XX CONFERENCIA ESPAÑOLA DE BIOESTADÍSTICA

CEB SEPTIEMBRE 2025 ELCHE

Organizan

Colaboran

ISBN: 978-84-18177-95-8

Índice general

1 Presentación

Este libro recoge los resúmenes de los trabajos presentados durante la **XX Conferencia Española de Bioestadística (CEB 2025)**, celebrada en Elche en septiembre de 2025.

La conferencia está organizada por el *Instituto Centro de Investigación Operativa* y la *Unidad Mixta de Investigación UMH-FISABIO (StatSalut)*, respondiendo a la petición de la *Sociedad Española de Bioestadística (SEB)*. El evento tiene como objetivo la difusión de los resultados obtenidos en el desarrollo y aplicación de métodos estadísticos y matemáticos en diversos ámbitos de las ciencias de la vida.

La CEB 2025 constituye un espacio de encuentro para investigadores, profesionales y estudiantes que desarrollan su labor en campos como la Biología, Medicina, Psicología, Farmacología, Ecología, Agricultura y otras disciplinas afines. Desde el Comité Organizador se ha puesto especial empeño en ofrecer un congreso de alto nivel científico, fomentando el intercambio de ideas, experiencias y colaboraciones futuras.

Este libro está estructurado en diferentes bloques temáticos, que incluyen conferencias plenarias, sesiones orales y sesiones de pósteres. Confiamos en que esta recopilación contribuya a consolidar el papel de la bioestadística como disciplina clave en el progreso del conocimiento científico y en la mejora de la salud y el bienestar social.

2 Comité Organizador

Eva María Navarrete Muñoz – Presidenta

Vocales:

- Laura Aixalà Perelló
- Xavier Barber Vallés
- Empar Casaña Escriche
- Irene Campos Sánchez
- Jonatan A. González
- Irene López Mengual
- María Asunción Martínez Mayoral
- Javier Morales Scouellamos
- Rocío Muñoz Sánchez
- María Paula Noce
- Desirée Valera Gran
- Begoña Gómez Pujante
- Coral Blanquer Piña

*Todas las personas integrantes del comité pertenecen a la **Universidad Miguel Hernández de Elche**.*

3 Comité Científico

Anabel Forte – Presidenta
Universitat de València

Vocales:

- Jose Antonio Quesada Rico
  Universidad Miguel Hernández
- Xavier Barber Vallés
  Universidad Miguel Hernández
- Eva María Navarrete Muñoz
  Universidad Miguel Hernández
- María Xosé Rodríguez Álvarez
  Universidad de Vigo
- Martí Casals
  Universitat de Vic
- Klaus Langohr
  Universitat Politècnica de Catalunya
- Marisa Estarlich Estarlich
  Universitat de València
- Juan Aparicio Baeza
  Universidad Miguel Hernández
- Cristina Lidón Moyano
  Universidad Internacional de Cataluña
- Alicia Quirós
  Universidad de León
- Michela Cameletti
  Università degli studi di Bergamo

4 Conferencias Plenarias

1. Geospatial Data Science for Public Health Surveillance

Paula Moraga

(*King Abdullah University of Science and Technology (KAUST)*)

Geospatial health data are essential to inform public health and policy. These data can be used to understand geographic and temporal patterns, identify risk factors, measure inequalities, and quickly detect outbreaks. In this talk, I will give an overview of statistical methods and computational tools for geospatial data analysis and health surveillance. Using dengue surveillance in Brazil as a case study, I will discuss data biases and availability issues in surveillance systems. I will also present modeling advancements to integrate complex health, climate, and digital data from different sources and resolutions to predict disease risk and detect outbreaks. Finally, I will discuss the importance of effective communication and dissemination to inform policymaking and improve global population health.

Keywords: dengue, disease surveillance, forecasting, nowcasting, spatial modeling

2. Statistical Consultancy and the Use of Plausible Intelligence

Mark J Brewer

(*Biomathematics & Statistics Scotland, BioSS Office, The James Hutton Institute*)

Does statistical consultancy get the credit it deserves? Is it seen as the poor relation of statistical methodological development? Is it just "easy", applying standard statistical techniques, simply and repeatedly? Can we now replace statistical consultants with Artificial Intelligence?

With examples gathered over many years, Mark will argue that statistical consultancy is commonly a vital part of the scientific process, and itself requires training, skills and experience above and beyond mathematical capability, programming skills and technical competence.

Keywords: statistical consultancy; professional skills vs technical skills

3. Regression in Contemporary Biostatistics: Evolution and Applications

Susana Pérez-Álvarez

(*Sociedad Española de Bioestadística*)

Regression is one of the most fundamental and versatile tools in statistics, and its development mirrors the growth of biostatistics itself. In this talk, I reflect on the evolution of regression methods through the lens of my own professional experience in **public health, diagnostic development, and clinical research.**

With the rise of genomics, I became increasingly aware of the importance of **multiple testing corrections**, from the conservative Bonferroni adjustment to the more nuanced False Discovery Rate (FDR), both critical in high-dimensional data contexts.

I then revisit **logistic regression**, which I first encountered in the context of public health and epidemiology. Its ability to estimate risks made it indispensable in early disease surveillance efforts, including applications in oncology and in HIV monitoring. A step beyond in this area, survival analysis also marked a turning point in my work, especially through the Kaplan–Meier estimator and Cox models, which extended regression into time-to-event contexts.

In the realm of in vitro diagnostics, I applied **Deming** and **Passing–Bablok regression** methods, now the standard for method comparison studies, meeting both scientific and regulatory demands.

Finally, my current focus includes **causal inference**, particularly in the context of clinical trials, where **propensity scores** and the **estimand framework** have expanded our ability to interpret treatment effects in both randomized and **real-world evidence (RWE)** settings.

By tracing this path, I aim to show how the evolution of regression methods has shaped both my understanding of **scientific rigor** and the role of biostatisticians in advancing science and healthcare through continuous **methodological innovation.**

Keywords: regression, real-world evidence (RWE), in-vitro diagnostics (IVD), biostatistics innovation.

5 Sesiones - Comunicaciones

4. Competing risks: cumulative incidence vs. Kaplan-Meier

Carmen Carazo-Díaz
(*Universidad Católica de Murcia (UCAM)*)

David Prieto-Merino
(*Universidad de Alcalá (UAH)*)

Luis Prieto Valiente
(*Universidad Católica de Murcia (UCAM)*)

A well-known limitation of Kaplan-Meier is its implicit assumption that censoring is independent of the event of interest. When competing risks exist (e.g., different, mutually exclusive causes of failure), treating the occurrence of alternative causes as censoring can lead to overestimating the probability of the cause of interest. In these cases, the appropriate nonparametric alternative is the Aalen-Johansen estimator, which generalizes Kaplan-Meier to multi-state or competing-risk data. The Aalen-Johansen estimator directly calculates the cumulative incidence function for each event type, avoiding the K-M assumption. Studies have empirically shown that using Kaplan-Meier to estimate incidence in the presence of competing risks can yield markedly upward biased values, whereas competing-risk-specific methods (such as Aalen-Johansen) produce accurate results.

A more intuitive way to assess the potential dependence between cancellation and event risk is to compare the risk in the cancellation context with the event risk in the general context. Calling K the former divided by the latter, this shows that $K = 1$ corresponds to K-M, while $K = 0$ corresponds to Aalen-Johansen. Other values of K, whether integer or fractional, allow the researcher to estimate the cumulative incidence more precisely.

Keywords: Kaplan-Meier, Aalen-Johansen, cumulative incidence function.

5. Testing without metric parameters. They are not an alternative to parametric parameters but rather a possible complement

Carmen Carazo-Díaz

(*Universidad Católica de Murcia (UCAM)*)

David Prieto-Merino

(*Universidad de Alcalá (UAH)*)

Luis Prieto Valiente

(*Universidad Católica de Murcia (UCAM)*)

It's common to read in textbooks that the Student t test (and ANOVA) is only valid when the variable involved is normally distributed in the sampled populations. Therefore, we should only use them after applying the normality test and seeing that there is no evidence against the normality hypothesis. If there is, we should use a nonparametric test.

It is true that the distribution of the Student t and Snedecor F statistics in random sampling follows the distribution specified by the corresponding distribution functions only if the variable in question is normally distributed in the sampled populations. However, other, equally true, facts must be taken into account, which require us to qualify this rule.

a) When the sample size exceeds, say, 30, the Central Limit Theorem guarantees a very approximately normal distribution of the statistic under study.

b) In no case can "normality tests"guarantee that the variable is normally distributed in populations.

c) In small samples (where the Central Limit Theorem is not applicable), the statistical power of the normality test is limited and very rarely provides strong evidence against normality.

d) The distribution of the statistics involved is remarkably robust to the non-normality of the original distribution.

e) Recall that the p-value of the test tells us how likely it is to find samples with values corresponding to statistics so far from the parameters proposed by the null hypothesis, or even farther, if it is true.

Parametric and nonparametric tests test different null hypotheses. Although the two hypotheses overlap to a large extent, they are not exactly the same. The nonparametric test provides the exact probability relative to the null hypothesis it poses. The parametric test calculates the exact probability relative to the hypothesis it poses, if the normality condition is met. If this normality does not exist (something that can never be known with certainty), the probability provided by the parametric test is an estimate (with a certain degree of unquantifiable imprecision) of the desired probability.

Each test can be applied, and the degree of evidence against its null hypothesis that each provides can be assessed, always keeping in mind that the evidence provided by the p-value against its null hypothesis varies gradually; it is not an "all or nothing" matter, separated by a threshold value such as 0.05 or 0.01, or any other agreement reached at any given time.

A careful analysis of the information provided by each type of test and the conclusions supported by the information provided by the test show that both types of tests are not mutually exclusive. In some cases, they are complementary. And in no case do they contain contradictory allusions. Therefore, it is necessary to change the criterion outlined in the first paragraph.

Keywords: nonparametric test, normality test, statistical significance.

6. Progression of hospitalized patients with severe influenza: Bayesian estimation of conditional probabilities in directed acyclic graphs

Lesly Acosta

(*Universitat Politècnica de Catalunya (UPC)*)

Carmen Armero

(*Universitat de València*)

This study presents a general Bayesian framework for estimating joint, conditional, and marginal probabilities associated to random events in directed acyclic graphs (DAGs) based on the multinomial-Dirichlet inferential process and simulation tools.

Influenza is a well-known seasonal illness in our society. Most people affected by this disease recover in one to two weeks without complications. However, it can be a particularly serious illness in older adults and individuals with chronic conditions. Our study focuses on the progression of 1306 hospitalized patients with severe influenza. The data for the analysis, included in the Primary Care Influenza Surveillance System of Catalonia (PIDIRAC), were obtained from a retrospective cohort study conducted between October 1, 2017, and May 22, 2018, across 14 participating hospitals.

All patients with severe influenza who came to the hospital were initially assessed by a physician who, depending on the patient's condition, recommended admission to either an intensive care unit (ICU) or a specific hospital ward. Some patients who were initially admitted to a ward were later transferred to the ICU, moved to a long-term care facility, died, or were discharged home after recovery. These three outcomes are considered absorbing states. Patients in the ICU could either die or, if their condition improved, be transferred to a second type of hospital ward, from which they could eventually transition to one of the previously mentioned absorbing states.

Transition probabilities between direct and sequential states were estimated using a multinomial-Dirichlet inferential approach. General probabilities of transitioning from admission to the absorbing states, as well as relevant inverse probabilities, were obtained through simulation techniques based on the posterior distribution of the resulting Dirichlet probabilities.

Our work provides valuable insights into the progression of severe hospitalized influenza cases and can support healthcare decision-making and hospital resource planning during the seasonal influenza outbreaks that we experience each year with the arrival of colder weather.

Keywords: Influenza; Initial, transient and absorbing states; Markovian probabilities; Multinomial-Dirichlet inferential process.

7. Enhancing disease risk estimation reliability in highly imbalanced datasets

A. Adin,

A. Aramburu,

M.D. Ugarte

(*Public University of Navarre, Spain*)

Disease mapping is a key area of spatial epidemiology that focuses on aggregated count data from non-overlapping areal units, aiming to uncover underlying spatial and spatio-temporal patterns of adverse health events. In this context, Bayesian hierarchical models incorporating spatially structured random effects are widely used to provide more precise and stable estimates of disease risks or rates by borrowing strength from neighboring regions and reducing the variability of crude estimates, particularly in areas with limited or sparse data.

However, estimating disease risk in highly imbalanced datasets presents significant challenges, particularly when dealing with rare diseases or fine-grained spatial data. This study focuses on enhancing the reliability of such estimates by exploring different techniques such as adaptive spatial aggregation based on either observed cases or the relative precision of posterior estimates. Our approach aims to produce robust and reliable risk maps.

In this work we present preliminary results from cancer mortality data analysis in Navarre and the Basque Country, along with a simulation study, designed to assess the effectiveness of our proposed methodology.

Keywords: Bayesian inference, Smoothing methods, Spatial statistics.

8. Variable Contribution in Survival Analysis

Anna Aguilera
(*Vall d'Hebron Institut of Oncology (VHIO)*)

Victor Navarro
(*Vall d'Hebron Institut of Oncology (VHIO)*)

Guillermo Villacampa (*Statistics Unit (VHIO) and Data Science Unit, SOLTI Cancer Group*)

Introduction: Despite the advancements in machine learning, model interpretability remains challenging when using complex databases. Understanding the contribution of individual variables in the model prediction is crucial for interpreting and validating machine learning algorithms both for binary and survival outcomes.

Methods: This study evaluates different classification and survival methods. For binary outcomes,four classification methods were evaluated: i) logistic regression (LR), ii) random forest classifier, iii)suport vector machine classifier (SVM), iv) extreme gradient boosting (XGBoost). For survival analysisthree methods were evaluated: i) random survival forests (RSF), ii) Gradient Boosting Survival (GBS),and iii) Cox proportional hazards. RSF and GBS are nonparametric, ensemble methods that handleright-censored data using survival-adapted splitting rules and Nelson-Aalen estimators. SVM identifiesthe optimal hyperplane to specify classes. For these machine learning models, SHAP values will be usedto quantify covariates contribution. On the other hand, since Cox proportional hazards model does notestimate covariate contributions, we use a permutation-based approach that evaluates variable impact bysystematically removing them and measuring changes in performance. Performance is assessed usingthe rsq() function from the survMisc package, which provides an approximation of the coefficient ofdetermination, indicating the proportion of variability in survival explained by the model. LR estimatesodds ratios, identifying key contributors and quantifying their effect.

Results: A database containing clinicopathological and genomic information from >2,000 patientswith HER2-positive breast cancer was used to evaluate the contribution of each component in predicting:(i) pathological complete response (pCR, a binary endpoint) and (ii) disease-free survival (a survivalendpoint). Preliminary results show how genomic information has a greater contribution to predict pCR,while clinicopathological information plays a more prominent role in predicting survival outcomes. Additionally,a simulation-based analysis will be conducted by replicating different scenarios in oncologyresearch. This data will allow us to assess the consistency of different methods in quantifying the contributionof each covariate. Results from these simulations, together with visualization methods to explainthis information, will be presented at the conference.

Conclusions: This study can help establish a framework for better interpreting machine learning algorithms and comparing them to classical statistical methods using homogeneous measurements. Improving the interpretation of each covariate's contribution is crucial for developing and implementing precision medicine.

Keywords: Survival Analysis, Machine Learning, SHAP values.

9. Prognostic factors in colorectal cancer: A ten-year analysis based on gender and tumour location using competing risk models

Desirée Martín-García

Marilina García-Aranda

Janire Gallejones-Eskubi

Eloísa Urrechaga

Maximino Redondo

José María Quintana

Urko Aguirre

(*IBIMA Plataforma BIONAND, Hospital Universitario Costa del Sol, Osakidetza, Universidad de Málaga, Biocruces Bizkaia, RICAPPS*)

In the 21st century, cancer remains a major global challenge across social, public health, and economic spheres. It is responsible for roughly one in six deaths worldwide (16.8 %) and one in four deaths (22.8 %) linked to non-communicable diseases. Among cancers, colorectal cancer (CRC), comprising tumors located in the colon, rectum, and anus, stands out as the third most frequently diagnosed and the second leading cause of cancer-related deaths globally. In 2022, CRC claimed over 665,000 lives, underscoring its significance as a public health concern. However, data on specific long-term outcomes are limited.

This study examined ten-year predictive factors and gender disparities in colorectal cancer (CRC) patients in the Basque Country, employing an integrative multivariate methodology. Survival analysis, often called time-to-event analysis, measures a population's longevity. This methodology is often right censored, leaving some patients' time to event unreported due to follow-up interruptions or time restrictions. Clinical survival data analysis sometimes includes competing risks (CRs) where a person dies from many reasons.

Competing risk models address several shortcomings in traditional survival analysis. In a cancer trial, these models account for unrelated deaths that hinder the desired outcome. These models evaluate risk unbiasedly by integrating the target event hazard with the overall survival probability using the cumulative incidence function (CIF). Unlike Kaplan-Meier methods, which exaggerate event probabilities by censoring competing hazards.

Thus, in our study cumulative incidence curves were constructed, and both univariate and multivariate Fine-Gray competing risk regression models were formulated to ascertain characteristics linked to an elevated risk of mortality. Competing-risk models provide reliable risk classification, guide treatment procedures, and elucidate prognostic markers in cancers such as colorectal cancer. By separating cancer-specific outcomes, these models mitigate bias and facilitate evidence-based decision-making.

We identified key baseline risk factors for 10-year mortality in CRC patients. Advanced age, alcohol consumption, pathological haemoglobin levels, poor pre-surgical health status (ASA IV), and the absence of adjuvant chemotherapy were common risk factors for both CRC-specific and non-cancer-related mortality. Beyond these shared factors, CRC-specific mortality was strongly associated with advanced TNM stages, poor surgical outcomes (R1-R2), pathological NLR levels in rectal tumours, and postoperative infectious complications in colon tumours. Male sex and a family history of neoplasms increased CRC-specific mortality in men, while anxiety and depression were relevant risk factors in women. For non-cancer-related mortality, additional risk factors included unemployment, comorbidities, and being employed in women.

Our study highlights key predictors of long-term mortality in CRC patients, emphasizing common risk factors for both cancer-related and non-cancer-related deaths, as well as tumour location- and sex-specific differences, supporting the need for personalized long-term follow-up strategies in CRC survivors. Competing-risk analysis is essential for accurate survival estimation, especially in conditions characterised by significant competing mortality. Its implementation averts erroneous results and augments the validity of clinical research.

Keywords: colorectal cancer, competing risk models, mortality.

10. Environmental Suitability Assessment for Aquaculture Site Selection

Laura Aixalà-Perelló
(*Miguel Hernández University of Elche*)

Aitor Forcada
(*University of Alicante*)

Linda Fourdain
(*University of Alicante*)

Irene Lopez-Mengual
(*Miguel Hernández University of Elche*)

Antonio López-Quílez
(*University of Valencia*)

Pablo Sánchez-Jerez
(*University of Alicante*)

Kilian Toledo-Guedes
(*University of Alicante*)

Xavier Barber
(*Miguel Hernández University of Elche*)

The sustainable development of aquaculture in the Mediterranean requires a structured approach to identifying optimal farming locations based on environmental stability and long-term viability. This study explores a comprehensive assessment framework that integrates key oceanographic variables—such as sea surface temperature, salinity, pH, and current velocity—to evaluate site suitability while minimizing exposure to extreme environmental conditions. Extreme events, such as heatwaves, sudden temperature fluctuations, and severe storms, can significantly impact fish survival and farm productivity, making it essential to identify areas with stable environmental conditions that are less prone to such risks.

A major focus of this research is the development of a quantitative methodology that synthesizes multiple environmental indicators into a unified suitability assessment. Various statistical techniques are examined to establish the relationship between these variables and fish survival, identifying areas with both favorable conditions for aquaculture and lower vulnerability to extreme environmental events. By combining these environmental factors, the study aims to construct a robust index that can be mapped across marine regions, offering a detailed evaluation of aquaculture potential while accounting for the risks posed by extreme fluctuations in environmental conditions.

The proposed approach provides a practical tool for optimizing site selection, supporting decision-making processes that enhance sustainability and resource management. By refining environmental suitability assessments and considering the impact of extreme events, this research contributes to improving aquaculture planning, reducing operational risks, and ensuring the long-term success of fish farming operations in the Mediterranean.

Keywords: Aquaculture site selection, Spatial analysis, Extreme events.

11. Generalized Confidence Intervals for Covariate-Adjusted Overlap Measures

A. Alcaraz-Penãlba

A.M. Franco-Pereira

M.C. Pardo

(*Complutense University of Madrid, Instituto de Matemática Interdisciplinar, University Hospital 12 de Octubre*)

The overlap coefficient (OVL) measures the similarity between two probability density functions and is widely used in medical diagnostics to evaluate biomarker effectiveness. This study compares parametric methods for constructing confidence intervals for the OVL, including a generalized pivotal quantities approach, and introduces an algorithm to calculate generalized confidence intervals for covariate-adjusted OVL measures.

Keywords: Overlap coefficient, Generalized pivotal quantity, Covariates.

