible la detección de fraude en leche. Se han empleado los algoritmos: support vector machines (SVM), artificial neural networks (NN) y random forests (RF), diseñando varios modelos con diferentes parámetros. La alta dimensionalidad de los datos y escasez de muestras ha hecho necesario un aumento de muestras mediante la técnica SMOTE y reducir dimensiones mediante principal component analysis (PCA). La aplicación de estas técnicas juntas y por separado ha generado cuatro escenarios de trabajo. La determinación del mejor modelo se basó en las métricas de exactitud, Kappa, sensibilidad y especificidad, además de priorizar el modelo más simple. Se escogieron como óptimos los modelos SVM con kernel lineal, consiguiendo un 100% de exactitud en especiación y un 96.3% en adulteración. Se ha demostrado la capacidad de los modelos seleccionados para detectar de fraude en leche con una exactitud mínima de un 90%. También se demuestra que reducir dimensiones y aumentar datos es el mejor escenario, mejorando la eficiencia. Por ende, se ha podido apreciar que PCA y SMOTE son buenas técnicas para dichas tareas. Finalmente se implementaron los modelos seleccionados en una herramienta web, facilitando la comprobación de fraude de una forma sencilla y rápida.

Keywords: *Machine learning*, fraude alimentario, adulteración alimentaria

R-shiny application of the evolution of COVID-19 in Catalonia with Bayesian spatio-temporal analysis

Satorra P[1], Tebé C[1]
[1]Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Badalona, Spain;

Data analysis and visualisation is an essential tool for exploring and communicating findings in medical research, especially in epidemiological surveillance. We developed a shiny application to visualise the risk of COVID-19 cases, hospitalisations and vaccination by basic health areas (ABS) of Catalonia throughout the pandemic period (*https://brui.shinyapps.io/covidcat_evo/*). All data came from the open data catalogue of the Government of Catalonia. We considered the number of reported COVID-19 cases and hospitalisations and the cumulative number of COVID-19 fully vaccination (second or one-shot doses), by ABS and week. For the analysis, we used spatio-temporal small area estimation methods to borrow strength from neighbouring areas and time points. In particular, we used Bayesian hierarchical spatio-temporal models estimated with integrated nested Laplace approximation (INLA) using the R-INLA package. Different models were estimated for the spatial, temporal and spatio-temporal random effects and the best model was selected in terms of the model likelihood (DIC and WAIC). Also, models were adjusted for demographic and socio-economic characteristics of the ABS. Results presented high heterogeneity in cases and hospitalisation incidence between ABS and along the waves of the pandemic, meanwhile cumulative fully vaccination was similarly distributed by ABS. Urban areas were found to have a higher incidence of COVID-19 cases and hospitalisations than rural areas, while socio-economic deprivation of the area was associated with a higher incidence of hospitalisations. The estimated spatial, temporal and spatio-temporal relative risks (RR) are visualised in an R-shiny application, for COVID-19 cases, hospitalisations and fully vaccination. The evolution of the posterior mean of the spatio-temporal interaction RR is presented in an animated map showing the different outbreaks in ABS, together with the plot of the evolution of the posterior mean of the temporal RR, for each week. Finally, the posterior mean of the spatial RR is presented in a static map showing the excess/lack of risk of each ABS in all the pandemic period. This application provides a useful epidemiological surveillance tool that can help to understand the spatial and spatio-temporal trends of the COVID-19 pandemic and the COVID-19 vaccination campaign in Catalonia, produced in a user-friendly and fashionable way that is accessible to everyone.

Keywords: Disease mapping, COVID-19, shiny

Sesión de pósteres

9 de febrero, 10:30 a 12:00

Dealing with complex sampling designs on the estimation of the ROC curve and the area under it

Amaia Iparragirre[1], Irantzu Barrio[1,2], Inmaculada Arostegui[1,2]
[1]Department of Mathematics, University of the Basque Country (UPV/EHU);
[2]BCAM-Basque Center for Applied Mathematics

Complex survey data are becoming more and more relevant in a number of fields. This type of data is collected from a finite population of interest following some complex sampling design. Among other purposes, complex survey data are increasingly used to develop prediction models. In this work, we focus on logistic regression models for dichotomous response variables. Given the impact that these models have in daily practice, ensuring good discrimination ability is essential, which in the framework of logistic regression models is commonly measured by the receiver operating characteristic (ROC) curve and the area under this curve (AUC).

Due to the complex nature of the sampling process, the straightforward application of the most widely used statistical techniques, which are typically designed to be applied to simple random samples, is usually not appropriate in the context of complex survey data. Therefore, in this work we propose new estimators to estimate the ROC curve and the AUC of logistic regression models accounting for complex sampling designs. For this purpose, we propose to consider design-based estimators of sensitivity and specificity parameters [1]. An extensive simulation study, in which different sampling designs have been considered, has been carried out in order to evaluate the behavior of the proposed estimators under different scenarios. In this simulation study, the estimates obtained by the new design-based estimators, as well as the estimates obtained by traditional estimators, are compared to the true population parameters. The results suggest the use of the proposed estimators to estimate the ROC curve and AUC in the context of logistic regression models fitted to complex survey data in order to obtain unbiased estimates. These estimators are implemented in the R-package `wROC`, which is available in the following GitHub repository: https://github.com/aiparragirre/wROC.

Keywords: receiver operating characteristic curve, area under the curve, complex survey data.

References

1. Iparragirre A., Barrio I., Aramendi J., and Arostegui I. (2022). Estimation of cut-off points under complex-sampling design data. *SORT-Statistics and Operations Research Transactions, 46(1)*, 137-158.

El gran impacto de INLA y SPDE para la modelización estadística espacial desde la perspectiva bayesiana

Carmen Guarner-Giner[1]
[1]Departamento de Estadística e Investigación Operativa, Universitat de València

La estadística espacial está presente en numerosas áreas de investigación y ha cobrado un gran impulso en los últimos años. Comprender la incertidumbre en la modelización espacial de variables presentes en multitud de estudios resulta esencial para conocer cómo se distribuyen dichas variables. Actualmente, el proceso de modelizar la variabilidad espacial se está realizando mediante estadística bayesiana. Sin embargo, los enfoques convencionales basados en la simulación Markov Chain Monte Carlo (MCMC) suelen ser intensivos desde un punto de vista computacional. En este trabajo se abordan la aproximación de Laplace integrada encajada (INLA) junto con las ecuaciones diferenciales parciales estocásticas (SPDE) y el método de elementos finitos, para la inferencia y predicción en modelos latentes gaussianos espaciales, destacando su gran impacto debido a la precisión y velocidad de las aproximaciones posibles.

Keywords: INLA, SPDE, Modelización espacial

References

Analysis of Big Data with the integrated nested Laplace approximation

Héctor López-Gómez[1], Virgilio Gómez-Rubio[1], Carlos Gil-Bellosta[2]
[1]Department of Mathematics, School of Industrial Engineering-Albacete, Universidad de Castilla-La Mancha, Albacete (Spain); Circiter S.L., Madrid (Spain) [2]

The integrated nested Laplace approximation (INLA) [1] has been widely adopted for Bayesian inference as it can provide estimates of the posterior marginal distributions of the latent effects and hyperparameters of the model in a fraction of the time required by other estimation methods such as Markov chain Monte Carlo (MCMC). In this work we propose the use of INLA in a Big Data environment in which models can only be fit to a part of the data. We will follow previous work [2] to be able to fit the model of interest to different chunks of the data, following a MapReduce approach [3], to estimate the marginals of the latent effects and hyperparameters with INLA. Next, for each latent effect and hyperparameter the marginals will be combined in order to approximate the same marginals had the model been fit to the whole dataset. The application of this approach will be exemplified on a simulated dataset.

Keywords: Bayesian inference, Big Data, MapReduce

References

1. Rue, H., Martino, S. and Chopin, N. (2009), Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71: 319-392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

2. Huang, Z. and Gelman, A. (2005). Sampling for Bayesian computation with large datasets. Technical report. Available at SSRN: http://dx.doi.org/10.2139/ssrn.1010107

3. Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM 51(1), 107-113.