12. Identifying critical exposure time windows using Bayesian hierarchical models: The effect of NO_2 on foetal cranial size

Ruben Amoros

Mari-Carmen Enguix

Carmen Iñiguez

(*Universitat de València*)

Distributed Lag Nonlinear Models (DLNMs) are a class of models useful in evaluating the relationship between time-varying exposures and an outcome of interest. In recent years, they have been widely used to identify critical periods of fetal growth vulnerability to air pollution. However, the DLNM framework relies on specifying a bi-dimensional function space, called a cross-basis, which heavily relies on domain-specific expertise and prior literature, which is still scarce in this context. Furthermore, by design, it cannot adequately account for the uncertainty in the assignment of each window-specific exposure level, potentially leading to overly precise parameter estimates with unrealistically narrow confidence intervals.

In this work, we propose a Bayesian hierarchical model to identify the exposure window most related to the outcome. We demonstrate its application using NO_2 exposure during pregnancy and its impact on biparietal diameter at birth. Additionally, we explore different approaches to exposure temporal modelling, such as cyclic splines and latent autoregressive structures, and their impact on critical window selection.

Keywords: Critical exposure window; Bayesian hierarchical model.

13. Effect of Physical Activity and Functional Capacity on the Evolution of HRQoL in Patients with COPD

I. Arostegui

(*University of the Basque Country UPV/EHU, Basque Center for Applied Mathematics - BCAM*)

J. Najera-Zuloaga

(*University of the Basque Country UPV/EHU*)

I. Barrio

(*University of the Basque Country UPV/EHU, Basque Center for Applied Mathematics - BCAM*)

J.M. Quintana

(*Galdakao-Usansolo Hospital*)

C. Esteban

(*Galdakao-Usansolo Hospital*)

There is no doubt that physical activity (PA), as a behavioural habit, is beneficial for health. Functional capacity (FC) refers to the ability to achieve a physical objective. While these two concepts are related, they represent different constructs. Considering their interaction could play a crucial role among people with chronic diseases, particularly those with chronic obstructive pulmonary disease (COPD). In this work, we propose a classification of COPD patients based on the relationship between FC (whatthey can do) and PA (what they actually do), and we analyse its impact on health-related quality of life(HRQoL).

This cohort study includes 512 patients with COPD with a 3-year follow-up. PA was measured as the average daily step count using an accelerometer, while FC was assessed using the 6-minute walking test. HRQoL was evaluated with the Saint-George's Respiratory Questionnaire (SGRQ). Sociodemographic and clinical variables were recorded and included as potential covariates. Optimal cut-off points for predicting mortality were identified for both FC (can) and PA (do) continuous measurements, leadingto a four-quadrant structure: CanDo, CanDon't, Can'tDo and Can'tDon't. We modelled the baselinestatus and the evolution of HRQoL across these four groups using longitudinal beta-binomial mixedeffectsregression models, incorporating individual-level random intercepts and slopes. Other significantcovariates were also considered in a multivariate framework.

The models revealed differences in baseline HRQoL, which was influenced by both FC and PA. Significant differences were observed not only in overall HRQoL, but also in the symptoms, activity and impact domains. The differences between those who can and those who can't persisted after adjusting for significant covariates, showing that baseline HRQoL is significantly affected by FC. Regarding theevolution of HRQoL, adjusted by other significant longitudinal factors, all quadrants showed significantimprovement over time, with no differences between them, except for the CanDon't group whichremained stable. These results indicate that improvement in HRQoL is more strongly associated withengaging in PA rather than simply having the capacity to do so. This finding highlights the importance ofpromoting PA to slow down the natural age-related decline in FC and, consequently, HRQoL over time.

The proposed quadrant structure effectively discriminated HRQoL status. Using both PA and FC together enabled the classification of COPD patients and provided insights into HRQoL at baseline and its 3-year progression, facilitating the development of targeted strategies to improve patient prognosis.

Keywords: beta-binomial mixed-effects model, physical activity, health status.

14. Performance of Propensity Score Matching, Inverse Probability of Treatment Weighting, and Targeted Maximum Likelihood Estimation for Causal Inference in Small Samples

Adrián Aurensanz-Crespo

(*Departamento de Métodos Estadísticos, Universidad de Zaragoza*)

Cristóbal M. Rodríguez-Leal

(*Facultad de Estudios Estadísticos, Universidad Complutense de Madrid*)

Jorge Castillo-Mateo

(*Departamento de Métodos Estadísticos, Universidad de Zaragoza*)

Jesús Asín

(*Departamento de Métodos Estadísticos, Universidad de Zaragoza*)

Rosario Susi

(*Departamento de Estadística y Ciencia de los Datos, Universidad Complutense de Madrid*)

Teresa Pérez Pérez

(*Departamento de Estadística y Ciencia de los Datos, Universidad Complutense de Madrid*)

Propensity Score Matching (PSM) is a widely used method to reduce bias when estimating treatment effects in observational studies. However, its performance may be suboptimal in small sample sizes, where poor matching can lead to increased bias and variance. Alternative methods, such as Inverse Probability of Treatment Weighting (IPTW) where all patients are involved in estimating treatment or doubly robust approaches like Targeted Maximum Likelihood Estimation (TMLE), can mitigate or resolve these issues. Understanding how these methods perform under different sample size constraints and study characteristics could help guide the choice of the most appropriate method depending on the scenario.

To assess the performance of IPTW, different PSM strategies, and TMLE approaches under different scenarios, when both treatment and outcome are binary, we simulate 27 different scenarios, varying sample size, outcome prevalence, and treatment proportion. Each method (IPTW, PSM variations, and TMLE) estimates the average treatment effect in different population groups, when applicable according to the methodology, to quantify bias, Mean Squared Error (MSE) and coverage. A comparative analysis of these methods will be conducted to assess the impact of various conditions on the performance of these causal inference techniques.

Keywords: Causal inference, propensity score, Monte-Carlo simulations.

15. Multivariate Statistical Characterization of Novel Fused Bergmann Glia in a Mouse Model of Multiple Sclerosis

C.A. Ávila-Zarza

(*Department of Statistics, University of Salamanca, Salamanca, Spain*)

P.G. Téllez de Meneses

(*INCYL, Institute for Neuroscience of Castilla y León, University of Salamanca, Spain; IBSAL, Institute of Biomedical Research of Salamanca, Salamanca, Spain*)

D. Díaz

(*INCYL, Institute for Neuroscience of Castilla y León, University of Salamanca, Spain; IBSAL, Institute of Biomedical Research of Salamanca, Salamanca, Spain*)

J.R. Alonso

(*INCYL, Institute for Neuroscience of Castilla y León, University of Salamanca, Spain; IBSAL, Institute of Biomedical Research of Salamanca, Salamanca, Spain*)

J. Valero

(*INCYL, Institute for Neuroscience of Castilla y León, University of Salamanca, Spain; IBSAL, Institute of Biomedical Research of Salamanca, Salamanca, Spain*)

E. Weruaga

(*INCYL, Institute for Neuroscience of Castilla y León, University of Salamanca, Spain; IBSAL, Institute of Biomedical Research of Salamanca, Salamanca, Spain*)

The integration of multivariate statistical methods is crucial for uncovering complex biological patterns in neurodegenerative disease research. In this study, we employ an advanced multivariate statistical framework to characterize the fusion events of Bergmann glia with bone marrow-derived cells (BMDCs) in Experimental Autoimmune Encephalomyelitis (EAE), a mouse model of multiple sclerosis (MS).

Given the mixed nature of our dataset—comprising both quantitative and qualitative information—we apply Optimal Scaling techniques, specifically Multiple Correspondence Analysis (MCA), to explore associations among key cellular features. Additionally, clustering analysis is performed to identify potential subgroups, enabling a deeper understanding of their heterogeneity.

By leveraging these multivariate techniques, our approach aims to provide an unbiased and comprehensive characterization of cellular fusion patterns in the cerebellum. This methodological framework not only enhances the interpretation of complex neurobiological data but also underscores the relevance of multivariate statistical modeling in biomedical research. Our findings will contribute to a refined understanding of cellular fusion dynamics, supporting future investigations into their role in neurodegeneration.

Keywords: Multivariate Data Analysis, Multiple Correspondence Analysis, Bergmann Glia Fusion.

16. Uncovering genetic architecture of regional white matter hyperintensities using a non-parametric multivariate GWAS framework

Kahina Baouche

(*Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*)

Patricia Genius

(*Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*)

Blanca Rodriguez-Fernández

(*Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*)

Tavia E. Evans

(*Department of Human Genetics, Radboud University Medical Center, Nijmegen, Netherlands*)

Diego Garrido

(*Department of Genetics, Microbiology and Statistics, University of Barcelona (UB), Barcelona, Spain*)

Natalia Vilor-Tejedor

(*Barcelona/betaeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*)

White Matter Hyperintensities (WMH) are hallmark magnetic resonance imaging (MRI) markers of cerebral small vessel disease and early predictors of cognitive decline and Alzheimer's disease (AD). Traditional genome-wide association studies (GWAS) typically apply univariate, parametric models that assume normality and may miss subtle or spatially distributed genetic signals. We introduce a non-parametric, multivariate GWAS framework, the Multivariate Asymptotic Non-Parametric Test of Association (MANTA), to jointly analyze region-specific WMH phenotypes without requiring distributional assumptions.

Our study included 1,388 cognitively unimpaired participants from the ALFA (Alzheimer's and Families) cohort (mean age 55.71 years, 61.31 % women). WMH volumes were regionally segmented (periventricular, deep, and juxtacortical regions) using a Bayesian inference algorithm applied to T1-weighted and T2-FLAIR MRI. Genetic data were obtained via the Illumina Neurochip array and imputed using the HRC r1.1 panel.

Using MANTA, we tested multivariate associations between genotype dosages and the three correlated WMH measures, adjusting for age, sex, education, total intracranial volume, and the first five genetic principal components. We identified five genome-wide significant loci ($p < 1 \times 10^{-8}$), including novel signals undetected by standard univariate GWAS. To validate the added value of our method, we conducted a parallel MANOVA analysis on quantile-normalized phenotypes, which failed to detect several of the MANTA significant loci, underscoring the robustness and power of our non-parametric approach. Gene-set enrichment of top hits pointed to neurodevelopmental and membrane organization pathways, consistent with known WMH biology and further supporting the biological validity of the findings.

MANTA provides a statistically rigorous alternative to conventional GWAS, particularly suited for multivariate phenotypes with non-Gaussian distributions, and offers a scalable framework for imaging-genetics applications in aging and dementia research.

Keywords: Alzheimer disease, Multivariate genome-wide association studies, White matter hyperintensities, non-parametric framework.

17. Between heatwaves and hope: a Bayesian view of sustainable aquaculture

X. Barber

(*Joint Research Unit UMH-FISABIO (StatSalut), Center of Operation Research, Universidad Miguel Hernández de Elche*)

L. Aixalà

(*Joint Research Unit UMH-FISABIO (StatSalut)*)

J. Aparicio

(*Joint Research Unit UMH-FISABIO (StatSalut)*)

D. Conesa

(*Department of Statistics and Operational Research, Universitat de València*)

I. López-Mengual

(*Joint Research Unit UMH-FISABIO (StatSalut)*)

A. Forcada

(*Department of Marine Sciences and Applied Biology, Universidad de Alicante*)

A. López-Quílez

(*Department of Statistics and Operational Research, Universitat de València*)

Pablo Sánchez-Juárez

(*Department of Marine Sciences and Applied Biology, Universidad de Alicante*)

Aquaculture in the Mediterranean is increasingly vulnerable to climate-induced extreme events, such as marine heatwaves and storms, which threaten fish health, survival, and farm productivity. This study develops advanced Bayesian spatio-temporal models to quantify the impact of these events on mortality and disease, while also identifying optimal farm locations and improving operational efficiency.

We employ a multivariate coregionalized framework using R-INLA, modeling mortality conditional on disease, and incorporating relevant environmental stressors and structural barriers. A novel hurdle modeling approach accounts for excess zeros and extreme event stages. In parallel, we adapt the Efficiency Analysis Tree (EAT) method to evaluate technical efficiency across farms using environmental and health indicators, producing decision-ready metrics.

Our approach bridges ecological modeling and operations research, offering predictive tools that support aquaculture resilience, site selection, and risk-aware management strategies in the face of increasing environmental volatility.

Keywords: spatio-temporal models, efficiency analysis tree, R-INLA, EAT, climate risk, fish farms.

18. Predicting Mortality Risk from Recurrent Hospitalizations with a Poisson-Renewal Model

Telmo Perez-Izquierdo

(*Department of Economic Analysis, University of the Basque Country UPV/EHU*)

Cristóbal Esteban

(*University Hospital of Galdakao*)

Irantzu Barrio

(*Department of Mathematics, University of the Basque Country UPV/EHU & Basque Center for Applied Mathematics*)

Predicting mortality risk in patients with chronic obstructive pulmonary disease (COPD) requires an accurate understanding of the relationship between recurrent hospitalizations and death. The literature has assumed either a Poisson (calendar-time) or renewal (gap-time) model to describe hospitalization events, but neither fully captures the underlying dynamics of the process. In this study, we propose a novel Poisson-renewal model that integrates the time-trend characteristics of Poisson processes with the gap-time nature of renewal models to better characterize hospitalization patterns.

We propose to factor the hazard of the hospitalization event into a Poisson (h_0^P) and a renewal (h_0^R) part. If h_0^P is constant, then the model is purely of renewal type. If h_0^R is constant, then the model is purely Poisson. We implement a piecewise exponential model for the Poisson component and a Weibull model for the renewal component.

Our methodology is applied to an observational cohort of 512 COPD patients from the Galdakao University Hospital, with a median follow-up of 4.7 years. We fit a joint model with a Weibull baseline hazard for death and the Poisson-renewal model for hospitalizations. We specified a gamma frailty distribution (mean 1, variance θ), and included age, sex, and lung function (DLCO, and FEV_1) as covariates for both hospitalization and death, with the number of hospitalizations 2 years prior to the first evaluation included only in the hospitalization model.

Results show that hospitalization and death risks are positively related. The hazard function indicates increasing hospitalization risk during follow-up. In particular, late-concentrated hospitalizations increase death risk more than early-dispersed hospitalizations; for instance, three late hospitalizations equate to the death risk of aging 67 more days compared to three early ones.

These findings highlight the importance of incorporating both calendar-time and gap-time into hospitalization models for improved mortality prediction. Our Poisson-renewal framework enables refined risk estimation and has potential applications in clinical decision-making.

Keywords: hospitalization process; joint model; death prediction; COPD.

19. Advancing Multivariate Modeling for Spatial Ordinal Survey-Based Data

M.A. Beltrán-Sánchez

M.A. Martínez-Beneito

A. Corberán-Vallet

(*University of Valencia*)

A Bayesian hierarchical individual-level model for the analysis of spatial ordinal data from health surveys has recently been presented (Beltrán-Sánchez, MA et al., 2024). However, health survey questions are typically grouped into thematic blocks; therefore, a multivariate analysis would allow these variables to be considered as true groups. In this work, we extend the previous proposed model to a multivariate framework following the proposal of the M-model (Botella-Rocamora, P et al., 2015). This extension allows us to jointly analyze sets of ordinal response variables which are likely to be correlated. In addition, the multivariate approach enables better estimation of geographical patterns associated with each response variable, estimates correlations among them and also helps us to control the individual-level effect.

We apply this methodology to the analysis of the Health Survey of the Region of Valencia in 2022 (HSRV2022) to describe the geographical distribution of different mental health indicators of interest in this region. Specifically, we analyze the block of questions from the 12-Item General Health Questionnaire (GHQ-12) (Goldberg & Williams, 1988) included in the HSRV2022. By using this methodology, ordinal data from health surveys can be easily summarized and exploited to a greater extent, taking into account the spatial nature of the data. The results could be used by public health practitioners to intervene specifically in those areas of the region with a higher risk than the rest.

Keywords: Bayesian inference, multivariate spatial modeling, ordinal survey data.

20. Assessing a Novel Approach to Cut-off Estimation Using Bootstrapping: A Simulation-Based Study

Lucia Blanc

(*Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Badalona, Spain*)

Natàlia Pallarès Fontanet

(*Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Badalona, Spain*)

Cristian Tebé

(*Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Badalona, Spain*)

Finding a cut-off for biological markers is essential for clinical diagnosis and prognosis in many pathologies. It allows population stratification and provides an objective criterion for risk assessment and treatment selection. However, establishing reliable and reproducible cut-offs remains challenging due to the uncertainty about the appropriateness of dichotomizing continuous variables and the inherent dependence on specific data, which limits generalizability across populations.

Conventional methods often introduce biases, such as the Youden index assumption of a single optimal threshold, leading to inconsistent results and overestimation of the biomarker effect. This compromises statistical precision and power, resulting in a significant loss of information.

To overcome these challenges, we propose a cut-off selection method that scans the entire range of biomarker values and assesses its discrimination capacity using a log-binomial model. Then, the process is replicated using bootstrap resampling to estimate the empirical distribution of the potential cut-off(s). This method improves robustness and reliability by prioritizing model likelihood over p-values using the Akaike Information Criterion. The empirical distribution of bootstrapped cut-off points will be visualized to identify optimal and stable thresholds.

To assess the proposed method, a simulation study will be conducted under six different scenarios, varying the biomarker distribution (normal or positively skewed) and the expected difference between positive and negative cases based on effect size (null, moderate, or large). The null scenario analysis will ensure that the approach does not generate false-positive cut-offs where none should exist and will demonstrate improved stability in cut-off estimation. Whereas, in moderate and large effect size scenarios, the method should correctly identify meaningful thresholds with greater precision.

The interpretation and validation of the results will be assessed using sensitivity, specificity, positive predictive value, negative predictive value, and accuracy, providing a comprehensive understanding of the method's diagnostic performance and its ability to identify cut-offs.

Keywords: Biological cut-off points optimization, bootstrapping resampling, Akaike Information Criterion (AIC), biostatistics.

21. Association models for count variables with excess zeros using Zero-Inflated Negative Binomial regressions: the case of telomere length and internalizing symptoms, TeloNeuro Project

Irene Campos Sánchez
(*Universidad Miguel Hernández de Elche*)

Desirée Valera Gran
(*Universidad Miguel Hernández de Elche; ISABIAL*)

Josep Xavier Barber
(*UMH-Fisabio (STATSALUT); Universidad Miguel Hernández de Elche*)

Maria Paula Noce
(*Universidad Miguel Hernández de Elche*)

Izaro Babarro
(*Biogipuzkoa Health Research Institute; Universidad del País Vasco*)

Sabrina Llop
(*CIBERESP - Instituto de Salud Carlos III; FISABIO - Universitat Jaume I - Universitat de València*)

Isolina Riaño-Galán
(*Universidad de Oviedo; Hospital Universitario Central de Asturias; CIBERESP*)

Eva María Navarrete Muñoz
(*Universidad Miguel Hernández de Elche; ISABIAL; UMH-Fisabio (STATSALUT)*)

Telomeres are known as biomarkers of cellular aging; however, there is very little evidence investigating their possible relationship with child neuropsychological development. The aim of the present study was to analyze the association between leukocyte telomere length (LTL) and psychoemotional problems at age 8 years. For this purpose, data from the INMA project (https://www.proyectoinma.org/) were used, including a total of 673 children from the Asturias, Gipuzkoa and Valencia cohorts. LTL was determined by qPCR in blood samples, and psychoemotional problems were assessed with the Strengths and Difficulties Questionnaire (SDQ). In addition, sociodemographic and lifestyle covariates were considered. The SDQ questionnaire provides a total score for psychoemotional problems and scores for two subscales (internalizing and externalizing symptoms), all of which are count-type variables. Normally, negative binomial (NB) regression models are used to analyze this type of variables. However, an excess of zeros was observed in the internalizing symptoms subscale, so it was decided to use zero-inflated negative binomial (ZINB) regression models to estimate the associations, adjusting for possible confounding factors. This method allowed the distribution of the data to be adequately modeled, improving the interpretation of the results.

Keywords: count, excess zeros, Zero-Inflated Negative Binomial, health studies.

22. Large Language Models for Improve the Efficiency and Scalability of Data Extraction in Medical Systematic Reviews

Lorenzo Carità

(*Statistics Unit, Vall d'Hebron Institut of Oncology (VHIO)*)

Anna Aguilera

(*Statistics Unit, Vall d'Hebron Institut of Oncology (VHIO)*)

Víctor Navarro

(*Statistics Unit, Vall d'Hebron Institut of Oncology (VHIO)*)

Maria Lee

(*Statistics Unit, Vall d'Hebron Institut of Oncology (VHIO)*)

Guillermo Villacampa

(*Data Science Unit, SOLTI Cancer Group*)

Introduction: Systematic reviews and meta-analyses are important sources of evidence in medicine, but manually extracting data from studies is time-consuming and resource-intensive. Large Language Models (LLMs) have the potential to substantially reduce the time required for data extraction and information restructuring. In this work we develop an LLM-powered approach for extracting and analyzing data from over 200 randomized clinical trials on ovarian cancer.

Methods: A systematic review was conducted to identify randomized clinical trials (RCTs) reported in ovarian cancer over the past twenty years. A structured prompt was designed to extract key information from the selected RCTs. Information from the following endpoints were extracted from each study: complete response (CR), objective response rate (ORR), duration of benefit (DoB), progression-free survival (PFS), and overall survival (OS), as well as information on the study design and patient characteristics. Extracted data were standardized and synthesized for meta-analysis, comparing treatment strategies and assessing their evolution over time. This information will be used to compare the efficacy of chemotherapy, PARP inhibitors (PARPi), immunotherapy, antiangiogenic agents and other treatments used in first- and second-line settings, while also mapping their temporal adoption in the treatment landscape.