Block-wise missing omics data integration using kernel methods

Ignacio Pérez, Ferran Reverter[1], Esteban Vegas[1]

[1] Department of Genetics, Microbiology and Statistics, Section of Statistics, Faculty of Biology, University of Barcelona, Diagonal, 643, 08028, Barcelona, Spain

Recent advances in high throughput techniques and the efforts of international consortia have made available several data sets that include multiple omics data and clinical information for many samples. Kernel methods provide a general framework for omics data integration, such as Multiple Kernel Learning (MKL). In this project, we introduce a procedure that expands the application of MKL to the case of blockwise missing data. We have developed an R package incorporating all the functions needed to implement this methodology based on an article of clustering on multiple incomplete datasets via Collective Kernel Learning (CoKL). An incomplete breast cancer data set has been used as an example to investigate how the integration of missing data works when using a classification MKL algorithm such as Simple MKL with outstanding results. On the other hand, and following the same approach, it has been tested in a larger incomplete data set with combined sets of omics and exposome data when using a regression MKL algorithm such as SEMKL with lower performance results. Overall, in this project we have developed a procedure that integrates the completion of a block-wise missing data set using kernel methods and its subsequent use in MKL-type predictive algorithms. Further investigation is still required to improve the MKL models as different strategies when computing the kernel matrices and the adequate choice of the MKL parameters could benefit both the computational time and the performance of the model.

Keywords: Omics data integration, block-wise missing data, kernel methods

Compare variable selection methods using Random Forest and Boruta to have a tool for predicting multi-resistance.

Janire Gallejones[1], Urko Aguirre[1], Eloisa Urrechaga[2], Miren Arantza Burzako[3], Ana Gual de Torrella[4]

[1]Research Unit, U. Hospital Galdakao Usansolo, Biosistemak Institute for Health Services Research. Network for Research on Chronicity, Primary Care and Health Promotion (RICAPPS); [2]Core Laboratory, U. Hospital Galdakao Usansolo. BioBizkaia Health Research Institute; [3]Healthcare Management Unit, U. Hospital Galdakao Usansolo. BioBizkaia Health Research Institute; [4]Microbiology Service, U. Hospital Galdakao Usansolo.

Introduction and Objectives: Antibiotic resistance is a global problem associated with high morbidity, mortality and costs. Such resistance can be prevented by interventions aimed at reducing the over-prescription of antibiotics to hospitalized patients. Due to the presence of multiple interacting risk factors underscores the significance of selecting suitable variables. When considering the selection of factors, it is crucial to consider not only the individual effects, but also the manner in which they have interacted with other variables. In this context, Random Forest and Boruta methods were applied for the identification of risk factors predicting 10-day multidrug resistance for UTI.

Methods: After processing and cleaning the data, an initial database is available with 48 variables. To perform the variable selection in this case we have used two different techniques: Random Forest and Boruta. RF is a method that consists of combining predictor trees. Boruta is a method that uses RF as the underlying algorithm, the relative importance of each variable is calculated and if a variable is below the threshold it is eliminated.

Results: We recruited 9049 patients from public hospitals belonging to the Basque Health Service with a diagnosis of UTI in 2021. Our study showed that 36.5 % of the participants were men with a mean age of 71 years. A total of 292 multiple drug resistance were observed, of which 108 were in men. In both methods, the default value (0.01) was used as the threshold. The results of the selection of variables: with RF, 21 variables were chosen and with Boruta, 17 variables.

Conclusion: The selection models have their own criteria to inform which factors are the best. This results in different rankings depending on the algorithm used. In this case, the variable selection of both methods differs slightly; this helps to guide us on which variables are useful for predicting multi-resistance.

Keywords: multi-resistance, variable selection.

References

1. Beata Zielosko, Lakhmi C. Jain, Urszula Stańczyk. *Advances in Feature Selection for Data and Pattern Recognition*.

2. Hui Liu, Suishan Qiu, Minghao Chen, Jun Lyu, Guangchao Yu, Lianfang Xue. *A clinical prediction tool for extended-spectrum β lactamase producing Enterobacteriaceae urinary tract infection*

Assessing the cause of death in patients with heart failure through a Bayesian competing risks survival model

Jesús Gutiérrez-Botella[1], Carmen Armero[2], María Pata[3], Thomas Kneib[4], Francisco Gude-Sampedro[5]

[1]jesus.gutierrez.botella@rai.usc.es, Biostatech, Advice, Training and Innovation in Biostatistics SL; GRID-BDS, University of Santiago de Compostela

[2]carmen.armero@uv.es, Department of Statistics and OR, Universitat de València

[3]mariapata6@biostatech.com, Biostatech, Advice, Training and Innovation in Biostatistics SL

[4]tkneib@uni-goettingen.de, Georg-August-Universität Göttingen

[5]francisco.gude.sampedro@sergas.es, Epidemiology Department, Clinical University Hospital of Santiago de Compostela

Heart Failure (HF) occurs when the heart is unable to pump blood around the body properly. It usually happens because the heart has become too weak or stiff. Cardiac Resynchronization Therapy (CRT) is a procedure which consists of implanting a device in the heart's chambers to help it to work in a more efficient way. This therapy has been shown to improve the short-term prognosis of HF patients, but there are scarce data about its long-term benefits.

The objective of this work is to study the cardiovascular and non-cardiovascular death in HF patients who underwent CRT in relation to some clinical and demographic variables.

We used a Bayesian competing risks survival model for the two causes of death, cardiovascular and non-cardiovascular, with cause-specific hazard functions for each one. A variable selection procedure have been performed for each sub-model [1], and some additional analyses have been done with the selected model: a sensitivity analysis for prior distribution of model coefficients, a selection of different specifications for the baseline hazard function [2], and model checking with Deviance Information Criterion (DIC) and Conditional Predictive Ordinates (CPOs) [3]. Bayesian estimation was performed using MCMC methods with JAGS software. Posterior outputs of interest such as cause-specific hazard functions, overall survival function, cumulative incidence function or transition probabilities are obtained and discussed.

Keywords: Cardiac Resynchronization Therapy; Cardiovascular and non-cardiovascular death; Conditional Predictive Ordinate.

References

1. García-Donato, G., Cabras, S. & Castellanos, M.E. (2023) Model uncertainty quantification in Cox regression. *Biometrics*, 79, 1726–1736.

2. Lázaro, E, Armero, C, Alvares, D. Bayesian regularization for flexible baseline hazard functions in Cox survival models. *Biometrical Journal*. 2021; 63: 7–26.

3. Ming-Hui Chen, Mário de Castro, Miaomiao Ge, Yuanye Zhang. Bayesian Regression Models for Competing Risks. From: *Handbook of Survival Analysis*, 2013. CRC Press.

The COVID-19 Pandemic Impact on Clinical Markers: a Spatio-temporal Approach

Manuel Moreno [1,2], Maria Antonia Barceló [1], Marc Saez [1], Josep Vidal [2]
[1] Universitat de Girona
[2] Institut Català de la Salut

The COVID-19 pandemic deepened existing social inequalities. Difficulties in access to health care during the pandemic had effects on disease diagnosis and monitoring, leading to poor health status, especially in the vulnerable population. Available research on the topic focuses on ecologic studies, some with methodological shortcomings, implying uncertainty about the validity of the findings [1]. Epidemiological designs based on longitudinal data of individuals alongside appropriate methodological tools are needed to provide robust findings.

The main objectives are to characterize the socioeconomic inequalities during the pandemic and their effect on the health of the population, especially the most vulnerable, and to apply spatiotemporal modeling algorithms to routinely collected data.

For such purposes data is supplied by the Catalan Health Institute (ICS) via the Information System for Research in Primary Care (SIDIAP) from 2015 to 2022. It contains curated data on the population registered with the public health system. The data include demographic information, diagnoses, drug prescriptions, and consultations at primary care facilities [2]. A Bayesian approach using INLA's spatiotemporal modeling algorithm [3] is implemented.