Results: Results on the time efficiency improvements of using LLMs compared to manual information extraction will be presented at the conference. Additionally, metrics to quantify the accuracy of LLM will also be presented.

Conclusions: This study explores the potential of LLMs to improve the efficiency and scalability of data extraction in systematic reviews. By automating data extraction, this method not only reduces the time required for systematic reviews but also improves scalability and reproducibility, offering significant potential for more efficient decision-making in clinical research.

Keywords: Large Language Models (LLMs), meta-analysis, data extraction.

23. Enhancing Epidemiological Risk Assessment: Incorporating Age in Spatio-Temporal Models for Suicide Emergency Calls

Javier Martín-Pozuelo

(*Department of Statistics and Operations Research, University of Valencia*)

Xavier Barber

(*Center of Operations Research, Miguel Hernández University*)

Antonio López-Quílez

(*Department of Statistics and Operations Research, University of Valencia*)

Miriam Marco

(*Department of Social Psychology, University of Valencia*)

In epidemiological research, disease mapping plays a crucial role in providing insight into the spatial distribution of different causes of disease or death. In this context, one of the most widely used methods is indirect standardisation. This technique is used to remove the effect of confounding factors that hinder the understanding of true disease risk patterns. While this approach simplifies comparisons between populations by removing the influence of a specific factor, it also has limitations by removing valuable information.

This study aims to explore an alternative to improve risk estimation by incorporating the factor into spatio-temporal modelling. In particular, we focus on age as a key factor in the analysis of suicide-related emergency calls in the Valencian Community. Integrating age into a model, rather than opting for its standardisation, allows us to capture its effect more accurately, allowing us to explore both the individual effect and its spatial and temporal interactions. This alternative not only improves risk estimates, but also provides a more complete perspective for public health decision-making, facilitating targeted interventions based on a thorough understanding of the underlying risk structures.

Keywords: indirect standardization, disease mapping.

24. Integrating spatio-temporal barriers into global marine species distribution models

Alba Fuster-Alonso

(*Institute of Marine Science (ICM-CSIC), Barcelona, Spain*)

Haavard Rue

(*King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia*)

David Conesa

(*Department of Statistics and Operations Research (VaBar), University of Valencia (UV), Valencia, Spain*)

Finn Lindgren

(*University of Edinburgh (UoE), Edinburgh, Scotland*)

Elias T Krainski

(*King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia*)

Species distribution models (SDMs) are key for understanding and predicting species' spatio-temporal patterns, especially under global change. Although SDMs have been widely applied on regional scales, there is a growing need to extend these models globally to capture the large-scale environmental and ecological processes driving marine species distributions. To move toward more realistic global marine SDMs, we need methodological improvements, such as accounting for Earth's spherical geometry and considering dispersal constraints due to geographic barriers.

We present a novel approach to global species distribution modeling by integrating spatio-temporal barriers into the INLA-SPDE (Integrated Nested Laplace Approximation - Stochastic Partial Differential Equation) framework. This method allows for efficient Bayesian inference of spatial and temporal dynamics while explicitly incorporating physical barriers and Earth's spherical geometry. By building a worldwide mesh on spherical coordinates and anisotropic covariance structures, we ensure biologically realistic connectivity across global ocean basins.

We assess the capabilities of our model through a comprehensive simulation study, evaluating its performance under both continuous and binary response scenarios. Furthermore, we apply our method to two case studies, Atlantic cod (Gadus morhua) and shallow coral species, which demonstrate improved spatial predictions and ecological realism compared to models that ignore barriers and Earth's spherical geometry. Our findings emphasize the importance of incorporating geographical and ecological constraints in global SDMs, providing a computationally efficient framework for ecological modeling and climate impact assessment in marine systems.

Keywords: global scale, barrier model, marine environment, and INLA.

25. Bayesian spatial modeling of interrupted time series: Assessing the impact of health alerts on opioid prescriptions in the Valencian Community, Spain

Celia Robles
(*HSR Unit, FISABIO*)

Adina Iftimi
(*Department of Statistics and Operations Research, University of Valencia*)

Isabel Hurtado
(*HSR Unit, FISABIO*)

Aníbal García-Sempere
(*HSR Unit, FISABIO*)

Opioid consumption in Europe has increased in recent years, raising significant public health concerns. Among these substances, fentanyl –particularly its immediate release formulations (IRF)– carries a high risk due to its rapid onset, potency, and potential for misuse, dependence, and overdose. In Spain, IRF fentanyl is officially approved solely for the management of breakthrough cancer pain (BCP) in patients already receiving chronic opioid therapy for cancer-related pain. However, emerging data indicate a widespread pattern of off-label prescriptions, including its use for non-cancer pain (NCP) and administration to opioid-naïve cancer patients, raising concerns about improper prescribing practices and associated health risks. To reduce inappropriate prescribing, Spanish health authorities implemented three key interventions in Valencia. These measures were analyzed using interrupted time series (ITS) analysis, revealing a temporary reduction in off-label prescribing.

ITS analysis employs segmented regression models to detect both immediate and long-term changes following an intervention. However, conventional ITS methods do not account for spatial variation, potentially failing to capture regional disparities in prescribing behavior. Building on this limitation, recent studies have introduced spatially interrupted time-series (SITS), which capture both spatial variations and temporal trends. However, modeling space-time interactions remains crucial, as policy impacts may manifest differently across regions and evolve over time. Knorr-Held models provide a robust Bayesian framework for capturing these interactions by integrating independent spatial and temporal effects while accounting for their interdependencies.

Our results highlight the importance of incorporating spatiotemporal effects in the analysis of public health interventions. The posterior distribution of the coefficients associated with spatial effects shows variability across departments, indicating that the response to interventions has not been homogeneous throughout the region. Additionally, the space-time interaction reveals dynamic patterns over time, suggesting that changes in opioid prescribing depend not only on the timing of policy implementation but also on regional factors that may amplify or mitigate their impact. Including these effects in the model allows for a more accurate evaluation of the implemented policies, facilitating the identification of high-risk areas and optimizing future intervention strategies.

Keywords: Monte Carlo Markov Chains, sparse model selection, credibility estimation, cost-effectiveness, health technology assessment.

26. `qPRAentry`: A comprehensive R package for quantifying the risk of plant pest entry

Maria Chiara Rosace

(*Department of Sustainable Crop Production, Università Cattolica del Sacro Cuore*)

Antonio Vicent

(*Centre de Protecció Vegetal i Biotecnologia, Institut Valencià d'Investigacions Agràries (IVIA)*)

Matteo Crotta

(*European Food Safety Authority (EFSA)*)

Ewelina Czwienczek

(*European Food Safety Authority (EFSA)*)

David Makowski

(*UMR MIA-PS University Paris-Saclay, INRAe*)

Vittorio Rossi

(*Department of Sustainable Crop Production, Università Cattolica del Sacro Cuore*)

Martina Cendoya

(*Centre de Protecció Vegetal i Biotecnologia, Institut Valencià d'Investigacions Agràries (IVIA)*)

The introduction of pests through the movement of plants and plant products across international borders represents a major threat to agriculture and ecosystem stability. Quantitative Pest Risk Assessment (qPRA) plays a fundamental role in evaluating and mitigating these risks. The entry phase of qPRA focuses on assessing the likelihood of a pest reaching a new area, a process that involves the identification and quantification of potential entry pathways. Despite advancements in pathway models, risk assessments are often constrained by data limitations, including the aggregation of trade data at the country level and the complexity of tracking re-exported commodities. To address these challenges, we introduce the comprehensive R package `qPRAentry` designed specifically for the initial phase of the qPRA, i.e., the entry step. The methodology implemented is based on the qPRAs developed by the European Food Safety Authority (EFSA), adapted to be applied in a flexible way according to the characteristics of each pest. The package includes functions to calculate the quantity of commodities imported from countries where a pest under assessment is present, and to implement a pathway model capable of assessing the risk of introduction of new pest founder populations. Using trade data, the developed functions allow an approximation of the quantity of potentially infested commodities remaining in a country of interest and their redistribution among the different regions of the country. In a second stage, the quantity of potentially infested commodity serves as an input to a pathway model that estimates the potential number of founder populations in each country or region of interest. The package provides flexibility in designing the pathway model, enabling parameter adaptation to the characteristics of the pest under assessment. In addition, functions for visualising the results are included, as well as Shiny applications that allow the qPRA entry step to be performed in a user-friendly way. Overall, the package improves the consistency and reproducibility of the qPRA entry step.

Keywords: Matrix completion, collaborative filtering, hypothesis testing, permutation test, movie recommender systems.

27. A computationally efficient procedure for combining ecological datasets by means of sequential consensus inference

M. Figueira

(*Department of Statistics and OR, University of Valencia, Spain*)

D. Conesa

(*Department of Statistics and OR, University of Valencia, Spain*)

A. López-Quílez

(*Department of Statistics and OR, University of Valencia, Spain*)

I. Paradinas

(*AZTI, Txatxarramendi Ugartea,Sukarrieta, Spain*)

In ecology and environmental sciences, combining diverse datasets has become an essential tool for managing the increasing complexity and volume of ecological data. However, as data complexity and volume grow, the computational demands of previously proposed models for data integration escalate, creating significant challenges for practical implementation. This study introduces a sequential consensus Bayesian inference procedure designed to offer the flexibility of integrated models while significantly reducing computational costs.

The method is based on sequentially updating some model parameters and hyperparameters, and combining information about random effects after the sequential procedure is complete. The implementation of the approach is provided through two different algorithms. The strengths, limitations, and practical use of the method are explained and discussed throughout the methodology and examples.

Finally, we demonstrate the method's performance using two different examples with real ecological data, highlighting its strengths and limitations in practical ecological and environmental applications.

Keywords: expert elicitation; recursive inference; sequential learning.

28. Bayesian Spatial Analysis of Socio-Economic Inequities in HPV Vaccination Coverage

María Herrero-Silvestre

(*Vaccine Research Department, Fisabio-Public Health, Valencia, Spain*)

Ana Corberán-Vallet

(*Department of Statistics and Operations Research, University of Valencia, Valencia, Spain*)

Cintia Muñoz-Quiles

(*Vaccine Research Department, Fisabio-Public Health, Valencia, Spain*)

Álvaro Briz-Redón

(*Department of Statistics and Operations Research, University of Valencia, Valencia, Spain*)

Alejandro Orrico-Sánchez

(*Vaccine Research Department, Fisabio-Public Health, Valencia, Spain*)

Mónica López-Lacort

(*Vaccine Research Department, Fisabio-Public Health, Valencia, Spain*)

Human papillomavirus (HPV), one of the most common sexually transmitted viruses, has been linked to several types of cancer. The Valencia Region included the HPV vaccine in its immunisation programme in 2008. However, vaccination coverage remains suboptimal in many areas. This study aims to identify neighbourhoods with low vaccination coverage and analyse the socio-economic factors that influence it.

To do that, a retrospective study was conducted using real-world data from the Valencia Health System Integrated Database (VID). Girls born in the region between 1995 and 2006 were included in the study (2009–2017). We used a Bayesian spatial clustering model to identify neighbourhoods with similar vaccination coverage. Covariates such as non-EU population, cadastral value and net income were included in the model to assess their associated effect.

A total of 25,204 girls were included, of whom 22,479 received the HPV vaccine. The 70 neighbourhoods analysed were grouped into 4 clusters, with average coverage ranging from 83.9 % to 91.8 %. In particular, neighbourhoods with a higher non-EU population, high cadastral values or low income had lower coverage, with decreases of about 16 % (95 % CI: 0.9 %-28.7 %), 16 % (95 % CI: 4.9 %-26.4 %) and 14 % (95 % CI: 2.8 %-23.9 %), respectively.

Keywords: HPV vaccination; Spatial clustering model; Bayesian analysis.

29. Early detection of Porcine Reproductive and Respiratory Syndrome (PRRS) using sow production data

M. Lorena Correa-Abendaño

(*Unitat mixta d'Investigació IRTA-UAB en Sanitat Animal, CReSA; Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Programa de Sanitat Animal, CReSA; Departament de Matemàtiques, Universitat Autònoma de Barcelona*)

Carles Vilalta

(*Unitat mixta d'Investigació IRTA-UAB en Sanitat Animal, CReSA; Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Programa de Sanitat Animal, CReSA*)

Amanda Fernández-Fontelo

(*Departament de Matemàtiques, Universitat Autònoma de Barcelona*)

Porcine reproductive and respiratory syndrome (PRRS) is a major swine disease with great economic impact. Effective early detection is crucial to mitigate disease outbreaks, but current surveillance approaches are limited. This work studies statistical methods for detecting PRRS outbreaks in pig production data. We analyse real-world farm data to identify patterns that may indicate early disease emergence. Our findings aim to improve PRRS surveillance and provide new and useful information to the swine industry.

Keywords: PRRS, disease surveillance, anomaly detection.

30. Comparative Analysis of Variable Selection Methods in Statistical Models

Jordi Cortés Martínez

(*Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya*)

Júlia Sierra Borras

(*Universitat Politècnica de Catalunya*)

Àlex Soler Carballo

(*Universitat Politècnica de Catalunya*)

Cristian Tebé Cordomí

(*Unitat de Suport i Recerca en Bioestadística, Institut de Recerca Germans Trias i Pujol*)

Introduction: Variable selection is a critical step in statistical modelling, as it affects the accuracy, interpretability, and computational efficiency of the model. Different methods have different assumptions, strengths and limitations, so they'll perform differently depending on the data. Through data simulation under different scenarios, methods such as Stepwise Selection, Bootstrap, Random Forest, Lasso, Elastic-Net, and Boruta will be evaluated to determine their effectiveness in selecting...

Methodology: To conduct a comprehensive comparison, synthetic datasets will be generated with both continuous and binary outcomes, different sample sizes and a range of variable types and associations. The selected variable selection methods will be applied to these datasets, and their performance will be assessed using key metrics such as Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value, which refer to the ability of each method to correctly identify the truly ...

Expected Results: It is anticipated that different methods will exhibit varying levels of sensitivity and specificity, depending on the data structure and complexity. The study will also examine how factors like outcome prevalence and sample size influence selection stability and model performance. While stepwise selection may struggle with high-dimensional data, Lasso and ElasticNet are expected to perform well in cases with correlated variables. Random Forest and Boruta, being tree-based metho...

Conclusion: The study aims to provide insight into the trade-offs of each selection method, and guide practitioners to the most relevant approach for their specific data and modelling needs. By systematically comparing these variable selection techniques, this research will contribute to the optimization of statistical models in biomedical fields.

Keywords: variable selection, modelling, linear regression, logistic regression.

31. Evaluation of longitudinal surrogate endpoints: a case study of schizophrenia

Gokce Deliorman
(*Complutense University of Madrid*)

Maria del Carmen Pardo
(*Complutense University of Madrid*)

Ariel Alonso
(*KU Leuven*)

Geert Molenberghs
(*Hasselt University*)

In clinical trials, surrogate endpoints provide earlier insights into treatment effectiveness, reducing follow-up time and costs. Most existing methods assess surrogate markers at a single time point, although in many settings they are repeatedly measured over time. In this work, we examine a scenario where both true and surrogate endpoints are measured longitudinally through clinical trial data from schizophrenia studies.

Keywords: Galecki's model, Individual causal association (ICA), Longitudinal outcomes, Surrogate endpoints.

32. Predicting Hospital Occupancy with Random Slope Poisson Models in Small-Area Studies

Naomi Diz-Rosales
(*Universidade da Coruña, CITIC*)

María José Lombardía
(*Universidade da Coruña, CITIC*)

Domingo Morales
(*Universidad Miguel Hernández de Elche, IUICIO*)

The COVID-19 pandemic has highlighted the critical importance of accurate healthcare demand forecasting. In this study, we introduce a small-area estimation approach based on a Poisson mixed model incorporating random slopes to estimate Intensive Care Unit bed occupancy. We develop the best empirical predictors and introduce bootstrap-based estimators for the mean squared error. Model parameters are estimated using maximum likelihood with a Laplace approximation, and mode predictors for random effects are calculated. Through simulation studies, the performance of the estimation algorithm, the accuracy of the predictors and the reliability of the error estimators are evaluated. The methodology is applied to Intensive Care Unit occupancy data from seven health districts in Galicia, Spain, covering the period from November 2020 to March 2022. The results demonstrate the robustness of the model at various stages of the pandemic, including different SARS-CoV-2 variants, public health interventions and vaccination rates. It also provides an introduction and evaluation of the integration of temporal correlations between random effects to improve the accuracy of Intensive Care Unit demand forecasting and allow for dynamic monitoring.

Keywords: COVID-19, Random slopes, Small Area Estimation.

33. A Piecewise Subdistribution Hazard Modeling with Time-Independent Covariates

Juan Carlos Espinosa-Moreno

(*University of the Basque Country (UPV/EHU) & BCAM - Basque Center for Applied Mathematics*)

Fernando García-García

(*BCAM - Basque Center for Applied Mathematics*)

Dae-Jin Lee

(*IE University, School of Science and Technology*)

We propose a piecewise extension of the discrete-time subdistribution hazard (SH) model for competing risks, originally introduced by Fine and Gray (1999), with time-independent covariates. Our method formulates the cumulative incidence function (CIF) via a weighted binomial Generalized Linear Model (GLM) with a complementary log-log (cloglog) link function. By discretizing follow-up time into intervals and assuming a piecewise-constant hazard within each, the model allows for flexible estimation of event-specific hazard rates while preserving the interpretability of regression coefficients. This extension directly incorporates time discretization and interval-specific covariate effects, enabling CIF estimation even in the presence of right-censored data. To address censoring, we integrate inverse probability of censoring weighting (IPCW) into the GLM framework, employing four alternative estimators for the censoring survival function, including both marginal and covariate-dependent (Cox-based) approaches.

We evaluate the performance of the proposed piecewise subdistribution hazard (PSH) model on synthetic competing risks datasets under a variety of censoring scenarios, ranging from complete data to right-censoring that is either independent or dependent on covariates, with censoring proportions set at 30 % and 50 %. Model accuracy is assessed in terms of coefficient bias, coverage probabilities, and predictive performance, measured using the integrated Brier score (IBS). Across all settings, PSH consistently recovers the true covariate effects with low bias and reliable coverage, while outperforming established competing risks methods – including cause-specific and subdistribution Cox models, as well as flexible machine learning approaches such as random survival forests – in terms of predictive error.

These results demonstrate that our proposal provides a competitive and interpretable alternative for modeling cumulative incidence functions in the presence of competing risks, particularly when follow-up time is naturally segmented or when smooth baseline hazard assumptions may not hold. Nonetheless, the method's application requires careful handling of censoring mechanisms to ensure the stability and convergence of the IPC-weighted GLM estimation.

Keywords: Competing Risks; Subdistribution Hazard; Piecewise Modeling; Censoring.

34. Quantifying, modeling, and characterizing brain perivascular spaces through deep learning and statistical inference for count data

Alba Fernandez-Bonet

(*Barcelonaβeta Brain Research Center, Barcelona, Spain*)

Patricia Genius

(*Barcelonaβeta Brain Research Center, Barcelona, Spain. Centre for Genomic Regulation, Barcelona, Spain*)

Blanca Rodríguez-Fernández

(*Barcelonaβeta Brain Research Center, Barcelona, Spain. Centre for Genomic Regulation, Barcelona, Spain*)

Armand Gonzalez-Escalante

(*Barcelonaβeta Brain Research Center, Barcelona, Spain. Hospital del Mar Research Institute, Barcelona, Spain*)

Tavia E. Evans

(*Barcelonaβeta Brain Research Center, Barcelona, Spain. Department of Human Genetics, Radboud University Medical Center, Nijmegen Netherlands*)

Natalia Vilor-Tejedor

(*Barcelonaβeta Brain Research Center, Barcelona, Spain. Centre for Genomic Regulation, Barcelona, Spain. Department of Human Genetics, Radboud University Medical Center, Nijmegen Netherlands*)

Background: Perivascular spaces (PVS) are fluid-filled compartments surrounding brain blood vessels, emerging as promising markers of small vessel disease and overall brain health. They play a key role in the glymphatic system, and have been linked to stroke, cognitive decline, and neurodegenerative diseases. MRI-defined PVS counts often exhibit heterogeneity due to overdispersion and excess of zeros, yet are commonly assessed through dichotomous classifications with limited statistical power. Traditionally, PVS are visually assessed by radiologists, which introduces subjectivity and limits statistical robustness. This study aims to apply a deep learning algorithm for PVS quantification and to identify a suited probability distribution for examining determinants of PVS burden in cognitively unimpaired individuals.

Methods: Convolutional neural network regression was used to quantify PVS counts by training and validating MRI-T2 contrasts images acquired from 322 individuals from the ALFA+ study across four brain regions: basal ganglia (BG), centrum semiovale (CSO), midbrain (MB), and hippocampus (HP). The performance of key count-based probability distributions (Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial) was evaluated through simplistic generalized linear regression models, adjusting for age and sex as confounders. Vuong's non-nested statistical test was applied to compare the standard distributions with their zero-inflated counterparts. Demographics, cardiometabolic and mental health factors, and polygenic risk scores associated with Alzheimer's disease (AD)-related pathways, metabolic traits, and sleep disturbances were examined as potential determinants of regional PVS. Interaction and stratified models by sex, amyloid status, and AD CSF amyloid/tau biomarker used for diagnosis classification framework (AT) groups were assessed to explore potential modifiers.

Results: Poisson and negative binomial distributions provided the best fit for modeling MRI-defined PVS counts in BG and HP, and CSO and MB, respectively. In the CSO, significant age-related increases in PVS counts and consistent sex differences were observed, with men showing higher counts. In the HP, models showed that sex, sleep duration, and APOE-ε4 were associated with greater PVS counts among individuals with higher amyloid burden. Nominal trends included systolic blood pressure and CAIDE-II (HP); particulate matter (PM2.5) (BG, CSO), and genetic predisposition to AD through cholesterol efflux pathway (MB), particularly in individuals with higher amyloid burden. Sex-specific models revealed a significant inverse association between BMI and PVS burden in men (BG), along with nominal effects

of genetic predisposition through complex lipoprotein (BG) and amyloid pathways (CSO, MB) also in men. Biomarker-based models highlighted associations for smoking in individuals with high amyloid and tau burden (CSO) and systolic blood pressure in those with low biomarker burden (HP), with additional nominal trends for PM2.5 (BG), sleep duration and depression (HP), and cholesterol efflux pathway (MB) in those with elevated amyloid burden.

Conclusion: These findings demonstrate the utility of deep learning-based automatic PVS quantification and advanced statistical modeling, including Vuong's test for distribution selection, in handling overdispersed and zero-inflated count data. By identifying regional and sex-specific determinants of PVS burden, and accounting for environmental and genetic factors, this approach enhances the robustness of epidemiological research on brain health and cognitive aging.

Keywords: Perivascular spaces, genetic of Alzheimer's disease, risk factors AD continuum.

35. Longitudinal Bayesian Analysis of Pain Perception in Mecp2-Heterozygous Female Mice: A Study of Disease Progression in Rett Syndrome

Javier Cuitavi
(*Universitat de València*)

Elena Martínez-Rodríguez
(*Universitat de València*)

María Abellán-Álvaro
(*Universitat de València*)

Moisés García-Arencibia
(*Consejería de Educación, Formación Profesional, Actividad Física y Deportes, Gobierno de Canarias*)

Mónica Santos
(*CNC: Universidade de Coimbra Instituto de Investigação Clínica e Biomédica de Coimbra*)

Lucía Hipólito
(*Universitat de València*)

Anabel Forte
(*Universitat de València*)

Carmen Agustín-Pavón
(*Universitat de València*)

Jose V. Torres-Pérez
(*Universitat de València*)

Rett Syndrome (RTT) is a neurodevelopmental disorder primarily affecting females, with symptoms that progressively impact motor and sensory functions. In this study, we applied a Bayesian normal linear mixed model (LMM) to longitudinal data from 2- to 6-month-old Mecp2 heterozygous (Mecp2-het) female mice to explore the evolution of sensory abnormalities associated with RTT. Our approach allowed us to account for the small sample sizes and variability in symptom onset by treating each mouse's disease progression individually while modelling the overall trend of symptom development.

We found significant variability in the timing and progression of symptom onset, with Mecp2-het females classifiable as early- or late-symptomatic based on the emergence of hallmark neurological features such as clasping, a hallmark of neurodegeneration, and gait abnormalities. This heterogeneity within the model highlights the need to stratify subjects by symptom onset in future RTT studies to better understand the diverse trajectories of disease progression. Bayesian modelling provided valuable insight into the transition of nociception, with a shift from pre-symptomatic hypersensitivity to apparent hyposensitivity as motor symptoms emerged.

Additionally, using our model, we observed a decrease in neuronal activation in the periaqueductal gray of 6-month-old Mecp2-het females following the hot plate test, indicated by lower FOS expression. This finding aligns with a reduction in cannabinoid receptor 1 (CB1) expression in this brain region, compared to wild-type controls. By modelling the data within a Bayesian framework, we were able to quantify the uncertainty in these estimates and the variability in sensory abnormalities across different stages of RTT.

Our study demonstrates the power of Bayesian longitudinal models in understanding disease progression in RTT, particularly in examining the changes in nociception over time. These results emphasize the

importance of the presymptomatic phase for understanding sensory dysfunction and provide a framework for addressing the challenges of pain assessment in RTT patients.

Keywords: Longitudinal data, small sample sizes, Bayesian Analysis, Rett Syndrome.

36. Optimizing dynamic predictions from joint models for multivariate longitudinal and time-to-event data via super learning approach

Arnau Garcia-Fernández

(*Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Institute for Research and Innovation in Health (IRIS)*)

Dimitris Rizopoulos

(*Erasmus MC, Department of Biostatistics, Rotterdam*)

Guadalupe Gómez Melis

(*Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Institute for Research and Innovation in Health (IRIS)*)

Time-to-event and longitudinal data are common in health studies, and joint models (JMs) provide a way to analyze both while allowing dynamic predictions that are updated over time. Applying JMs to multivariate longitudinal data is computationally challenging. Ensemble methods like super learning (SL) combine algorithms to obtain optimal predictions and offer a solution for these problems. This work explores the use of SL for deriving dynamic predictions in multivariate JMs.

Keywords: Joint models, dynamic predictions, super learning.

37. On the evaluation of time-dependent discrimination ability for survival models

Leire Garmendia Bergés

(*BCAM - Basque Center for Applied Mathematics*)

Irantzu Barrio

(*Department of Mathematics, University of the Basque Country UPV/EHU*)

Guadalupe Gómez Melis

(*Statistics and Operations Research Department and Institute for Research and Innovation in Health (IRIS), Universitat Politècnica de Catalunya - BarcelonaTech (UPC)*)

Survival models are often used to predict the course of future individuals. The time-dependent AUC (AUC(t)) is commonly used to quantify the ability of a survival model to correctly predict future events at a fixed time t. This work aims to analyze the asymptotic properties of the Conditional Inverse Probability of Censoring Weighting (CIPCW) estimator and to compare its behavior to other AUC(t) estimators under conditions different from those analyzed in the literature.

Keywords: Survival analysis; Discrimination ability; Time-dependent AUC.

38. Exploring partial least squares and resampling methods to identify latent risk profiles of cerebrovascular progression in healthy asymptomatic individuals

Patricia Genius

(*Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*)

Blanca Rodríguez-Fernández

(*Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*)

Mahnaz Shekari

(*Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*)

Carole Sudre

(*MRC Unit for Lifelong Health and Ageing, Department of Population Science and Experimental Medicine, University College London, London, United Kingdom*)

Juan Domingo Gispert

(*Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*)

Natalia Vilor-Tejedor

(*Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*)

Analyzing high-dimensional epidemiological data presents major statistical challenges due to multicollinearity among predictors and complex covariance structures with outcome variables. In neuroepidemiology, such complexity necessitates robust multivariate modelling approaches to reveal risk profiles associated with early structural brain changes in asymptomatic individuals.

In this study we investigate risk profiles linked to the progression of white matter hyperintensities (WMH), a core marker of cerebral small vessel disease (CSVD), in a cohort of 111 cognitively unimpaired (CU) middle-aged individuals at risk of Alzheimer's disease (AD). Participants exhibited progression or stability in global WMH volumes over a 3-year follow-up (Progressors + Stables: $\Delta\text{WMH}_{v2-v1} >$ Percentile $\Delta\text{WMH}_{v2-v1} = 0 + 0{,}10$). Predictor variables included cerebrospinal fluid (CSF) biomarkers, exposure to air pollutants, AD pathway-specific polygenic risk scores (PRS) and cardiovascular risk factors. Cognition was evaluated at both time-points.

We employed principal component analysis on regional WMH volumes to capture shared variance across brain regions (i.e., patterns of regional WMH progression). Partial least squares (pls) regression was applied to estimated latent components linking predictors sets to annualized WMH progression. To assess model stability and variable importance, we implemented a non-parametric bootstrap procedure (1000 iterations, 80 % resampling). Latent constructs were then examined for their association with cognitive outcomes, using univariate linear regression and interaction models.

While pls-derived latent factors exhibited limited robustness under resampling, specific variables consistently emerged as significant predictors of WMH progression. These included CSF biomarkers of vascular dysfunction, neurodegeneration, amyloidosis and inflammation, alongside cardiovascular risk factors, all of which remained significant under bootstrapped 97.5 % confidence intervals. Subsequent linear modelling revealed that these markers were also associated with worse memory performance, independent of age, sex, baseline cognition and AD genetic risk factors.

Our findings showed that even in the presence of statistical noise and weak latent structure, biologically and clinically meaningful predictors of WMH progression and cognitive decline can be identified using robust multivariate modelling. The study underscores the importance of incorporating multiblock statistical frameworks and resampling strategies in epidemiological research, particularly when sample sizes are constrained, and predictor sets are high-dimensional.

Keywords: Multivariate modelling; Epidemiology; Latent variable modelling, Partial Least squares.

39. Addressing bias in statistical inference based on epidemiological registry data

Dídac Gallego Alabanda

(*Universitat Politècnica de Catalunya - BarcelonaTech (UPC)*)

Guadalupe Gómez Melis

(*Statistics and Operations Research Department and Institute for Research and Innovation in Health (IRIS), Universitat Politècnica de Catalunya - BarcelonaTech (UPC)*)

Oleguer Plana-Ripoll

(*Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital and National Centre for Register-based Research, Aarhus University, Aarhus, Denmark*)

Hospital records are commonly used to study diseases at the population level. When the goal is analyzing disease incidence, survival analysis methods are typically applied, as some individuals may be censored, meaning that the onset of the disease is never observed. However, estimating factors such as the age of disease onset (e.g., through the survival function) using all the available data may lead to biased results due to potential misinterpretation of these records.

The hospital records considered in this talk include, for each individual visiting the hospital for a specific disease, the reason for the visit (e.g., the disease of interest), the date of the visit, and a unique patient identifier. Such records are commonly used in countries, like Denmark, both for administrative and research purposes. When civil registry data are also available (including information such as birth-date and sex) it becomes possible to construct a database that records each individual's age at the time of hospital visits, enabling the estimation of the age of disease incidence. A common approach is to identify the first recorded hospital visit for a given disease as the moment of diagnosis. However, including individuals with incomplete information (e.g., those whose records only begin years after their birth) can lead to gaps in visit histories. Consequently, defining the first observed visit as the time of disease diagnosis may introduce bias into incidence age estimation, as these "first visits"might actually correspond to recurrent visits.

This talk explores the bias that arises when estimating disease onset age using hospital visit records as described. We introduce a notation that frames this issue within the counting process framework, representing each individual as a counting process that increments with hospital visits related to the disease of interest. To address these biases, we propose a methodology that implements a "washout"window at the start of the recording period, filtering out individuals who might otherwise be misclassified as incident cases. Finally, we validate our approach through a simulated population, evaluating both the magnitude of these biases and the effectiveness of the washout window method.

Keywords: counting process; disease incidence; register-based epidemiology.

40. "INLA con cosas": a toolbox for approximate Bayesian inference

V. Gómez-Rubio

(*Department of Mathematics, Universidad de Castilla-La Mancha*)

H. López-Gómez

(*Department of Mathematics, Universidad de Castilla-La Mancha*)

G. García-Donato Layrón

(*Department of Economy and Finance, Universidad de Castilla-La Mancha*)

The integrated nested Laplace approximation (INLA) can be used for Bayesian inference on models that can be expressed as latent Gaussian Markov random fields (GMRFs). In particular, INLA will provide estimates of the posterior marginals of model parameters and latent effects in a short time. Furthermore, INLA is implemented in the R-INLA package for the R statistical software. INLA has proven to be very useful for fitting a wide range of highly structured models such as spatial and temporal ones.

There are a number of models that INLA is not able to fit. For example, mixture models and double hierarchical models. In order to be able to conduct Bayesian inference for a larger class of models, some authors have combined INLA with more general algorithms. For example, INLA have been integrated into Markov chain Monte Carlo methods as well as importance sampling.

In addition to model fitting, INLA can be used for model assessment and selection. An important example is that of computing the posterior probability of models as this is important for variable selection, for example. This approach will be illustrated using different examples.

Keywords: Bayesian inference, model selection, variable selection.

41. Bayesian Spatio-Temporal Methods for Analysing Land Use Changes

Carmen Guarner
(*University of Valencia*)

Mario Figueira
(*University of Valencia*)

David Conesa
(*University of Valencia*)

Antonio López-Quílez
(*University of Valencia*)

Tamás Krisztin
(*International Institute for Applied Systems Analysis (IIASA)*)

Changes in land use patterns have significant environmental and socio-economic impacts, making it crucial for policymakers to understand their causes and consequences. This study, part of the European LAMASUS (Land Management for Sustainability) project, presents a methodological synthesis for treating land use data using a Bayesian approach within spatial and spatio-temporal modelling structures. It provides a comprehensive framework for understanding and managing land use changes through the implementation of key models, downscaling techniques, and Big Data solutions.

Keywords: Land use, spatio-temporal models, Big Data

42. Longitudinal Clustering for Symptom Trajectories in First-Episode Psychosis: A Comparison of Techniques

Laura Julià
(*IDIBAPs, University of Barcelona*)

Sergi Mas
(*IDIBAPs, University of Barcelona, CIBERSAM*)

Manuel Jesús Cuesta-Zorita
(*Hospital Universitario de Navarra, IdiSNA, CIBERSAM*)

Benedicto Crespo-Facorro
(*Hospital Universitario Virgen del Rocío, IBiS, CIBERSAM*)

Javier Vázquez-Bourgon
(*Hospital Universitario Marqués de Valdecilla – IDIVAL, Universidad de Cantabria, CIBERSAM*)

Carlos Spuch
(*IIS-Galicia Sur, CIBERSAM*)

Ana Gonzalez-Pinto
(*Hospital Universitario Alava, UPV/EHU, BIOARABA, CIBERSAM*)

Angela Ibañez
(*Hospital Universitario Ramón y Cajal, Universidad de Alcalá, CIBERSAM*)

Judith Usall
(*Parc Sanitari Sant Joan de Déu, CIBERSAM*)

Cristina Romero-López-Alberca
(*University of Cadiz*)

Ana Catalan
(*Basurto University Hospital, University of the Basque Country, King's College London, CIBERSAM*)

Anna Mané
(*Hospital del Mar Research Institute, UPF, CIBERSAM*)

Silvia Amoretti
(*VHIR, CIBERSAM*)

Josep Lluís Carrasco
(*University of Barcelona*)

Identifying response phenotypes in first-episode psychosis (FEP) is a key objective of the FarmaPRED-PEP project, which aims to improve understanding of antipsychotic treatment heterogeneity. This study compares non-parametric and parametric clustering techniques applied to symptom trajectories measured

with the Positive and Negative Syndrome Scale (PANSS) in the PEPS cohort. We evaluate three clustering techniques using Euclidean distance: hierarchical clustering with Ward's method, k-means, and Partitioning Around Medoids. Internal validation metrics, including the silhouette index, Dunn index, and Calinski–Harabasz Index, guide the selection of the optimal number of clusters.

For parametric modelling, we apply Latent Class Mixed Models with three different time structures: linear, quadratic, and cubic splines. The best-fitting model is determined using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and entropy. We evaluate cluster stability via bootstrapping and generate simulated PANSS trajectories to assess the methods' ability to retrieve true underlying subgroups. This comparison provides insights into the most reliable approaches for stratifying FEP patients based on symptom evolution, contributing to a more personalized treatment framework.

Keywords: longitudinal clustering, latent class models, first-episode psychosis.

43. Effect Size-Driven Pathway Meta-Analysis for Gene Expression Data

Pablo Pedro Jurado Bascón

Juan Antonio Villatoro García

Pedro Carmona Sáez

(*Department of Statistics and Operations Research, University of Granada; GENYO. Centre for Genomics and Oncological Research: Pfizer/University of Granada/Andalusian Regional Government, PTS Granada, Spain*)

The steadily increase in publications that share omics datasets in public repositories has become a great advantage for studies that try to merge/find common results for a given health disorder, therefore meta-analysis is the current go-to models to find new insight in omics sciences, specifically for gene expression, as the primary source of studies in this field.

The common approach is to make a meta-analysis of each gene individually, but it must be noted that this leads to a lack of statistical power due to noise and artifacts, while not all genes are shared across most studies, losing information when joining genes between cases. To address these limitations, we propose GSEMA (Gene Set Enrichment Meta-Analysis), a novel methodology that leverages single-sample enrichment scoring techniques to aggregate gene expression data into pathway-level matrices. Working with pathways instead of genes preserves information for missing genes in some studies, and it is known that pathway analysis provides more consistent and relevant results.

Pathway matrices were calculated for a variety of collection of studies: Systemic Lupus Erythematosus, Parkinson's Disease and Simulated data. Effect size based meta-analysis was then applied to these cases, and the results were compared to traditional and current methods. We could prove that GSEMA outperforms other methodologies when controlling false positives rates while providing meaningful biological interpretations.

GSEMA is implemented as an R package available in CRAN repository at https://cran.r-project.org/web/packages/GSEMA/index.html.

Keywords: Meta-Analysis, Gene Expression, Omics Data Integration, Gene Set Enrichment.

44. The k-sample problem with left-truncated and right-censored data

Adrián Lago

Jacobo de Uña-Álvarez

Juan Carlos Pardo-Fernández

(*Universidade de Vigo*)

The comparison of distributions is a long-studied problem in Statistics that has provided practitioners with many different and useful methods. In this work, we extend the classical Kolmogorov–Smirnov and Cramér–von Mises tests to data subject to left truncation and right censoring. Left truncation usually arises as a result of delayed entry times or cross-sectional sampling schemes, which allow individuals to experience the event of interest before being observed. Under right censoring, the exact failure time of some individuals is not precisely recorded, as some are only known to have failed after a certain point in time. These two phenomena are not only non-exclusive but also appear in many different contexts, such as AIDS studies, clinical trials, demography, and epidemiology. As a consequence, it is important to develop proper statistical techniques that account for them and correct the sampling bias they induce, which makes the observed sample unrepresentative of the underlying population.

We will introduce new Kolmogorov–Smirnov and Cramér–von Mises-type tests for the k-sample problem with left-truncated and right-censored data. Theoretical results will establish their asymptotic behaviour under both the null and alternative hypotheses, proving their consistency. Due to the practical difficulty of applying these theoretical results, a bootstrap resampling scheme will be proposed to approximate the null distribution of the new tests, which will be validated through Monte Carlo simulations. The practical performance of the proposed tests under the alternative hypothesis will also be studied in a simulation study, where the classical log-rank test will be considered for comparison. The new tests will prove to be omnibus in nature, in contrast to the well-known rank-based tests, which, to this day, are the only statistical procedures available to compare distributions under left truncation and right censoring. Finally, a real dataset from a clinical trial will be analysed.

Keywords: Censoring, k-sample problem, truncation.

45. Goodness-of-fit methods with complete and right-censored data: the `GofCens` package

Klaus Langohr

(*Universitat Politècnica de Catalunya-BarcelonaTech; Institut de Recerca i Innovació en Salut (IRIS), UPC*)

Arnau Garcia-Fernández

(*Universitat Politècnica de Catalunya-BarcelonaTech*)

Mireia Besalú

(*Universitat Politècnica de Catalunya-BarcelonaTech; Institut de Recerca i Innovació en Salut (IRIS), UPC*)

Guadalupe Gómez Melis

(*Universitat Politècnica de Catalunya-BarcelonaTech; Institut de Recerca i Innovació en Salut (IRIS), UPC*)

Goodness-of-fit techniques are essential for validating parametric models and ensuring that the underlying assumptions align with the observed data. Traditionally, goodness-of-fit tests have been developed for complete datasets, where all sample observations are fully observed. Widely used tests, such as the Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling tests, measure discrepancies between the empirical and theoretical distribution functions. When data are censored, these statistics can be adapted by using the Kaplan-Meier estimator in place of the empirical distribution function, thus enabling their application to incomplete data. However, computing the corresponding p-values for censored data is challenging and not always straightforward, often requiring advanced resampling techniques or simulation-based approaches.

In this work, we present the R package `GofCens`, a comprehensive tool for performing goodness-of-fit tests and generating diagnostic plots for both complete and right-censored data. The package tackles the computational challenges in the presence of censored data using bootstrap methods to compute p-values for the statistical tests. It supports the most commonly used parametric models in survival analysis, making it a useful tool for researchers and practitioners. We begin by reviewing the theoretical foundations underlying the implemented tests and visualization tools. Following this, we illustrate the practical utility of `GofCens` using real-world survival data from professional basketball players, highlighting its functionality and user-friendly design. We conclude by discussing the current capabilities of the package and identifying areas for future enhancements.

Keywords: Goodness of fit, right-censored data, R package.