Focusing on individuals with chronic conditions, it was observed that some clinic indicators worsened, before and during the pandemic, with differences between men and women (REGICOR index: +9·4% vs. +3·7% and Barthel index: −14·1% vs. −16·6%). Underscoring the importance of continuous care for individuals with chronic conditions.

Keywords: Spatio-temporal, COVID-19, REGICOR.

References

1. Barceló, M.A., Saez, M. (2021). Methodological limitations in studies assessing the effects of environmental and socioeconomic variables on the spread of COVID-19: a systematic review. *Environmental Sciences Europe*, 33(1).
2. Recalde, M., Rodríguez, C., Burn, E., Far, M., García, D., Carrere-Molina, J., Benítez, M., Moleras, A., Pistillo, A., Bolíbar, B., Aragón, M., & Duarte-Salles, T. (2022). Data Resource Profile: The information system for research in primary care (SIDIAP). *International Journal of Epidemiology*, 51(6).
3. Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2), 319–392.

GLOBAL SPECIES DISTRIBUTION MODELS FOR PENGUIN SPECIES

Marc Moreno[1], Alba Fuster[2], Miriam Gimeno[2], Francisco Ramirez[2], Marta Coll[2], Maria Bas[2]

[1] Universitat Autònoma de Barcelona (UAB);
[2] Institut de Ciències del Mar (ICM-CSIC), Barcelona;

Species distribution models (SDMs) have found widespread application in ecology, serving as valuable tools for estimating and predicting the spatial and temporal distribution of target species. A promising alternative to traditional SDMs is the Bayesian Additive Regression Trees (BART) classification tree method. BART is a non-parametric Bayesian regression approach that relies on a sum-of-trees model. In this study, we present the application of BART to analyze the distribution patterns of two penguin species, namely *Pygoscelis adeliae* and *Eudyptula minor*. Specifically, we model the probability of their presence using presence/pseudo-absence data. The model employed in this study is outlined below:

$$Y_i \sim Ber(\pi_i), \;\; i = 1, \dots, n,$$

$$\phi^{-1}(\pi_i) = \sum_{j}^{m} g_j\big(X, T_j, M_j\big),$$

where, Y_i is the response variable (presence/pseudo-absence of species) in each observation i associated with a Bernoulli probability distribution; π_i is the probability of presence linked to the predictor by a link function ϕ^{-1}; then, g_j is the $j-th$ tree of the form $g_j\big(X, T_j, M_j\big)$, where m is the total number of trees, X is a vector of multiple covariates, T_j represents a binary tree structure consisting of a set of interior decision rules and a set of terminal nodes, and $M_j = u_1, \dots, u_b$ denote a set of parameter values.

The results obtained from our analysis indicate that the potential future optimal habitat for *P. adeliae* may face a bigger reduction compared to *E. minor*. Furthermore, the insights derived from this study underscore the importance of implementing proactive conservation measures for penguin species conservation. In conclusion, the BART model has the potential to emerge as a valuable tool for projecting species optimal habitat suitability and guiding targeted conservation efforts.

Keywords: SDM, BART, penguins

References

Statistical modeling to adjust for time trends in platform trials utilising non-concurrent controls

Pavla Krotka[1], Martin Posch[1], Marta Bofill Roig[1]
[1]Center for Medical Data Science, Medical University of Vienna

Platform trials enhance drug development by offering increased flexibility and efficiency. They evaluate the efficacy of multiple treatment arms, with the added benefit of permitting treatment arms to enter the trial over time and to stop early based on interim data. Efficacy is usually assessed using a shared control arm. For arms entering later, the control data is divided into concurrent and non-concurrent controls (NCC), referring to control patients recruited while the given treatment arm is in the platform and before it enters, respectively. Including NCC can reduce the sample size and increase power, but also lead to bias in the effect estimates, if there are time trends.