46. Unravelling the dynamics of *Diaporthe amygdali* on almond crops: A hierarchical generalised additive modelling approach

Elena Lázaro

(*Departamento de Estadística e Investigación Operativa, Universidad de Valencia*)

Mayte Novellón

(*Centro de Protección Vegetal y Biotecnología, Instituto Valenciano de Investigaciones Agrarias*)

Carolina Francia

(*Instituto Nacional de Investigación Agroepecuaria (INIA) La Estanzuela, Uruguay*)

Mónica Berbegal

(*Instituto Agroforestal Mediterráneo, Universitat Politècnica de València*)

Josep Armengol

(*Instituto Agroforestal Mediterráneo, Universitat Politècnica de València*)

Antonio Vicent

(*Centro de Protección Vegetal y Biotecnología, Instituto Valenciano de Investigaciones Agrarias*)

Understanding the biology of plant pathogens and the epidemiology of their diseases requires experimental protocols and advanced statistical methods. This study focuses on *Diaporthe amygdali*, a fungal pathogen responsible for branch cankers on almond trees, a crop where knowledge of this pathogen remains scarce. To fill this gap, replicated experiments were carried out using two almond-associated isolates of *D. amygdali* (PHAL 4 and PHAL 45) to evaluate independently mycelial growth, pycnidia development, conidia germination, host infection, and lesion progression under different temperature and humidity conditions.

To analyse these issues, we explored Hierarchical Generalised Additive Models (HGAMs) to characterise the non-linear relationships between both covariates and response variables, as well as their interactions in *D. amygdali* dynamics. This approach also allowed to adapt different response distributions and account for variability between isolate-experimental replicates evaluating different random effect specifications. The modelling framework was implemented using the `mgcv` package in R.

We specifically explored different hierarchical structures to quantify variability between experimental replicates: (i) random intercepts to capture baseline differences, (ii) random slopes to allow covariate effects to vary between replicates, and (iii) random interaction effects to explore how the combined influence of covariates varied between replicates. Models were evaluated using AIC, adjusted R^2, deviance, and number of parameters, with AIC as the selection criterion. The selected models were used to predict the response variable for each replicate, assessing overall trends for covariate effects and replicate-specific variability, while confidence intervals were estimated to quantify uncertainty.

Our results highlight the value of HGAMs in understanding the dynamics of *D. amygdali* on almonds, as they account for covariate effects while quantifying replicate-specific variation, improving the generalisation of results. The random replicate effect was significant in all models, with random interaction effect models showing the best AIC. Although statistically significant differences between replicates were observed, the results were interpreted from a biological and epidemiological perspective based on predictions. This framework improves our knowledge of this pathogen under different temperature and humidity conditions, and provides useful information for developing effective management strategies.

Keywords: Hierarchical models, Non-linear regression, Plant pathogen dynamic.

47. Development and validation of prognostic models in phase I oncology clinical trials

Maria Lee Alcober

(*Statistics Unit, Vall d´Hebron Institute of Oncology (VHIO); Department of Statistics and Operations Research, Universitat Politècnica de Catalunya (UPC)*)

Elena Garralda

(*Department of Medical Oncology, Vall d'Hebron University Hospital, Vall d'Hebron Institute of Oncology (VHIO)*)

Guillermo Villacampa

(*Data Science Unit, SOLTI Cancer Group; Statistics Unit, Vall d´Hebron Institute of Oncology (VHIO)*)

Klaus Langohr

(*Department of Statistics and Operations Research, Universitat Politècnica de Catalunya (UPC)*)

Introduction: Selecting patients for phase I clinical trials is increasingly complex, especially with the growing use of immune checkpoint inhibitors (ICIs) and targeted agents (TAs) in novel trial designs. While a minimum life expectancy is often required for enrollment, there is no standardised method for estimating it in clinical practice. Existing prognostic scores focus on overall survival rather than short-term outcomes like three-month survival and often rely on dichotomised variables, which may limit their accuracy. To address these challenges, this study aimed to develop and validate an online prognostic calculator for patients with refractory advanced solid tumors, offering independent applicability for both ICIs and TAs, leveraging continuous variables for better precision, and providing a user-friendly tool to support clinical decision-making, as well as some machine learning-based prognostic models such as random survival forests.

Methods: Overall survival (OS), measured from the first phase I dose to death, was the primary endpoint. We identified key prognostic factors using univariate Cox models without dichotomisation and selected variables via LASSO regression. A stratified Cox model accounted for different baseline hazards by treatment (ICIs vs. TAs), with interactions assessed via ANOVA. Kaplan-Meier curves visualised risk stratification. The Phase I Prognostic Online (PIPO) tool, built as an R Shiny app, used restricted cubic splines to refine survival estimates. Internal validation was conducted via bootstrapping and cross-validation, while external validation used an independent dataset. Performance was assessed through discrimination (C-statistic), calibration (calibration plots, Hosmer-Lemeshow test), overall performance (Brier score), and clinical utility (decision curve analysis).

Results: A total of 921 patients treated at Vall d'Hebron Institute of Oncology (799 training, 122 validation) were included. Six variables independently associated with OS were ECOG, number of metastatic sites, liver metastases, dNLR, albumin, and LDH. PIPO demonstrated good discrimination and calibration in the training set, though external validation showed reduced calibration and clinical utility. Incorporating restricted cubic splines did not significantly improve model performance.

Conclusions: This study provides a replicable framework for developing and validating prognostic models to refine patient selection in phase I trials. Future work will focus on improving calibration and comparing PIPO with other prognostic tools in external validation.

Keywords: prognostic models, phase 1 clinical trials, oncology, model validation.

48. Equations to re-estimate meta-analysis with different adherence scenarios. A simulation study applied to nutritional interventions

Miguel Ángel López-Espinoza

(*Universidad Adventista de Chile*)

David Prieto-Merino

(*Universidad de Alcalá*)

Meta-analyses (MA) of randomized clinical trials (RCTs) provide high-quality evidence of effectiveness due to their controlled methodology. However, in pragmatic RCTs of nutritional interventions, adherence to assigned treatments varies, limiting the accuracy of intention-to-treat (ITT) analysis in measuring true efficacy. As-treated (AT) analysis grouping participants based on actual adherence, could address this issue, but many RCTs report only ITT results. We propose a statistical method to estimate AT effects from ITT-reported data.

Objectives: To develop a statistical method to estimate the efficacy of interventions with Meta-Analysis of RCTs by correcting the ITT estimates assuming different simulated adherence scenarios.

Keywords: Adherence; RCT; Meta-Analysis; Intention to treat analysis; Monte-Carlo simulations.

49. Severe Storms and Extreme Wave Events in the Western Mediterranean Sea

Irene Lopez Mengual

(*Centre of Operations Research, Miguel Hernández University of Elche*)

Carolina Gramcianinov

(*Institute for Coastal Systems Analysis and Modeling, Helmholtz-Zentrum Hereon, Germany*)

Joanna Staneva

(*Institute for Coastal Systems Analysis and Modeling, Helmholtz-Zentrum Hereon, Germany*)

Xavier Barber

(*Centre of Operations Research, Miguel Hernández University of Elche*)

Extreme marine events, such as severe storms and high waves, pose significant risks to coastal and marine activities, particularly in the Western Mediterranean Sea. Understanding these events is crucial due to their increasing frequency and intensity under climate change. This study analyzes extreme wave events and their impacts on coastal ecosystems and aquaculture, focusing on oceanographic variables and their interrelations.

Using **Marine Copernicus datasets**, we examine wind speed, current velocity, and wave height to detect extreme events through a percentile-based threshold method (e.g., 95th or 99th percentile). Events are identified based on their persistence, temporal continuity, and spatial filtering to ensure robust detection. Associated variables such as spectral wave energy, peak period, and mean wave direction are also considered.

To enhance predictive capabilities, we apply a **multivariate Bayesian model**, integrating wave height (wave direction and bathymetry), current velocity (current direction and bathymetry), and **sea level pressure** (wind direction and atmospheric pressure). This approach allows us to better understand the interdependencies among oceanographic variables during extreme events.

Our findings provide insights into long-term trends, environmental influences, and potential hazards to marine-based industries, particularly aquaculture. This study contributes to improving coastal risk assessment and adaptation strategies in the face of increasing climate-induced extreme events.

Keywords: climate change, extreme events, aquaculture.

50. Beyond Leroux and BYM: A Unified Model for Spatial Analysis

Elena García Herbello

(*Master of Biostatistics, University of Valencia*)

Miguel A. Martinez-Beneito

(*Department of Statistics and Operations Research, University of Valencia*)

The Leroux CAR distribution and the BYM model have been among the most widely used tools for the spatial analysis of areal data over the past three decades. Despite their widespread adoption, neither model has demonstrated a clear advantage over the other in terms of model fit. Consequently, the choice between them often depends on personal preference or the software used for implementation.

The key distinction between the Leroux and BYM models lies in how they handle spatial structure. In the Leroux model, the precision matrix is a linear combination of an identity matrix, representing non-structured variability, and the precision matrix of an intrinsic conditional autoregressive (ICAR) distribution. In contrast, the BYM model applies this linear combination to the variance-covariance matrix, utilizing the ICAR variance-covariance structure instead of its precision matrix.

In this work, we propose a comprehensive model that encompasses both the Leroux and BYM models as special cases. Additionally, our model includes three other spatial models that may also be suitable for analyzing areal data. A key feature of our approach is a parameter that allows the model to transition between these different structures, enabling the estimation of posterior probabilities for each submodel. This provides a practical way of quantifying the relative support for the Leroux, BYM, and alternative models in a given dataset.

To illustrate our methodology, we apply the proposed model to mortality data for all causes included in the abridged list of the Spanish National Institute, focusing on the municipalities of the Valencian Region. This analysis provides empirical posterior probabilities for each model in a real-world context. Moreover, our findings offer valuable insights into the characteristics of these spatial models, aiding their selection in future studies.

Keywords: Leroux CAR distribution; BYM model; Gaussian Markov Random Fields.

51. Multiple Correspondence Analysis for federated learning

Maider Mateo-Abad
(*Biogipuzkoa HRI; RICAAPS*)

Jone Renteria
(*Biogipuzkoa HRI*)

Clara Sanchez
(*Osakidetza; Biogipuzkoa HRI; RICAAPS*)

Rafael Rotaeche
(*Osakidetza; Biogipuzkoa HRI; RICAAPS*)

Federated learning is commonly used as a machine learning technique, allowing multiple entities to collaborate within a shared framework without exchanging individual raw data. The use of this approach is increasing due to health data privacy needs. In this context, meta-analysis is widely used to evaluate the results obtained from each entity. However, the complexity of implementation can vary depending on the statistical method used. Nevertheless, some statistical models, such as Principal Component Analysis have been adapted to be used by federated learning. In this work, a proposal for the application of Multiple Correspondence Analysis (MCA) is presented, which is applied to multivariate categorical data coded in the form of an indicator matrix.

Considering k entities or sites, denote D_k as each local dataset with n_k samples, therefore $D = D_1 \cup, \ldots \cup D_k$ is the global dataset, with a $n = n_1, \ldots, n_k$ sample size, and Q variables. Let $\mathbf{Z}$ be asuper-indicator of size $nx\mathbf{J}$, where the columns, $J = \sum_{q=1}^{Q} J_q$, are dummy variables representing thetotal number of categories of all Q variables.

Our proposal is to first, construct from each D_k, frequency tables with the variables to be included in the MCA. Second, share in a global server all these aggregated tables, and then disaggregate them back into individual cases to obtain anonymous D_k datasets. Subsequently, combine all D_k datasets in order to construct the global dataset $D = D_1 \cup, \ldots \cup D_k$. The super-indicator matrix $\mathbf{Z}$ can be constructed to include it in the MCA analysis. Once the results of the analysis are obtained, the coordinates of the resulting lower-dimensional space are stored in such a way that they are transferred to the local datasets, in order to assign them to the corresponding individuals. This process ensures data privacy because no individual data is shared, and it also maintains the statistical power, as the MCA is applied to the global dataset, but without any identification information.

The proposed method is going to be applied in the context of a multicenter project, LOXO (PI22/01665 funded by the ISCIII). Another gap for federated learning approaches is data heterogeneity existing between different health care systems. In order to overcome this, for the LOXO project an agreed common data model (CDM) has been defined first. Therefore, all local datasets will have the same variables and structure.

Keywords: Federated learning, Multiple Correspondence Analysis, health data privacy, LOXO project

52. Using Joint Models for Longitudinal and Time-to-Event Data to Estimate the Causal Effect of Liver Transplantation on Survival in Hepatocellular Carcinoma Patients

Pedro Miranda Afonso

(*Department of Biostatistics, Erasmus University Medical Center*)

Hao Liu

(*Starzl Transplant Institute, University of Pittsburgh Medical Center*)

Michele Molinari

(*J.C. Walter Jr Transplant Center, Houston Methodist Hospital*)

Dimitris Rizopoulos

(*Department of Biostatistics, Erasmus University Medical Center*)

Liver transplantation (LT) is the only curative treatment for selected patients with unresectable hepatocellular carcinoma (HCC). However, due to organ scarcity, patients must often wait for a suitable graft, during which they may become ineligible due to tumour progression or clinical deterioration. A predictive model identifying patients at the highest risk of waitlist dropout and those who would benefit most from LT could improve organ allocation. Transplant-related survival benefit, defined as the additional survival time gained from LT compared to waitlist survival, provides a comprehensive metric to guide allocation. Estimating this causal effect requires addressing the observational nature of transplant data and time-varying confounders. To address these challenges, we developed a joint model for longitudinal and time-to-event data that dynamically predicts individualised transplant-related survival benefit in HCC patients. Unlike alternative approaches, such as the G-formula, structural marginal models andtargeted maximum likelihood estimation, our model makes stronger assumptions about the biomarkermeasurement process but remains non-parametric for competing processes like censoring and visit times.

We analysed data from 7,471 HCC patients listed in the US Scientific Registry for Transplant Recipients (SRTR) between 2012 and 2022, of whom 4,786 received a liver. We developed a Bayesian joint model to associate the pre-transplant trajectories of three well established predictors—the serum level of tumour α-fetoprotein (AFP), the tumour burden score (TBS), and the model for end-stage liver disease (MELD) score—with the risk of death before and after transplantation. Our model predicts a patient's survival probabilities with and without transplantation, which are then used to estimate liver transplant survival benefit. The model is implemented in the freely available R package `JMbayes2`. It provides unbiased estimates of the causal effect of transplantation on individual survival using observational SRTR data without explicitly requiring a model for the transplant assignment mechanism.

Thisprediction model represents an advancement in optimizing liver transplant decisions, promoting fairerorgan allocation, and improving overall survival for waitlisted HCC patients.

Keywords: Conterfactual prediction, joint model, multivariate longitudinal data

53. A Bayesian Shared-Parameter Joint Model for Multiple (Un)Bounded Longitudinal Markers, Competing Risks, and Recurrent Events Using Patient Registry Data

Pedro Miranda Afonso

Dimitris Rizopoulos

Anushka K. Palipana

Emrah Gecili

Cole Brokamp

John P. Clancy

Rhonda D. Szczesniak

Eleni-Rosalina Andrinopoulou

(*Erasmus University Medical Center, Duke University, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cystic Fibrosis Foundation*)

Motivated by a cystic fibrosis study, we developed a Bayesian shared-parameter joint model for multiple continuous (possibly bounded) longitudinal markers, a recurrent event process, and multiple competing terminal events. The model supports various forms of association, discontinuous risk intervals, and both gap and calendar timescales. Our efficient `C++` implementation allows fast fitting even for complex models and large datasets. The model is available in the R package `JMbayes2`.

Keywords: competing risks, multivariate longitudinal data, recurrent events.

54. Understanding driver emotions: A Bayesian network perspective

Dorota Młynarczyk
(*Universitat de València*)

Gabriel Calvo
(*Universitat de València*)

Francisco Palmí-Perales
(*Universitat de València*)

Carmen Armero
(*Universitat de València*)

Virgilio Gómez-Rubio
(*Universidad de Castilla-La Mancha*)

Úrsula Martínez-Iranzo
(*Universitat Politècnica de València*)

Ana de la Torre García
(*Universitat Politècnica de València*)

This study explores the application of Bayesian networks (BN) to the field of driver behaviour modelling (DBM) within a Bayesian reasoning. DBM are computational frameworks designed to analyse and predict the behaviour of drivers taking into account various factors, such as the driver's state of mind and the environmental conditions in which driving takes place. These models aim to improve traffic safety, enhance driver assistance systems, and optimize decision-making in autonomous vehicles. A Bayesian BN is represented by a directed acyclic graph (DAG) that encodes probabilistic dependencies among a set of relevant random variables and their unknown characteristics, defining a joint probability distribution.

In this study, the set of variables was chosen based on an analysis of the factors that may influencedrivers' mental states. The selected variables capture key physiological measures, including averageheart rate, its variability, and breathing frequency. The DAG structure was then determined using theBayesian Information Criterion (BIC).

Keywords: Directed acyclic graph, Driver behavioural model, Driver mental states.

55. Efficient Simulation with Nimble for High-Dimensional Spatial Multivariate Models

Paloma Botella-Rocamora

(*Dirección General de Salud Pública, Conselleria de Sanitat, Generalitat Valenciana*)

Jordi Pérez-Panadés

(*Dirección General de Salud Pública, Conselleria de Sanitat, Generalitat Valenciana*)

Francesc Botella-Quijal

(*Dirección General de Salud Pública, Conselleria de Sanitat, Generalitat Valenciana*)

Miguel A. Martínez-Beneito

(*Departamento de Estadística e I.O., Universitat de València*)

In recent years, multivariate disease mapping models have become a key research field for biostatisticians and spatial epidemiologists. These models enable the joint geographic representation of multiple diseases, incorporating the relationships between them. This approach not only improves risk estimation but also provides additional epidemiological information of great value, such as the correlation matrix between diseases.

For a certain collection of multivariate disease mapping models, known as M-models, it would be practically feasible to include any high number of diseases in a joint analysis. However, technical/computational limitations in their estimation often lead researchers to focus on a specific subset of diseases that share common characteristics or risk factors. This prior selection of diseases hinders the exploration of possible unknown or undocumented relationships and, therefore, the search for new spatially...

The development of new simulation tools, such as Nimble, which allows for simulation customization using MCMC with high computational efficiency, paves the way for posing new challenges, such as the estimation of large-scale spatial multivariate models that simultaneously include a high number of diseases.

In this work, we illustrate the performance and results of this proposal through the multivariate analysis of the outcomes obtained by applying this type of model to a selection of 146 causes of mortality on one hand and premature mortality on the other (73 in men and 73 in women), based on municipal-level mortality data from the Mortality Registry of the Comunitat Valenciana.

Keywords: Epidemiology, Bayesian modeling, Multivariate Disease Mapping.

56. Poisson Regression with Robust Variance vs. Logistic Regression in R: The Case of Maternal Employment Status and Children's Sensory Reactivity in the InProS Project

Rocío Muñoz Sánchez

Irene Campos Sánchez

Empar Casaña Escriche

Desirée Valera Gran

Iris Juárez Leal

Miriam Hurtado Pomares

Eva María Navarrete Muñoz

(*Universidad Miguel Hernández de Elche*)

The InProS project (https://inteo.umh.es/inpros/) is a cross-sectional, population-based study conducted with children aged 3 to 7 years in the province of Alicante. The objective of this study is to compare the estimated results of the association between maternal employment status (actives vs. not active) and the presence of atypical Sensory Reactivity (SR) in children, using logistic regression and Poisson regression with robust variance.

To estimate the odds ratios (OR) accounting for potential confounding factors, the R software packages "base" and "epiDisplay" were used. To estimate the prevalence ratio (PR) considering potential confounders, Poisson regression with robust variance was applied, based on Huber's sandwich method, using the R packages `base, sandwich` and `lmtest`.

SR was assessed using the Spanish adaptation of the Short Sensory Profile (SSP), defining atypical SR as a total SSP score < 155. Maternal employment status was collected through the closed-ended question, "What is your current employment status?" allowing classification into active or not active.

In both cases, the models were adjusted by considering variables that showed a p-value $< 0{,}20$ and/or generated changes greater than 10 % in the magnitude of the association when included in the model. These variables were: maternal education level, age, country of birth, and marital status; paternal education level and employment status; and child-related factors such as sex, age, sleep quality, and adherence to the Mediterranean diet.

The prevalence of atypical SR was 29,4 %. The logistic regression results showed an $OR = 1{,}58$; $95\,\%\ CI$: $0{,}98 - -2{,}55$; $p = 0{,}055$, while the Poisson regression yielded a $PR = 1{,}32$; $95\,\%\ CI : 0{,}99 - -1{,}78$; $p = 0{,}062$. Comparing the results, we observe that the effect size is larger for the OR than for the PR, which could indicate an overestimation bias.

Keywords: Prevalence ratio; Poisson regression with robust variance; Odds ratio; Logistic regression; Sensory reactivity.

57. Second Order Markov multistate models: inference proposal, simulation study and Covid-19 patients application

Josu Najera-Zuloaga

(*Department of Mathematics, University of the Basque Country UPV/EHU*)

Mireia Besalú

(*Departament d'Estadística i Investigació Operativa and Institute for Research and Innovation in Health (IRIS), Universitat Politècnica de Catalunya - BarcelonaTech (UPC)*)

Guadalupe Gómez Melis

(*Departament d'Estadística i Investigació Operativa and Institute for Research and Innovation in Health (IRIS), Universitat Politècnica de Catalunya - BarcelonaTech (UPC)*)

Multistate models provide an effective framework to describe the clinical course of a disease and are widely used in research to model the progression of patients between different health states. Traditional analysis typically rely on a first-order Markov assumption, where the future evolution of the stochastic process depends only on the current state. However, this assumption may not always be appropriate for describing real-world clinical scenarios. To address this limitation, second-order Markov models can be considered, incorporating additional historical information by assuming that the future evolution depends not only on the current state at a given time but also on the state at the preceding time.

Accounting for an additional preceding time increases the dimensionality of the transition probabilities, posing challenges for estimation and inference. To address this issue, extended transition probability matrices for each state have been proposed, and extensions of the Chapman-Kolmogorov equations have been developed to compute n-step second-order transition probabilities.

In this work, we present two non-parametric estimators for these transition probabilities and establish their consistency under the assumption of homogeneity. The first estimator is defined as the ratio of the total number of conditional transitions to the number of individuals at risk, while the second is obtained as the mean of time-specific conditional probability estimates. We derive their asymptotic distributions and provide consistent variance estimators. Additionally, we conduct simulation studies to compare the performance of the proposed estimators under several controlled scenarios.

Finally, we apply the proposed method to a multistate model describing transitions among different health states in more than 3,000 COVID-19 patients hospitalized in the Barcelona metropolitan area during March and April 2022. The multistate model accounts for 14 possible transitions among seven distinct post-admission states. Our findings highlight statistically significant differences in patient progression to non-invasive and invasive mechanical ventilation based on initial pneumonia severity, differences that would remain undetected under a first-order Markov assumption.

Keywords: Multistate Models, Second Order Markov Models, Transition Probabilities estimation.

58. Reconstruct Individual Patient Data from Published Kaplan-Meier Curves to Perform a Meta-Analysis in CAR-T Cell Therapy

Victor Navarro Garcés

(*Statistics Unit, Vall d´Hebron Institute of Oncology (VHIO)*)

Pere Barba

(*Department of Hematology, Vall d'Hebron Barcelona Hospital Campus, Vall d'Hebron Institute of Oncology*)

Sergi Camarillas

(*Department of Hematology, Vall d'Hebron Barcelona Hospital Campus, Vall d'Hebron Institute of Oncology*)

Gloria Iacoboni

(*Department of Hematology, Vall d'Hebron Barcelona Hospital Campus, Vall d'Hebron Institute of Oncology*)

Guillermo Villacampa

(*Statistics Unit, Vall d´Hebron Institute of Oncology (VHIO)*)

Introduction: Meta-analysis of survival outcomes in clinical trials are often limited by the lack of individual patient data (IPD). This is especially relevant in CAR T-cell therapy for relapsed/refractory B-cell acute lymphoblastic leukemia (R/R B-ALL), where most trials are single-arm and do not report hazard ratios. To address this, we reconstructed IPD from published Kaplan-Meier curves to enable statistical comparisons across studies.

Methods: A systematic review was performed and 40 clinical trials were identified evaluating CAR T-cell therapy in R/R B-ALL. The primary endpoint for this study was overall survival (OS), secondary endpoints were minimal residual disease-negative complete remission (MRDneg-CR) and complete remission rate (CRR). Kaplan-Meier curves were digitized and used to reconstruct individual patient-level data. Validation confirmed high alignment between reconstructed and published survival curves. Cox models showed that 4-1BB co-stimulatory domain constructs were associated with better OS and response outcomes compared to CD28-based constructs.

Results: This study highlights the feasibility of reconstructing IPD from Kaplan-Meier curves to enable robust meta-analyses when direct treatment comparisons are unavailable.

Keywords: Reconstructed Kaplan-Meier curves, Survival Analysis, Meta-Analysis, Oncology.

59. Spatial analysis of changes in summer water column stratification in the Mediterranean: potential effects of climate change on finfish aquaculture

Marina Ortiz Navarro
(*University of Alicante*)

Javier Atalah
(*Cawthron Institute*)

Pablo Sanchez-Jerez
(*University of Alicante*)

Xavier Barber
(*Miguel Hernández University of Elche*)

Kilian Toledo Guedes
(*University of Alicante*)

Due to the increase in CO_2 emissions and other greenhouse gases, the oceans are experiencing warming caused by the global phenomenon known as climate change. This effect is particularly evident in enclosed seas such as the Mediterranean Sea. In this context, marine heatwaves have become an increasingly frequent and intense phenomenon, severely impacting the stability of marine ecosystems. This study analyzes the temporal trends in temperature behavior across the Mediterranean within the framework of global warming.

Using data from Copernicus Marine Services, this study examines the effects of climate change on the thermal structure of the water column along a latitudinal and longitudinal gradient in the Mediterranean. The research focuses on 12 locations where marine aquaculture is practiced in different countries (Spain, Tunisia, Greece, Croatia, Turkey, Cyprus, and Egypt). Specifically, it analyzes seasonality across two periods, as well as the assessment of thermal anomalies during the summer season. To determine the sustainability of marine aquaculture systems across the Mediterranean Sea, a climate risk indicator will be applied to these locations using R Studio, alongside the interpretation of water column graphics. The study will also analyze the relationship between thermal trends and environmental and geographic parameters using Bayesian statistics.

A change in the deepening of the 25ºC isotherm over time is detected, influenced by latitude and longitude in relation to oceanographic characteristics. The pattern of thermal changes could significantly impact European aquaculture in the near future, as thermal stress may lead to production issues. Therefore, the aquaculture industry must be able to adapt to changes in marine ecosystem services to maintain sustainable production. Consequently, it is crucial to develop climate change mitigation strategies across the Mediterranean region.

This study forms part of the ThinkInAzul programme (https://thinkinazul.es/) and it was supported by MCIN with funding from the European Union NextGenerationEU (PRTR-C17·1) and Generalitat Valenciana (THINKINAZUL/2021/044-TOWARDS and THINKINAZUL/2021/021-MODESTA).

Keywords: Mediterranean Sea, Aquaculture, climatic risk, climate change.

60. Statistical methods to ensure fair comparisons between groups in real-world data: COVID-19 waves case study

Natalia Pallarés

Cristian Tebé

(*Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (GTP), Badalona, Spain*)

Randomised clinical trials (RCTs) are considered the gold standard for studying the effectiveness of interventions or treatments because randomisation ensures similar baseline characteristics and eliminates confounding that may bias the relationship between exposure and the outcome. Observational studies do not use randomisation, resulting in differences between groups in measured or unmeasured characteristics that could confound this association.

Several statistical methods have been developed to control for confounding in observational studies. In addition to traditional multivariable regression, the range of methods that use the propensity score (PS) to correct for baseline differences between groups includes using the PS as an adjustment covariate, the propensity score matching (PSM), and the inverse probability weighting (IPW).

When examining differences in in-hospital mortality between COVID-19 waves, we need to make the waves comparable in terms of subject's baseline risk to ensure fair comparisons. We will illustrate how these methods can be applied in this scenario using different R packages: Matchit and WeightIt to calculate and apply PSM and IPW, and the alternative MatchThem when working with missing data; cobalt to check covariate balance after matching or weighting and survey to fit models that correctly estimate standard errors when using matching or weighting.

We will compare the results of the different methods, present arguments for and against each one of them, and provide recommendations for comparing treatment or exposure groups in observational studies.

Keywords: COVID-19, propensity score, inverse probability weighting.

61. Comparing the spatial distribution of different football players

Francisco Palmí-Perales

(*Department of Statistics and Operative Research, Universitat de València*)

Virgilio Gómez-Rubio

(*Department of Mathematics, Universidad de Castilla-La Mancha*)

Jesús Lagos

(*Scoutanalyst S.L., Madrid*)

Scouting in football relies on identifying players with similar profiles. This study introduces a novel spatial similarity index to enhance player scouting by leveraging spatial data analysis. Player movements and occupation are quantified, enabling the comparison of spatial distributions through the creation of a spatial index. The proposed index is built using the Lee's spatial cross-correlation statistic, providing a rigorous measure of similarity between players based on their location in the field.

To illustrate the utility of the proposed approach, the study applies the spatial similarity index to data from the Spanish La Liga (2019–2020 season). A clustering analysis of the designed index groups players according to their spatial characteristics, demonstrating the method's ability to classify players into traditional football positions while capturing finer role-specific details. The proposal offers a reproducible and scalable approach for data-driven player evaluation by sharing all the materials to reproduce a toy example.

Keywords: Spatial statistics, similarity index, sports data analysis.

62. Modeling endoscopic performances on detecting advanced neoplasia through cancer screening colonoscopies

Jon Aritz Panera
Noemie Travier
Montse Garcia
Nuria Vives
Laura Sanchez
Marcos Castejon
Berta Casas
Ángeles Coloma
Sónia Abenia
Gemma Binefa

(*Cancer Screening Unit, Catalan Institute of Oncology; Early Detection of Cancer Group, IDIBELL*)

The colorectal cancer screening programme of the Catalan Institute of Oncology invites its target population through a letter every 2 years to undergo a faecal immunological test (FIT). In case of a positive test, they are recommended to undergo a colonoscopy at the endoscopic unit of each patient's reference hospital. The main aim of the study is to evaluate endoscopist performance of follow-up colonoscopies after a positive FIT test. Analysing the differences between endoscopists and endoscopic units is crucial to improve detection rates of premalignant and malignant lesions.

The dataset included colonoscopies performed between 2021 and 2024, excluding repeated colonoscopies and those performed in other hospitals or private institutions. Endoscopists who performed fewer than 30 colonoscopies were also excluded. Missing data were handled by imputation or case exclusion. The dependent variable was binary and indicated the presence of advanced neoplasia, the target lesion of the programme. Independent variables included age, sex, FIT score and colonoscopy quality indicators (Boston number, withdrawal time, caecal intubation rate and bowel preparation solution). Name of endoscopist and endoscopic unit were random effect variables. A generalised linear mixed model (GLMM) with binomial distribution and logit function was used to assess factors associated with advanced neoplasia detection. Validation included comparison with logistic regression (LR), Hosmer-Lemeshow test, Brier score, calibration curves and cross-validation.

Among 82 endoscopists, the GLMM estimated significant variability in advanced neoplasia detection, identifying 28 with a random effect significantly different from zero. Of these, 15 had a positive effect, indicating higher detection rates than expected under the assumed random-effects distribution, while 13 had a negative effect. Endoscopic units contributed less variability, but one of 11 hospitals showed significantly lower performance.

Endoscopists identified with higher and lower random effects could be matched stratified by hospital to work together and learn techniques from the higher performing endoscopist. Further analysis should be done to understand the effects of each endoscopic unit to assess whether their performance is due to hospital materials, endoscopist experience or patient characteristics.

Keywords: Colorectal cancer screening, colonoscopy, mixed models.

63. Surfing the Wave of Data: Statistical Insights from World Surf League Competition Data

Jon Aritz Panera

(*Cancer Screening Unit, Catalan Institute of Oncology, L'Hospitalet de Llobregat, Barcelona, Spain; Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)*)

Lore Zumeta-Olaskoaga

(*Biogipuzkoa Health Research Institute, Donostia-San Sebastián; Department of Physiotherapy, Faculty of Health Sciences, University of Deusto*)

Surfing competitions have been held for decades, beginning with the IPS World Circuit in 1976, followed by the ASP World Tour from 1983 to 2014, and now continuing as the World Surf League (WSL). Since 2008, the WSL has maintained comprehensive records of all professional surfing competitions. These data could serve as a valuable resource for research on athlete performance and competition dynamics. This study explores the statistical analysis of competition data, focusing on identifying performance patterns, evaluating consistency, and detecting trends across seasons and surfers.

Motivated by this opportunity, our work aims to bring statistical rigor to the surfing community.To achieve this, we first collected high-quality surf data using web scraping tools and processed it intoanalysis-ready datasets. We then provide statistical insights to answer key questions about competitivesurfing performance: Do shorter surfers perform better than taller competitors? Is there a bias towardsregular-footed surfers in wave scoring? Is there a strategy that works better in certain waves? Doesheight affect performance in back-side versus front-side maneuvers?

We collected data from the WSL men's and women's Championship Tours, through web scraping:annual competition schedules (2008-2018), each year's championship outcomes, and surfer biographicalinformation. After cleaning and integrating the data, we constructed two final data sets, one for men's andone for women's competitions, consisting of 8625 and 4908 performances, respectively, and 41 variables.Our analyses began with exploratory techniques using descriptive statistics and data visualization. Afterinitial exploration and appropriate data structure, we implemented statistical modeling approaches toanalyze championship outcomes. These included conditional logistic regression and generalized mixedmodels. Preliminary results indicated a competitive advantage for regular-footed surfers, who showedconsistently higher advancement rates compared to their goofy-footed counterparts.

To our knowledge, this is a novel work that presents the application of statistical models to analyzecompetitive surfing using WSL records. Although our findings are exploratory, they reveal some performancepatterns. Incorporating key environmental variables (e.g. wave height, wind conditions, tides) andmore complete surfer biographical records would substantially improve the anayses, providing deeperinsights of the key factors that contribute to a surfer's success in competition.

Keywords: surfing, competition data, performance analysis, time trends, sports statistics.

64. Health services use of a cohort of COPD patients from a pre-pandemic to the SARS-CoV-2 pandemic period

José M. Quintana

(*Unidad de Investigación, Hospital Galdakao-Usansolo, Bizkaia, Spain*)

María J. Legarreta

(*Instituto Biosistemak, Bilbao, Bizkaia, Spain*)

Nere Larrea

(*Instituto Biosistemak, Bilbao, Bizkaia, Spain*)

Irantzu Barrio

(*UPV/EHU, Leioa, Bizkaia, Spain*)

Amaia Aramburu

(*Hospital Galdakao, Servicio de Respiratorio, Bizkaia, Spain*)

Cristobal Esteban

(*Hospital Galdakao, Servicio de Respiratorio, Bizkaia; Instituto Bio Bizkaia, Barakaldo, Bizkaia, Spain*)

Background: SARS-CoV-2 pandemic had negative effects on health services access. We evaluate the differences in health services use among patients with chronic obstructive pulmonary disease (COPD) during the period of 2017–2019 compared to the COVID pandemic period of 2020–2022.

Methods: Cohort of patients recruited from different hospitals who had an admission due to COPD exacerbation. Sociodemographic and clinical data were collected from all participants at 2016. Follow up was performed at 2017–19 and 2020–22 with those who agreed to participate, focusing on their use of health services (number of hospital admissions by any cause, visits to Emergency Room, consultations with primary care physician, nurse, or medical specialist). A sample of patients in the form of paired data was generated where time 1 corresponds to the years 2017-2019 and time 2 of the same patient corresponds to the 2020-2022 period. From these data, multivariate negative binomial regression models were developed for all the number of service usage even data with random effects for patients. Models were adjusted by study period age, Charlson Comorbidity Index, other comorbidities, previous admissions and having a SARS-CoV-2 infection or hospital admission by it on pandemic period 2.

Results: The mean age of the 384 participants was 69.2 years (SD: ±9.8), with men constituting 72.1 % of the sample. We observed a statistically significant reduction in the number of hospital admissions (Risk Ratio (RR):0.71), emergency visits (RR:0.75), and face-to-face visits with primary care doctors (RR:0.51) from the first period to the second period. However, there was no significant change in the number of face-to-face consultations with primary care nurses or pneumologists. Having a SARS-CoV-2 infection or being admitted for it during the second period was associated with an increase in hospital admissions (RR=1.36), emergency visits (RR=1.4), and face-to-face consultations wit

Conclusion: SARS-CoV-2 pandemic had an important negative effect on health care access on patients with COPD. On the one hand, access to the use of most health services in these patients decreased significantly. On the other hand, having had a SARS-COV-2 infection or a hospital admission by it was related to a greater use of these health services.

Keywords: SARS-CoV-2, pandemic, health services, COPD.

65. AdinData: A Dynamic Framework for Ageing Studies Using REDCap and R Software

Jone Renteria

(*Biogipuzkoa HRI*)

Maider Mateo-Abad

(*Biogipuzkoa HRI, Network for Research on Chronicity, Primary Care and Health Promotion (RICAPPS)*)

Kalliopi Vrotsou

(*Biogipuzkoa HRI, Osakidetza, Network for Research on Chronicity, Primary Care and Health Promotion (RICAPPS)*)

Amaia Calvo

(*Biogipuzkoa HRI*)

Ander Matheu

(*Biogipuzkoa HRI*)

Itziar Vergara

(*Biogipuzkoa HRI, Osakidetza, Network for Research on Chronicity, Primary Care and Health Promotion (RICAPPS)*)

The global population is living longer, with the pace of ageing accelerating more rapidly than everbefore. Particularly, developed countries are experiencing an increase in the number of older individualsand healthy ageing presents a significant socioeconomic challenge in the sustainability of universalhealthcare systems. For this reason, research in this field is increasingly advancing. At the BiogipuzkoaHRI, ageing is a cross-cutting research priority, where the interaction between different groups promotesthe study of characteristics and determinants of the ageing process through a multidisciplinary approach.To support these efforts and understand the aging process, it is essential to analyze and manage data fromdiverse sources in an integrated way, which requires robust technical infrastructures.

The AdinData framework aims to consolidate ageing-related information from multiple and diverseresearch lines into a unified system. Its main objective is to produce a comprehensive data platformwhere users can request data through a simple online form via REDCap. The innovation of this frameworkis the integration of diverse tools, including REDCap and R, maximizing the automatization ofthe process from user request to data delivery, with R handling the pre-processing, validation, and postprocessingsteps, ensuring alignment with each petition requirements. This includes the application ofspecific randomization techniques for patient selection based on the characteristics of every request. Thisframework minimizes human interaction through API integration, maximizes efficiency by centralizingall information in one place, and ensures adherence to data protection regulations and the guidelines ofthe Ethics Committee for Research with Medicines of Euskadi (CEIM, for its acronym in Spanish).

The comprehensive, integrated and fully accessible dataset produced within the AdinData frameworkcomprises longitudinal data from diverse sources and research projects. It currently includes 2744participants aged 64 and older from Gipuzkoa Health Centers, 53 % of whom are women, with 45 % of theparticipants providing follow-up information. The dataset integrates determinants of active and healthyaging, fragility studies, and biological samples stored in the Basque Biobank. It also includes standardizedscales administered by healthcare professionals to assess fragility, functionality, health status, socialsupport, and more. Combining information from biological samples with observational data can providesignificant added value in addressing researchers' questions. The inclusion of new data can further enrichthe dataset, making it increasingly comprehensive and valuable for upcoming research projects.

Keywords: ageing, data integration, data management

66. Can mortality prediction be enhanced integrating patient clinical and medical imaging data?

Lander Rodriguez-Idiazabal

(*Basque Center for Applied Mathematics (BCAM)*)

Marta Avalos

(*Bordeaux Population Health Center, University of Bordeaux*)

Jose M. Quintana-Lopez

(*Galdakao-Usansolo Hospital, Osakidetza*)

Irantzu Barrio

(*Department of Mathematics, University of the Basque Country (UPV/EHU)*)

Electronic health records provide a rich source of data like sociodemographics, clinical history, biological markers, and medical images for predicting clinical outcomes. However, integrating these diverse data types poses methodological challenges, often compromising interpretability. This study evaluated the predictive performance of COVID-19 mortality models with different levels of interpretability and data integration.

This study analyzed a retrospective cohort of 5,504 adults who tested positive for SARS-CoV-2and visited the Galdakao-Usansolo hospital emergency department between March 2020 and January2022. The dataset included sociodemographic information, baseline comorbidities, vaccination records,biological data, and mortality status. To address missing biological data, variables with over 40 % missingnesswere excluded, while the remaining missing values were imputed using the k-nearest neighborsmethod. Additionally, frontal chest X-ray images were available for 965 patients, though they lackedlabeled annotations.

The baseline COVID-19 mortality classifier was based on a cost-sensitive Lasso penalized additivelogistic regression. Continuous variables were categorized into quantiles, and 0 degree splines were usedas additive functions. Cross validation was combined with stability selection to calculate the empiricalselection probability for each basis function. This classifier was compared against three alternative models:(i) a convolutional neural network trained on the available X-ray images; (ii) a semi-interpretablemodel combining the baseline classifier with a convolutional neural network; and (iii) a fully connectedneural network which had clinical and biological data combined with the output from a convolutionalneural network as inputs.

Integrating X-ray images and patient data improved mortality prediction in COVID-19 patients.The deep learning model that combined clinical, biological and radiological data achieved the highestpredictive performance, though the semi-interpretable model, which fused the baseline classifier withthe convolutional neural network, demonstrated comparable predictive ability. These findings highlightthe potential of semi-interpretable machine learning approaches and underscore the value of integratingdiverse data sources to enhance predictive modeling and improve healthcare delivery.

Keywords: Machine Learning, COVID-19, Mortality Prediction.

67. Causal inference and positivity assumption violation. Effect on estimation of average treatment effect

Cristobal M. Rodríguez-Leal
(*Universidad Complutense de Madrid*)

Adrián Aurensanz-Crespo
(*Universidad de Zaragoza*)

Jorge Castillo-Mateo
(*Universidad de Zaragoza*)

Jesús Asín
(*Universidad de Zaragoza*)

Rosario Susi
(*Universidad Complutense de Madrid*)

Teresa Pérez
(*Universidad Complutense de Madrid*)

Valid causal estimates rely on four assumptions that must be met: no interference, consistency, exchangeability (conditional exchangeability in case of observational studies), and positivity. Any violation of them can lead to biased estimates. The positivity assumption is defined as the probability of receiving each value of the intervention conditional on confounding covariates being greater than zero. It can be tested empirically. A practical violation of the positivity assumption occurs when some subjects almost always or never receive the intervention. Such a violation can occur for several reasons, one of which is limited overlap in the covariate distribution, as evidenced by poor overlap in the estimated propensity score between treated and untreated subjects. Propensity score is the probability of receiving the intervention conditioned to confounding covariates.