For platform trials with continuous endpoints without interim analyses, a regression model has been proposed that utilizes NCC and adjusts for time trends by including the factor "period" as a fixed effect. Here, periods are defined as time intervals bounded by any treatment arm entering or leaving the platform. It was shown that this model leads to unbiased effect estimates and asymptotically controls the type I error (T1E) rate regardless of the time trend pattern, if the time trend affects all arms in the trial equally and is additive on the model scale [1]. However, if interim analyses are included, the definition of the factor periods becomes data dependent and the number of periods to adjust for depends on previous results. Furthermore, due to early stopping the sample sizes in different arms become outcome dependent, and therefore the effect estimates are no longer unbiased. This can affect the adjustment for time trends in the linear model, and the T1E rate might no longer be controlled.

In this work, we present two extensions of this model. First, we propose an alternative definition of the time covariate by dividing the trial into fixed-length data-independent calendar time intervals. Second, we propose alternative models to adjust for time trends. In particular, we consider: accounting for dependency between closer time intervals by adjusting for autocorrelated random effects; and employing spline regression to model time with a smooth polynomial function. We implement the proposals in the `NCC` R-package [2] and evaluate their performance in terms of the T1E rate and statistical power in a simulation study under a wide range of scenarios.

Keywords: Platform trials, Non-concurrent controls, Statistical modeling

References

1. Bofill Roig M., Krotka P., et al. (2023). On model-based time trend adjustments in platform trials with non-concurrent controls. *BMC Med. Res. Methodol.,* 22(1), 1-16.

2. Krotka P., et al. (2023) NCC: An R-package for analysis and simulation of platform trials with non-concurrent controls. *SoftwareX,* 23, 101437.

Análisis estadístico de la supervivencia de pacientes con implante Cardioband

Sandra González Martín[1 2], Ana Pardo Sanz[1], José Luis Zamorano Gómez[1]
[1]Servicio de Cardiología, Hospital Universitario Ramón y Cajal;
[2]Facultad Estudios Estadísticos, Universidad Complutense de Madrid

Las enfermedades cardiovasculares son una de las principales causas de discapacidad y mortalidad, incluyendo la insuficiencia tricúspide, la cual puede tratarse con válvulas Cardioband. Dado que los datos estaban censurados, se aplicaron técnicas como Kaplan-Meier, regresión de Cox, estimador Nelson-Aalen y Random Survival Forest para analizar el tiempo de supervivencia hasta la muerte. El objetivo principal del estudio fue evaluar la efectividad de Cardioband como alternativa a la cirugía para tratar la insuficiencia tricúspide (IT), y los objetivos secundarios incluyeron la revisión de técnicas de análisis de supervivencia, depuración de datos, selección de variables, evaluación de modelos y comparación de supervivencia entre grupos, para ello se emplearán tanto técnicas clásicas como de Machine Learning para lograr estos objetivos. El reclutamiento de los pacientes con IT se llevó a cabo entre 2015 y 2023 en el Hospital Ramón y Cajal y se realizó un seguimiento a los 6 meses y al año de la intervención, donde 37 pacientes fueron tratados con Cardioband, 5 Tric-Valve y 42 sin intervención. Como resultado se identificaron covariables relevantes para la supervivencia de los pacientes con IT, destacando el sexo, altura, ingresos, antecedentes de infarto de miocardio y enfermedad pulmonar obstructiva crónica. En los pacientes Cardioband, el sexo, los ingresos, tasa de filtrado glomerular post y algunas variables del anillo fueron influyentes. Por lo que, con todo esto, se concluyó una mayor supervivencia en el sexo femenino en comparación con el masculino y en pacientes sin antecedentes de riesgo en ambos grupos analizados.

Keywords: Enfermedades cardiovasculares, insuficiencia tricúspide, análisis de supervivencia

References

1. Pardo Sanz A et al. (2022). Long-term outcomes of percutaneous tricuspid annuloplasty with Cardioband device. Eur Heart J Cardiovasc Imaging, 23(7)

2. Pickett KL et al. (2021). Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. BMC Medical Research Methodology, 21(216).