The objective is to quantify the bias of different causal inference techniques, such as inverse probability weighting, propensity score matching, and targeted maximum likelihood estimation, on the estimation of the average treatment effect and the average treatment effect on the treated, for a binary treatment and different scenarios with different degrees of propensity score overlap.

We simulate five scenarios varying the degrees of propensity score overlap, using known data-generating mechanisms and considering heterogeneity effect sizes among treated patients. This allows us to obtain both the true average treatment effects and the average treatment effects on the treated in each scenario. The selected models (inverse probability weighting, various propensity score matching strategies, and targeted maximum likelihood estimation) are used to estimate these effects applying Monte Carlo simulations to assess their performance.

Keywords: Causal inference, positivity assumption, Monte Carlo simulations, propensity score.

68. Dispersion patterns of particulate and dissolved organic matter near aquaculture facilities in the Southeastern Iberian Peninsula

Irene Ruíz García

Marina Ortíz Navarro

Daniel Vidal Barrachina

Pablo Sánchez-Jerez

(*University of Alicante, Spain*)

Fish farms release significant amounts of **particulate and dissolved organic matter** (POM and DOM) into the surrounding marine environment, primarily from uneaten feed, fish waste, and biofouling. In pelagic systems, the dispersion and accumulation of these organic inputs are strongly influenced by local hydrodynamic conditions, particularly currents and turbulence. Strong currents can enhance the dilution and transport of organic matter, reducing localized impacts and facilitating its incorporation into wider biogeochemical cycles. In contrast, weak currents or enclosed areas may lead to organic matter accumulation, potentially causing oxygen depletion, changes in microbial communities, and eutrophication. Understanding the interaction between fish farm emissions and water movement is crucial for mitigating environmental impacts and ensuring the sustainability of aquaculture operations.

A grid of sampling points was established at the cardinal points around an aquaculture facility, both near the installation (200 m) and further away (1000 m). **Water samples** were collected using a Niskin bottle at depths of 12 and 24 meters. Particulate and dissolved organic matter in its nitrogenous forms was analyzed. A current meter was deployed, and the relationship between current intensity and direction was modeled using hierarchical Bayesian models.

The results highlight the need to establish **minimum current thresholds** for the proper dispersion of particulate matter and the importance of incorporating spatial correlations to accurately adjust dispersion models for effective environmental management of aquaculture.

This study forms part of the ThinkInAzul programme (https://thinkinazul.es/) and it was supported by MCIN with funding from the European Union NextGenerationEU (PRTR-C17·I1) and Generalitat Valenciana (THINKINAZUL/ 2021/044-TOWARDS.

Keywords: Aquaculture, particulate organic matter, dissolved organic matter, hydrodynamics, marine environment.

69. Re-randomization in clinical trials: An innovative strategy to improve recruitment and efficiency

Naiara Santos

Guillermo Villacampa

(*SOLTI, Barcelona, Catalonia, Spain*)

Background: The success of clinical development in oncology over the last decade has generated a competitive landscape with many ongoing randomized controlled trials (RCTs). In this context, the accrual capacity becomes one of the most relevant barriers for the completion and success of RCTs. The failure in accrual has several implications: i) decreased statistical power of the study, ii) negative effects on reliability, iii) waste of resources, etc. In this study, we explore the implementation of re-randomization as an innovative strategy to improve patient recruitment while maintaining the integrity of the statistical analysis.

Methods: The re-randomization is a strategy for RCT in which participants can be re-enrolled and re-randomized in multiple treatment episodes, if the inclusion criteria were met. We examined the statistical methodology underlying re-randomization, emphasizing its advantages in terms of sample size efficiency and statistical power. The PATRICIA trial (NCT02448420), a phase II study in breast cancer where patients were randomized 1:1 to receive i) trastuzumab plus palbociclib and endrocrine therapy or ii) trastuzumab plus chemotherapy, was used as a case study.

Results: To evaluate the i) the feasibility of the approach, ii) the potential gain in statistical power, iii) the potential bias on the estimation of treatment effect and iv) the correct control of type I error rates, a simulation-based analysis was conducted. These simulations provide insights into the required sample size and the expected power of the design under different assumptions. Results from these simulations, as well as data from the PATRICIA trial, where 10 % of patients were re-randomized, were summarized.

Conclusions: The re-randomization design can improve the recruitment rate in RCTs, lower trial costs, and shorten the recruitment period under specific circumstances. At the same time, it still provides an unbiased estimate of treatment effect while maintaining correct type I error rates.

Keywords: Randomised Controlled Trials, Re-randomization, Accrual failure.

70. D-Optimal Designs for Ethanol Pharmacokinetics with Multiple Intakes

M. Teresa Santos-Martín

(*Department of Statistics, IUFFyM, University of Salamanca*)

Juan M. Rodríguez-Díaz

(*Department of Statistics, IUFFyM, University of Salamanca*)

Irene Mariñas-Collado

(*Department of Statistics and Operations Research and Mathematics Didactics, University of Oviedo*)

This study addresses the complex dynamics of alcohol elimination in the human body, a crucial aspect in both forensic and clinical contexts. Existing models for estimating blood alcohol concentration (BAC), such as Widmark's model, are often oversimplified by assuming linear elimination kinetics, which limits their practical applicability. To overcome these limitations, a new nonlinear pharmacokinetic model is proposed that explicitly accounts for both absorption and elimination phases of alcohol metabolism, offering a more realistic description of BAC evolution over time.

The model accommodates multiple alcohol intakes, being initially developed for two distinct intake events but easily extendable to more complex drinking patterns. Special attention is given to the accurate estimation of model parameters, highlighting the need for carefully planned experimental designs. In this regard, optimal experimental design (OED) methodologies are employed to determine D-optimal designs under various scenarios, depending on the timing of the second intake.

Furthermore, a sensitivity analysis of the model parameters is performed, and the impact of correlated observations on optimal design is analyzed. Interestingly, a strong linear relationship between the support points of D-optimal designs is identified, which can be exploited to construct quasi-optimal designs. These designs retain high efficiency while being significantly easier to compute, offering a practical tool for experimental planning.

Altogether, this work provides a flexible and robust framework for improving BAC estimation and designing informative experiments in studies of ethanol pharmacokinetics, with potential applications in forensic science, clinical research, and public health.

Keywords: D-optimality, ethanol pharmacokinetics, non-linear models.

71. The Overlooked: Addressing Survivorship Bias in Capture-Recapture Models

Blanca Sarzo

(*Área de Ambiente y Salud, FISABIO (Spain)*)

Ruth King

(*School of Mathematics, University of Edinburgh (UK)*)

Rachel McCrea

(*Department of Mathematics and Statistics, Lancaster University*)

Survivorship (or selection) bias arises within statistical analyses where the observed data are subject to some underlying selection process prior to entry into the sampled data. For example, within capture-recapture studies, a primary selection mechanism is the survival until initial capture time. The common Cormack-Jolly-Seber (CJS) model conditions on the first time an individual is observed, leading to potential survivorship bias. While the issue of survivorship bias has been well studied in many fields there has been little exploration within the capture-recapture framework. In particular, we focus on individual (continuous) random effect CJS models, where it is assumed that individuals have different survival probabilities, specified to be from some common underlying distribution. We discuss the implications of the survivorship bias within the data collection process, and describe a novel modelling approach that accounts for the survivorship bias within an ecologically sensible manner. Using simulated data we demonstrate the significant impact of ignoring the survivorship bias present in the data. We fit the corrected model to a guillemot data set and demonstrate that even with relatively mild selection bias, the individual heterogeneity variability is substantially underestimated when ignoring this survivorship bias.

Keywords: Conditional model; Continuous random effects; Individual heterogeneity.

72. `flowchart`: an `R` package for creating participant flow diagrams integrated with `tidyverse`

Pau Satorra

Cristian Tebé

João Carmezim

Natàlia Pallarès

(*Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP)*)

Background: In health research, a patient flowchart is the best way to show the flow of participants in a study when reporting results, as stated by the CONSORT guideline. There are several packages in R for drawing flowcharts using different approaches, but in general the programming is quite complex and the numbers need to be entered manually or parameterized beforehand, making reproducibility difficult.

Results: We have created a new `R` package called `flowchart` that provides a different approach integrated into the tidyverse framework. It allows you to create many different types of flowcharts in an easy and much more reproducible way because it automatically adapts to the data. The main idea behind the package is to create flowcharts from an initial dataset by combining different basic functions with the pipe operator ($|>$ or $\%>\%$). These functions are highly customizable and allow the user to create reproducible flowcharts in different study settings.

Conclusion: `flowchart` is available in the Comprehensive `R` Archive Network (CRAN) in its current version 0.7.0 (`https://cran.r-project.org/web/packages/flowchart`) with more than 7000 downloads. The package is designed to facilitate the creation of flowcharts with `R`. We will illustrate how users can create reproducible flowcharts for the entire flow of participants in studies in different settings in just a few steps and with a tidy workflow.

Keywords: flowchart, real world data, clinical trials.

73. Bayesian variable selection for the mean inspiratory pression: A comparison between a traditional and a latent variable aproach

Carolina Mulet

(*Rey Juan Carlos University*)

María Eugenia Castellanos

(*Rey Juan Carlos University*)

Gonzalo García-Donato

(*University of Castilla-La Mancha*)

Usually, in applied studies, we are more interested on hypothetical constructs rather than measurable variables. These are commonly referred to as latent variables and we can reconstruct their idiosyncratic behaviour through means of several observable variables or proxies. Our interest relies on the relevance of each latent variable while explaining the response of interest, the mean inspiratory pression.

In this variable selection problem, the definition of the latent variables arises naturally as we find out some regressors which highly correlate with each other. In particular, we are interested on the effect of some individual variables, sex and age, and other non-observed constructs: the antrophometric status, respiratory behaviour and physical condition. We approximate the latent variables via some measures and technical indicators.

From the Bayesian model averaging framework, we propose the grouped Bayes factor, a weighted average of the corresponding Bayes factors, to measure the importance of the latent variable in the variable selection problem. Moreover, we discuss the role of the model prior probability taking into account the relation between the observed indicators. Finally, we compare the methodology proposed with some traditional variable selection procedures, showing the importance of considering the latent variable structure in a problem with these characteristics.

Keywords: Latent variables; Grouped Bayes factor; Model averaging.

74. Bayesian Therapy: A Good Treatment for Potentially Underpowered Trials?

Cristian Tebé
(*Germans Trias i Pujol Research Institute and Hospital*)

Natalia Pallarés
(*University of Barcelona*)

Francesc Escrihuela-Vidal
(*Bellvitge University Hospital, Bellvitge Biomedical Research Institute*)

Sean W X Ong
(*University of Melbourne, University of Toronto*)

Isabel Oriol
(*Bellvitge University Hospital, Bellvitge Biomedical Research Institute*)

Sara Grillo
(*Hospital de la Santa Creu i Sant Pau*)

Miquel Pujol
(*Bellvitge University Hospital, Bellvitge Biomedical Research Institute*)

Jose M Miró
(*Hospital Clinic-IDIBAPS, University of Barcelona, CIBERINFEC, Instituto de Salud Carlos III*)

Steven Y C Tong
(*University of Melbourne*)

Jordi Carratalà
(*Bellvitge University Hospital, Bellvitge Biomedical Research Institute*)

Introduction: Staphylococcus aureus bacteraemia is an important cause of morbidity and mortality.Recently, two randomised clinical trials have evaluated the effect of adding fosfomycin to standardtherapy in adult patients with methicillin-resistant (BACSARM trial) and methicillin-sensitive (SAFO trial)S. aureus bacteraemia. Although differences in the primary endpoints did not reach statisticalsignificance, the use of fosfomycin was associated with a significant reduction in rates of persistentbacteraemia in both studies. We combined the BACSARM and SAFO trial cohorts and conducted aBayesian analysis to estimate the probability of treatment success of the addition of fosfomycin.

Methods: A new database was created with the common individual of all subjects included in theBACSARM and SAFO trials. Two treatment groups were defined: patients receiving combination therapywith fosfomycin versus those receiving daptomycin or cloxacillin monotherapy. The primary endpointwas clinical success at 8 weeks after randomisation, defined as a composite of survival, clinicalimprovement and freedom from recurrence. A hierarchical log-binomial model was used to estimate thecrude and adjusted (including clinical prognostic factors as covariates) estimates of the effect ofcombination therapy from a Bayesian and frequentist perspective. For the Bayesian approach, aminimally informative prior distribution (N (0,5)) centered on no effect was used (RR of 1.0). To examinethe impact of the prior distribution on the primary analysis different scenarios were assessed.

Results: The pooled cohort included 178 patients in combination and 191 in monotherapy. The successrate at 8 weeks was 72.5 % and 65.5 % respectively. In the Bayesian analysis, the success rate ratio (RR)was

1.10 with a 95 % credible interval of 0.97 to 1.26, and a posterior probability of benefit ofcombination therapy versus monotherapy of 91.8 %. In the covariate-adjusted model, the probability ofbenefit was 66.6 % (RR 1.02, 95 %CrI 0.94-1.11). The posterior probability of improvement for secondaryendpoints were 100 % and 98.9 % for persistent bacteremia at day 3 and day 7 respectively, and 81.7 %for mortality at 60 days. In the frequentist analysis the differences observed in the estimates were minimal.

Conclusion: The results obtained using the Bayesian approach were similar to those obtained using thefrequentist approach, but the estimated probability distribution enriched the analysis and clinicaldiscussion. The combination of the two studies suggests that adjuvant treatment with fosfomycin maybe of clinical benefit in patients with S. aureus bacteraemia.

Keywords: Bayesian analysis, fosfomycin, Staphylococcus aureus, bacteraemia, clinical trials.

75. Diagnostic Methods for Generalized Linear Models with Interval-Censored Covariates

Andrea Toloba

Klaus Langohr

Guadalupe Gómez Melis

(*Statistics and Operations Research Department and Institute for Research and Innovation in Health (IRIS), Universitat Politècnica de Catalunya·BarcelonaTech*)

Interval censoring occurs when a random variable is only known to lie within an interval instead of being observed exactly. It is typically found in time-to-event data, but can also occur in other situations. For example, in analytical chemistry, limits of detection (LoD) and quantification (LoQ) may be involved in the quantification process of analytes. This is the case of our motivating study, in which a generalized linear model (GLM) is used to assess the effect of circulating carotenoids on various cardiometabolic parameters, such as glucose levels. The quantity of circulating carotenoids is interval-censored, as it falls within a range determined by the LoD and LoQ of each metabolite from the carotenoids family.

Diagnostic methods for generalized linear models are not easily adapted when one of the covariates is interval-censored. Consider the general expression of a GLM, $g(\mu_i) = \eta_i = \alpha + \beta' x_i + \gamma Z_i$ where $\mu_i = E[Y_i|X_i, Z_i]$ is the mean response, Z_i is interval-censored within $[z_{li}, z_{ri}]$, x_i is a vector of fully observed covariates, and the response Y_i is also fully observed. We denote by $(Y_i, Z_i, ZL_i, ZR_i, X_i)$ the ith component of a sample from (Y, Z, ZL, ZR, X) and use lowercase letters to represent realizations of the random variables. Residuals are typically based on the predicted mean response, $\hat{\mu}_i = g^{-1}(\hat{\alpha} + \hat{\beta}' x_i + \hat{\gamma} Z_i)$, which cannot be directly computed because the value of Z_i is unknown. In fact, the residuals themselves are interval-censored within a range given by $g^{-1}(\hat{\alpha} + \hat{\beta}' x_i + \hat{\gamma} z_{li})$ and $g^{-1}(\hat{\alpha} + \hat{\beta}' x_i + \hat{\gamma} z_{ri})$. Our goal is to properly account for this uncertainty.

We extend the deviance and quantile residuals using Turnbull's nonparametric maximum likelihood estimator of the distribution function of Z. These residuals are studied as a novel approach for assessing the choice of the distribution of the response variable Y. We also propose diagnostic tools to assess whether the selected link function g adequately captures the relationship between the linear predictor η_i and the mean response μ_i. A common approach is to plot the working responses $\zeta_i = \eta(\hat{\mu}_i) + \frac{\partial \eta}{\partial \mu}(\hat{\mu}_i) \cdot (y_i - \hat{\mu}_i)$ against $\eta_i = \eta(\hat{\mu}_i)$, which should look linear if the link function is correctly specified, although we will have to overcome its dependence on $\hat{\mu}_i$. In addition, we adapt the Component-Plus-Residual plots of $r_i = (\zeta_i - \hat{\eta}_i) + \hat{\beta}_j x_j$ against x_j to evaluate if the covariate effects in $\eta_i = \alpha + \beta' x_i + \gamma Z_i$ satisfy the assumed linearity.

Keywords: interval censoring; censored covariate; model diagnosis.

76. Concordance analysis and prognostic value of non-invasive biomarkers using repeated measurements in patients diagnosed with Human Papillomavirus (HPV)-related oropharyngeal cancer

Sara Tous
(*Institut Català d'Oncologia (ICO)-Hospitalet, IDIBELL; CIBERESP*)

Beatriz Quirós
(*Institut Català d'Oncologia (ICO)-Hospitalet, IDIBELL; CIBERESP*)

Miquel Àngel Pavón
(*INCA-lab, ICO-Hospitalet, IDIBELL*)

Beatriz Cirauqui
(*Oncologia Mèdica, B-ARGO, ICO-Badalona, IGTP*)

Anna Sumarroca
(*Otorrinolaringologia, Unitat Funcional de Cap i Coll, Hospital del Mar, IMIM*)

Xavier León
(*Hospital de Sant Pau*)

Ricard Mesía
(*ICO-Badalona; CARE program*)

Lorena Arribas
(*Unitat Funcional de Nutrició Clínica, ICO-Hospitalet*)

Marc Oliva
(*Oncologia Mèdica, ICO-Hospitalet*)

Laia Alemany
(*Institut Català d'Oncologia (ICO)-Hospitalet, IDIBELL; CIBERESP*)

Oropharyngeal cancer (OPC) has a 5-year survival rate of 40–50 %, with little improvement over 30 years. Risk factors include tobacco and alcohol and HPV infection as major carcinogens. HPV-related OPC differ from negative cases in epidemiological, molecular, and clinical aspects, affecting younger non-smokers and non-drinkers patients (pts) with a better prognosis. However, $\sim$20 % of HPV-related OPC pts respond poorly to treatment and should not undergo de-intensification therapies. No biomarkers currently identify these pts and recurrence patterns remain poorly understood.

Prognostic biomarkers have been analysed in tumour tissue, while non-invasive samples need to be further explored. Additionally, genetic alterations in recurrent HPV-related OPC are unclear, both the emergence of non-HPV-related mutations post-treatment and the increase of its aggressiveness.

To address these gaps, we recruited a prospective cohort of OPC pts ($n = 279$) in four hospitals from Barcelona. Informed consent was obtained and oral rinses, tumour swabs, and blood at different time points (pre-, post-treatment, and at one- and two-years post-diagnosis, or upon recurrence) were collected. Viral HPV DNA was analysed from oral rinses and tumour swabs, while anti-E6/E7 HPV16 (E6/E7 HPV16) and Circulating Tumour DNA (ctDNA) from blood. Tumour biopsies were obtained at diagnosis and if recurrence for HPV DNA detection, viral load, and p16INK4A expression. A subgroup also provided stool samples pre-, post-treatment, and two months later, with oral and intestinal microbiome analysis

and dietary/oral hygiene questionnaires. Nutritional, body composition, and treatment-related data was obtained from medical records.

This study aims to: 1) assess concordance at diagnosis of viral detection, viral load and p16INK4a expression in tumours with serological results for E6/E7 HPV16 and ctDNA, and viral detection in oral samples; 2) assess concordance at diagnosis and follow-up of serological results for E6/E7 HPV16 titers using two techniques, Luminex and ELISA; 3) assess concordance at diagnosis and follow-up of E6/E7 HPV16 and ctDNA detection in blood; 4) estimate variations in E6/E7 HPV16 titers and ctDNA detection in blood and viral load in oral samples and blood during pts follow-up in HPV-positive OPC pts; 5) evaluate the prognostic/predictive value of E6/E7 HPV16 titers and ctDNA in blood and viral load in oral samples and blood at baseline, follow-up, and recurrence in HPV-positive OPC pts; 6) characterize oral microbiome from tumour swabs and gut microbiome from stool samples, and evaluate their association with HPV status and viral load in tumours and saliva in OPC pts pre- and post-treatment; and 7) determine nutritional, body composition, and toxicity profiles based on HPV status pre- and post-treatment. We will present the statistical approaches used to answer the aims of the study.

Keywords: Oropharyngeal cancer prognosis, HPV-related biomarkers, repeated measurements.