Listado de participantes

En orden alfabético:

№	Nombre	Afiliación	e-mail
1.	Agnès Pérez Millan	Hospital Clínic de Barcelona	agperez@recerca.clinic.cat
2.	Alan Domínguez	ISGlobal	alan.dominguez@isglobal.org
3.	Alba Fuster Alonso	Instituto de Ciencias del Mar (ICM) - CSIC	afuster@icm.csic.es
4.	Álvaro Díaz Pérez	Universidad Carlos III de Madrid	100493165@alumnos.uc3m.es
5.	Amaia Iparragirre	Universidad del País Vasco (UPV/EHU)	amaia.iparragirre@ehu.eus
6.	Andrea Toloba	Universitat Politècnica de Catalunya	andrea.toloba@upc.edu
7.	Antia Enriquez Yurrebaso	UPV/EHU, BCAM	antia.e.y@gmail.com
8.	Armand Gonzalez	BBRC	agonzalez@barcelonabeta.org
9.	Arnau Garcia Fernandez	UPC	arnau.garcia.fernandez@estudiantat.upc.edu
10.	Belén García Martínez		garciabelencita2001@gmail.com
11.	Blanca Paniello Castillo	ISGlobal	blanca.paniello@isglobal.org
12.	Blanca Rius Sansalvador	IDIBELL	brius@idibell.cat
13.	Blanca Rodríguez-Fernández	BarcelonaBeta Brain Research Center	blancarf1995@gmail.com

№	Nombre	Afiliación	e-mail
14.	Carlos Javier Peña de los Santos	IIS INCLIVA	cpena@incliva.es
15.	Carmen Guarner Giner	Universitat de València	guargi@alumni.uv.es
16.	Claudia Armengol Arcas		carmengola@uoc.edu
17.	Claudia Tielas Sáez		claudiatielassaez@hotmail.es
18.	Cristóbal Manuel Rodríguez Leal	Universidad Complutense de Madrid	cristo07@ucm.es
19.	Daniel Fernández Martínez	UPC	daniel.fernandez.martinez@upc.edu
20.	David Moriña Soler	UB	dmorina@ub.edu
21.	Dorota Mlynarczyk	Universitat Autònoma de Barcelona	dorotaanna.mlynarczyk@uab.cat
22.	Edmond Géraud	Universitat Politècnica de Valencia	egeraud@doctor.upv.es
23.	Esther García Lerma	IDIBELL	egarcial@idibell.cat
24.	Esther Tercero Atencia	Universitat de Valencia	esther14tercero@gmail.com
25.	Garazi Retegui Goñi	Universidad Pública de Navarra (UPNA)	garazi.retegui@unavarra.es
26.	Gokce Deliorman	Universidad Complutense de Madrid	gdeliorm@ucm.es
27.	Gonzalo Aparicio Rodríguez	Universidad Complutense de Madrid	gonzaapa@ucm.es
28.	Guillermo Villacampa Javierre	Instituto de Investigación Vall d'Hebron	villacampa.u@gmail.com
29.	Héctor López Gómez	Universidad de Castilla-La Mancha	hector.lopez@uclm.es
30.	Ignacio Pérez Blasco	UOC	nachopb6@hotmail.com

№	Nombre	Afiliación	e-mail
31.	Irene Vañó Segarra	Institut d'Investigació Pere i Virgili	irenevanyosegarra@gmail.com
32.	Ivan Vergara Fernández	Universitat Oberta de Catalunya	ivergaraf@uoc.edu
33.	Janire Gallejones Eskubi	Biosistemak-Hospital Galdakao	gallejonesjanire@gmail.com
34.	Jesús Gutiérrez Botella	Universidade de Santiago de Compostela	jesusgutierrezbotella@gmail.com
35.	Jesús Vicente Giménez de los Aguirre	Universidad de Granada	intertato_@hotmail.com
36.	Joana Llauradó Pont	ISGlobal	joana.llaurado@isglobal.org
37.	João Pedro Carmezim Correia	Germans Trias i Pujol Research Institute and Hospital (IGTP)	jcarmezim@igtp.cat
38.	Jon Aritz Panera Carracedo	Institut Català d'Oncologia ICO	jonaritzp@gmail.com
39.	Jordi Cortes Martinez	UPC	jordi.cortes-martinez@upc.edu
40.	Jorge Castillo-Mateo	Universidad de Zaragoza	jorgecm@unizar.es
41.	Jorge Mestre Tomas	Universidad de Valencia (UV)	jorgemartineztomas@gmail.com
42.	Lander Rodriguez Idiazabal	BCAM - Basque Center for Applied Mathematics	lrodriguez@bcamath.org
43.	Lasai Barreñada	KU Leuven	lasai.barrenadataleb@kuleuven.be
44.	Leire Garmendia Bergés	BCAM - Basque Center for Applied Mathematics	lgarmendia@bcamath.org
45.	Leire Unamuno Celayeta	UPC-UB	leire.u1998@gmail.com
46.	Lucía Yubero Fernández	UAM	luciayufe@gmail.com
47.	Mahnaz Shekari	Barcelonabeta Brain Research Center	mshekari@barcelonabeta.org

№	Nombre	Afiliación	e-mail
48.	Maialen Otamendi Garitano	Biosistemak Institute for Health Systems Research	otamendimaialen2@gmail.com
49.	Manuel Alejandro Moreno Vasquez	Universitat de Girona	u1982406@campus.udg.edu
50.	Marc Moreno Candón	Universitat Autònoma de Barcelona (UAB)	marc.gaara@gmail.com
51.	Marcos Quijal	ISGlobal	marcos.quijal@isglobal.org
52.	María Gascón Pérez	Hospital Galdakao-Usansolo. Biosistemak.	mgascon@kronikgune.org
53.	Maria Grazia Pennino	IEO-CSIC	grazia.pennino@ieo.csic.es
54.	María del Pilar González Barquero	Universidad Carlos III de Madrid	100469005@alumnos.uc3m.es
55.	Mario Figueira Pereira	Universidad de Valencia	Mario.Figueira@uv.es
56.	Marta Bofill Roig	Medical University of Vienna	marta.bofillroig@meduniwien.ac.at
57.	Marta Lafuente Sánchez	UOC	marta.lasa09@gmail.com
58.	Matilde Francisco	UPC	matilde.martins.da.palma@upc.edu
59.	Melina Alves-Sampaio	UOC	malves-sampaiob@uoc.edu
60.	Miguel Ángel Beltrán Sánchez	Universitat de València	mianbel@alumni.uv.es
61.	Miguel Ángel López González	Universitat Oberta de Catalunya	malopezgonzalez@uoc.edu
62.	Naiara Santos	IDIBELL	nsantos@idibell.cat
63.	Natàlia Pallarès Fontanet	IGTP	npallares@igtp.cat
64.	Pablo Castillo Jiménez	Germans Trias i Pujol Research Institute and Hospital (IGTP)	pabloandres.castillo@autonoma.cat
65.	Patricia Genius Serra	BarcelonaBeta Brain Research Center (BBRC)	pgenius@barcelonabeta.org

№	Nombre	Afiliación	e-mail
66.	Pau Satorra Herbera	Germans Trias i Pujol Research Institute and Hospital (IGTP)	psatorra@igtp.cat
67.	Paula Osorio Recordà		paulaosoriorecorda@gmail.com
68.	Pavel Hernández Amaro	Universidad Carlos III de Madrid	pahernan@est-econ.uc3m.es
69.	Pavla Krotka	Medical University of Vienna	pavla.krotka@meduniwien.ac.at
70.	Pedro Miranda-Afonso	Erasmus MC	p.mirandaafonso@erasmusmc.nl
71.	Sandra González Martín	Universidad Complutense de Madrid / Hospital Ramón y Cajal	sandragmdm@gmail.com
72.	Sofía Aguilar Lacasaña	ISGlobal	sofia.aguilar@isglobal.org
73.	Virginia Munoa	Idibell	vmunoa@idibell.cat
74.	Xavier Garcia Cusco	Departament de Salut	xaviergarcus@gmail.com