77. Exploring the association between underreported mental disorders and atmospheric pollution

Leire Unamuno

(*Centre de Recerca Matemàtica, Barcelona*)

Amanda Fernández-Fontelo

(*Universitat Autònoma de Barcelona (UAB)*)

Pedro Puig

(*Centre de Recerca Matemàtica, Barcelona*)

David Moriña

(*Universitat Autònoma de Barcelona (UAB)*)

Recent research highlights a significant link between exposure to environmental pollutants such as ozone (O_3), nitrogen dioxide (NO_2), and particulate matter (PM_1, $PM_{2.5}$, and PM_{10}) and an increased risk of mental health issues, particularly in children and adolescents. These pollutants have been associated with cognitive and emotional difficulties, contributing to behavioral and developmental disorders, ADHD, anxiety, eating disorders, and other mental health conditions. While emerging studies suggest a similar impact on adult mental health, the evidence remains less conclusive due to limitations in national mental health registries and exposure data. These findings emphasize the need for further research and stronger public health policies to mitigate environmental risks, particularly in urban areas where exposure levels are highest.

The Integrated Nested Laplace Approximation (INLA) algorithm provides an efficient deterministic alternative to Markov Chain Monte Carlo (MCMC) for Bayesian inference. This algorithm is specifically designed for latent Gaussian Models (LGMs), offering highly accurate results with a considerably lower computational cost than MCMC.

We are interested in modelling the incidence of mental disorders over the time period of 2017–2022 in each of the 379 basic areas conforming the Catalan public health system. Let X_t represent the actual disease incidence at time t of a given disease, which follows a normal distribution with mean μ and variance σ^2, i.e. $X_t \sim \mathcal{N}(\mu, \sigma^2)$. Nevertheless, the registered incidence Y_t observed in the data is subject to underreporting, leading to a mixture model:

$$Y_t \sim (1-\omega)\cdot\mathcal{N}(\mu,\sigma^2) + \omega\cdot\mathcal{N}(q\mu, q^2\sigma^2)$$

where ω is the probability of underreporting and q is the underreporting intensity.

Each observation y_i can be modeled as a function of predictors x_{ij} and latent functions $f_r(z_{ir})$, with the mean μ_i defined by:

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{k}\beta_j x_{ij} + \sum_{r=1}^{s} f_r(z_{ir})$$

where β_0 is the intercept, $\beta_1,\ldots,\beta_k$ are coefficients for fixed covariates (e.g. pollutant concentrations), z_{ir} allows for temporal and spatial dependencies, and $f_1,\ldots,f_s$ are smooth functions that capture the non-linear effects of covariates.

Keywords: Bayesian inference; Mental health; Underreported data.

78. Classification in the 3D Kendall Shape Space with Applications in Anthropometry

Jorge Valero
(*Instituto de Biomecánica de Valencia (IBV), Universitat Jaume I*)

Vicent Gimeno i Garcia
(*Universitat Jaume I*)

M. Victoria Ibáñez
(*Universitat Jaume I*)

Amelia Simó
(*Universitat Jaume I*)

The development of technology in recent years has made it possible to acquire increasingly larger volumes of data. In particular, advancements in cloud computing and commercial 3D body scanning systems allow for working with 3D point clouds obtained from 3D body scanning, which capture the full complexity of body surfaces.

The increase in both the quantity and complexity of body data has brought new challenges for statistical analysis techniques. However, in most applications of human anthropometry, standard multivariate analyses are performed without considering the curved nature of shape space.

In this work, we propose a novel methodology that extends Linear Discriminant Analysis (LDA) to Kendall's Shape Space for the classification of human body shapes. Our approach adapts LDA to the non-Euclidean geometry of shape data, incorporating parallel transport to improve the estimation of shape variability across different groups. By leveraging shape-aware statistical tools, our method enhances classification performance and provides deeper insights into human body morphology. These advancements have potential applications in ergonomic design, the apparel industry, and medical diagnostics.

Keywords: Statistical shape analysis, Anthropometry, Linear Discriminant Analysis.

79. Surveillance of Acute Gastroenteritis following the 2024 DANA in the Comunitat Valenciana

Jordi Pérez-Panadés
(*Dirección General de Salud Pública, Generalitat Valenciana*)

Paloma Botella-Rocamora
(*Dirección General de Salud Pública, Generalitat Valenciana*)

Juan Juaneda
(*Dirección General de Salud Pública, Generalitat Valenciana*)

Francesc Botella
(*Dirección General de Salud Pública, Generalitat Valenciana*)

Aurora López-Maside
(*Dirección General de Salud Pública, Generalitat Valenciana*)

F. Javier Roig
(*Dirección General de Salud Pública, Generalitat Valenciana*)

Ana Boned
(*Dirección General de Salud Pública, Generalitat Valenciana*)

Bernardo Guzmán
(*Centro de Coordinación de Alertas y Emergencias Sanitarias (CCAES), Ministerio de Sanidad*)

María José Sierra
(*Centro de Coordinación de Alertas y Emergencias Sanitarias (CCAES), Ministerio de Sanidad*)

On October 29, 2024, an Upper-Level Isolated Depression (DANA) struck the Comunitat Valenciana (CV), causing severe damage to infrastructure, essential services, and severely affecting the population. In response, the Public Health Administration of the CV activated a protocol to assess health risks arising from the event, with special attention to infectious diseases. The primary objective of this protocol was to propose mitigation and control measures, including the surveillance of Acute Gastroenteritis (AG) in the affected areas.

A total of 75 municipalities, home to approximately one million inhabitants (20 % of the CV population), were impacted to varying degrees by the DANA. These municipalities belonged to 49 of the 241 Basic Health Zones (BHZ) in the region. For epidemiological surveillance, the territory was divided into two areas: the DANA Zone, which included the affected BHZ, and the Non-DANA Zone, which covered the rest of the region.

Syndromic surveillance of AG in Primary Care within the DANA Zone was conducted by analyzing daily incidence rates based on diagnoses recorded in electronic medical records. The AG consultation rate in the Non-DANA Zone was used as a reference to detect anomalies in consultation patterns.

A hierarchical Bayesian model was applied to assess daily consultation rates in each affected BHZ over the previous two years, comparing them to the reference region. This model incorporated both the historical deviations of each BHZ relative to the Non-DANA Zone and the natural variability of consultation rates over time. Including the consultation rate of the Non-DANA Zone in the model allowed for control of potential seasonal fluctuations in AG incidence, whose intensity and timing vary each year.

Daily thresholds were established based on the AG consultation rate in the Non-DANA Zone and the estimated predictive distribution for each BHZ, defining expected case numbers in each area. These daily

thresholds enabled continuous monitoring of AG consultation rates. Any deviation beyond the credibility limits—whether below (indicating potential underreporting due to healthcare service disruptions) or above (suggesting an anomalous increase in cases)—allowed for the identification of unusual patterns requiring further investigation.

Keywords: Epidemiology, Bayesian modeling.

80. Assessment of the Effectiveness of a Digital Solution for Patients with Advanced Chronic Diseases using Hurdle Models

Ania Gorostiza

(*Instituto de Investigación en Servicios de Salud Biosistemak, B Accelerator Tower (BAT), Bilbao, Spain*)

Borja García-Lorenzo

(*Instituto de Investigación en Servicios de Salud Biosistemak, B Accelerator Tower (BAT), Bilbao, Spain*)

Introduction: Population ageing and medical advances have increased life expectancy of individuals with chronic diseases, including advanced chronic diseases (ACDs). Patients living with ACD often experience temporary or permanent functional limitations, resulting in increased healthcare resource use. Integrated supportive care, particularly through digitally enabled interventions, can help maintain or improve quality of life, promote independence, and optimise healthcare utilisation from an early stage. Within the EU-funded ADLIFE project, an integrated, digitally enabled toolbox has been developed and implemented to support healthcare planning and delivery for ACD patients. This study aimed to evaluate the effectiveness of the ADLIFE intervention in real-life settings, using a reduction in the number of emergency room (ER) visits as a proxy for the appropriateness of care in real-life settings for ACD patients.

Methods: A multicentre, quasi-experimental, non-randomised, unblinded and controlled study was conducted across four pilot sites in Spain, the United Kingdom, Denmark and Israel. Patients over 55 years of age with ACD (e.g. chronic obstructive pulmonary disease and/or chronic heart failure) were recruited in 2023-2024 and followed for 3 to 14 months. The intervention group received the ADLIFE intervention, while the control group received standard of care (SoC). To minimise selection bias due to the non-randomised design and to ensure that they received SoC, the control group was selected retrospectively using propensity score matching. Sociodemographic, clinical and resource use information was collected from electronic health records and the ADLIFE digital platforms.

The primary outcome was defined as the number of ER visits, considered a proxy for intervention effectiveness. A descriptive analysis was first performed, followed by an unadjusted comparison of ER visit rates between groups. Adjusted analysis was then performed to assess the effect of the ADLIFE intervention. Given the discrete nature of the primary variable and the presence of excess of zeros, multivariate hurdle regression models were fitted. These models combined a logistic component, which estimated the odds of having an ER visit, with a zero-truncated count component, which modelled the number of ER visits among individuals attending at least once. Analyses were adjusted for potential confounders, including patient follow-up. Moreover, interaction terms were included to explore heterogeneity of the ADLIFE intervention effect across pilot sites.

Results: The ADLIFE intervention was associated with significant reduction in the likelihood of a first ER visit. However, no significant differences were found in the reduction of the number of subsequent ER among patients who had already used emergency services. The effect varied across pilot sites, suggesting that local characteristics and implementation processes may influence outcomes. Statistically significant differences of the ADLIFE intervention may not have been found in some sites due to limitations in sample size and patient follow-up.

Conclusions: This research provides scientific evidence for a digital solution that provides integrated and personalised care for patients with ACD. The use of hurdle models allowed a comprehensive evaluation of this complex, zero-inflated outcome, contributing to a better understanding of the intervention's impact.

Keywords: Advanced chronic diseases, supportive care, digital health, effectiveness evaluation, hurdle models.

81. Estimating the direct social costs of dementia-related neuropsychiatric symptoms using two-part regression models

Lore Zumeta-Olaskoaga
(*Biogipuzkoa Health Research Institute, Universidad de Deusto*)

Oliver Ibarrondo
(*Osakidetza Basque Health Service, Debagoiena Integrated Health Organization*)

Raúl Del Pozo
(*University of Castilla la Mancha*)

Igor Larrañaga
(*Biosistemak Institute for Health Systems Research*)

Ander Zapiain
(*Department of Social Welfare, Provincial Council of Gipuzkoa*)

Javier Mar
(*Biogipuzkoa Health Research Institute*)

From a social perspective, the behavioral dimension of dementia is particularly significant because it affects mood and behavior through neuropsychiatric symptoms (NPS). These symptoms increase the economic burden and reduce the quality of life for both patients and caregivers. Besides, they lead to greater institutionalization, misuse of medications, and social costs. However, the population impact of NPS is underestimated because of under-recording and separate data collection by Health and Social Services. The lack of integration of social and health data makes it difficult to obtain comprehensive data and to study the social burden of dementia and NPS. Consequently, formal social costs (i.e. direct non-healthcare costs) have only been partially considered in the literature.

In this study, we examine the excess formal social costs associated with dementia-related neuropsychiatric symptoms, using region-wide population data. Specifically, we analyze health and social data on individuals aged 60 and older in Gipuzkoa. We apply two-part models, weighted using entropy balancing, to estimate formal social costs. These models effectively capture the distribution of costs. The first part of the model estimates the marginal effect of social costs being greater than zero, while the second part estimates the expected value of social costs given that they are greater than zero. Combining them yields the average costs for each group and the adjusted excess cost.

Our results indicate that the burden of caring for neuropsychiatric symptoms is greater than previously suggested in the literature. These symptoms more than triple the social costs of dementia due to increased use of residential care and formal coverage, reaching more patients than earlier studies estimated. The higher prevalence of dementia and neuropsychiatric symptoms among individuals with lower socioeconomic status highlights a health inequality, which is partially mitigated by greater use of social benefits.

Keywords: two-part models; dementia; formal social costs.

82. Multivariate analysis of the effects of environmental enrichment in mice: characterising hippocampal plasticity.

Sánchez-Monreal S

(*University of Salamanca, Salamanca, Spain; 2 Institute of Neuroscience of Castilla y León, INCyL, Salamanca, Spain; Institute of Biomedical Research of Salamanca, IBSAL, Salamanca, Spain*)

Cocho T

(*University of Salamanca, Salamanca, Spain; 2 Institute of Neuroscience of Castilla y León, INCyL, Salamanca, Spain; Institute of Biomedical Research of Salamanca, IBSAL, Salamanca, Spain*)

Rodríguez-Iglesias N

(*Achucarro Basque Center for Neuroscience, Vizcaya, Spain; University of the Basque Country EHU/UPV, Vizcaya, Spain*)

Escudero-Solano C

(*University of Salamanca, Salamanca, Spain; 2 Institute of Neuroscience of Castilla y León, INCyL, Salamanca, Spain; Institute of Biomedical Research of Salamanca, IBSAL, Salamanca, Spain*)

de Castro L A

(*University of Salamanca, Salamanca, Spain; 2 Institute of Neuroscience of Castilla y León, INCyL, Salamanca, Spain; Institute of Biomedical Research of Salamanca, IBSAL, Salamanca, Spain*)

Alonso J R

(*University of Salamanca, Salamanca, Spain; 2 Institute of Neuroscience of Castilla y León, INCyL, Salamanca, Spain; Institute of Biomedical Research of Salamanca, IBSAL, Salamanca, Spain*)

Ávila-Zarza C

(*University of Salamanca, Salamanca, Spain; 2 Institute of Neuroscience of Castilla y León, INCyL, Salamanca, Spain; Institute of Biomedical Research of Salamanca, IBSAL, Salamanca, Spain; Department of Statistics, University of Salamanca, Salamanca, Spain*)

Díaz D

(*University of Salamanca, Salamanca, Spain; 2 Institute of Neuroscience of Castilla y León, INCyL, Salamanca, Spain; Institute of Biomedical Research of Salamanca, IBSAL, Salamanca, Spain*)

Valero J

(*University of Salamanca, Salamanca, Spain; 2 Institute of Neuroscience of Castilla y León, INCyL, Salamanca, Spain; Institute of Biomedical Research of Salamanca, IBSAL, Salamanca, Spain*)

The hippocampus, a brain region essential for memory, presents a remarkable capacity for change, known as plasticity. This is supported by the generation of new neurons and the dynamic formation and elimination of synaptic connections, allowing the establishment of circuits that encode information about changing environments. However, the standard laboratory environment for mice is minimally variable. In this study, we investigate how an enriched and dynamic environment—featuring tubes, shelters, and exercise wheels—modulates hippocampal plasticity. We analyse data on the axons of newly generated neurons (mossy fibres), synaptic terminals, and microglia, the cells involved in synaptic remodelling.

Multivariate statistical techniques were employed to explore this plasticity from an integrative perspective. Specifically, dimensionality reduction methods were used to examine the structure of the data, and classification analyses were applied to identify experimental subgroups based on neurobiological features. This multivariate approach enables a more comprehensive characterisation of neuroplastic responses to

environmental stimulation, going beyond the limitations of classical analytical methods traditionally used in the field.

Preliminary findings reveal inter-individual differences that point to the existence of distinct experimental groups, underscoring the value of this approach in uncovering latent structures within complex biological data. This study represents a pioneering application of multivariate methods in the investigation of hippocampal plasticity and opens new avenues for identifying composite neurobiological markers. Future directions include expanding this framework by integrating additional data layers to strengthen its exploratory potential in experimental neuroscience.

Keywords: hippocampal plasticity, multivariate analysis, environmental enrichment.

83. Machine Learning and Gender-Based Violence: What Machine Learning algorithm best predicts Gender-based Violence outcome?

Maria Jose Meisel
Zharick S. Molina
Lihki Rubio
Karen Florez

(*Universidad del Norte, Colombia*)

Gender-based violence is the violence that occurs because of a person's gender. The World Health Organization1 estimates that 1 in every 3 women, around 736 million, has suffered GVB in their lifetime. Latin America has a 25 % lifetime prevalence of women aged 15-49. Colombia in Latin America is no exception. The Attorney General of Colombia2 reported that 3 women every hour, 128 a day, and 47 thousand were victims of violence. This study predicts the outcome of GBV using ML algorithms.

This study uses the Colombian Institute of Health dataset to estimate a hospitalization or death outcome of the GBV event. The dependent variable is the outcome of the event; these may be no severe outcome, hospitalization, or death. The independent variables are age, occupation, social security, social status and social role of the victim, area, place, and scope in which the event occurs, weapon used in the event, mental health actions that take place after the event, protective orders that are emitted, type of event, and sexual violence. The ML algorithms implemented are classification models based on nearest neighbor, Bayes classification models (Gaussian Naive Bayes), logistical regression with Lasso and Ridge penalty, random forest, XGBoost, and support vector machine (classic linear, stochastic gradient descent, and Random Fourier Features).

This study uses stratified K-fold cross-validation to allow a better evaluation of each of the ML algorithms. Synthetic Minority Over-sampling Technique (SMOTE) is used to balance the outcomes in the dependent variable because this shows a great disparity (83.00 % no severe outcome, 16.13 % hospitalization, and 0.008 % death). These algorithms are going to be evaluated with AUC and CPU time in seconds. These algorithms are going to be evaluated with AUC and CPU time in seconds to establish which algorithm best predicts the dependent variable.

Keywords: Machine Learning, Gender-Based Violence.

84. Multivariate estimation approaches for the Network Scale-Up Method

Antía Enríquez Yurrebaso

(*uc3m-Santander Big Data Institute (IBiDat), Universidad Carlos III*)

Rosa E. Lillo

(*uc3m-Santander Big Data Institute (IBiDat), Universidad Carlos III, 2Department of Statistics, Universidad Carlos III*)

Belén Pulido

(*uc3m-Santander Big Data Institute (IBiDat), Universidad Carlos III*)

The Network Scale-Up Method (NSUM) is a relatively new statistical technique for indirectlyestimating the size of hard-to-reach or hidden populations by leveraging individuals' social networks. Rather than surveying members of the hidden group directly, NSUM asks respondents how many peoplethey know and how many of them belong to a specific population of interest. This approach has beenapplied to estimate the prevalence of sensitive or stigmatized groups, such as individuals who have hadabortions, committed suicide, or work in sex industries.

NSUM gained particular relevance during the COVID-19 pandemic, when direct data collectionwas often infeasible. International initiatives like CoronaSurveys demonstrated the method's potentialto estimate infection rates and health-related behaviors across countries, providing timely insights forpublic health responses.

Despite its usefulness, traditional NSUM estimators are typically limited to analyzing one hiddenpopulation at a time and often ignore dependencies between groups. Furthermore, existing robust estimatorsare not designed to handle multivariate data, making them vulnerable to contamination whenmultiple populations are studied simultaneously. To address these limitations, two novel families of estimatorsare introduced. The first employs copulas to capture dependencies between subpopulations, whilethe second leverages statistical depth functions to develop robust multivariate estimators.

These new methods are validated using both synthetic and real datasets. Preliminary findingsfrom a study funded by the Instituto de las Mujeres (Spanish Ministry of Equality) are also presented,applying these enhanced NSUM techniques to investigate public perceptions of gender equality and thedistribution of caregiving responsibilities. Results show that estimators using copulas effectively modelcorrelations between hidden groups, while the robust multivariate estimators offer improved stability andaccuracy in contaminated data environments.

Keywords: Aggregated relational data, multivariate estimators, robust estimators.

6 Sesión Especial: Jóvenes Investigadores/as

85. Evaluation of longitudinal surrogate endpoints: a case study of schizophrenia

Gokce Deliorman

(*Department of Statistics and O.R, Complutense University of Madrid*)

Maria del Carmen Pardo

(*Interdisciplinary Mathematics Institute (IMI), Complutense University of Madrid*)

Ariel Alonso

(*L-BioStat, KU Leuven*)

Geert Molenberghs

(*I-BioStat, Hasselt University*)

In clinical trials, surrogate endpoints provide earlier insights into treatment effectiveness, reducing follow-up time and costs. Most existing methods assess surrogate markers at a single time point, although in many settings they are repeatedly measured over time. In this work, we examine a scenario where both true and surrogate endpoints are measured longitudinally through clinical trial data from schizophrenia studies.

Keywords: Galecki's model, Individual causal association (ICA), Longitudinal outcomes, Surrogate endpoints.

86. Early detection of Porcine Reproductive and Respiratory Syndrome (PRRS) using sow production data

M. Lorena Correa-Abendaño

(*Unitat mixta d'Investigació IRTA-UAB en Sanitat Animal, CReSA*)

Carles Vilalta

(*Institut de Recerca i Tecnologia Agroalimentàries (IRTA), CReSA*)

Amanda Fernández-Fontelo

(*Departament de Matemàtiques, Universitat Autònoma de Barcelona*)

Porcine reproductive and respiratory syndrome (PRRS) is a major swine disease with great economic impact. Effective early detection is crucial for outbreak control, yet current surveillance approaches are limited. This study explores statistical techniques to detect anomalies in pig production data. By analyzing real-world farm data, we identify patterns signaling early disease emergence. Our findings aim to enhance disease monitoring and provide valuable insights for the swine industry.

Keywords: PRRS, disease surveillance, anomaly detection.

87. Optimizing dynamic predictions from joint models for multivariate longitudinal and time-to-event data via super learning approach

Arnau Garcia-Fernández

(*Departament d'Estadística i Investigació Operativa, Institute for Research and Innovation in Health (IRIS), Universitat Politècnica de Catalunya - BarcelonaTech (UPC)*)

Dimitris Rizopoulos

(*Department of Biostatistics, Erasmus MC, Rotterdam*)

Guadalupe Gómez Melis

(*Departament d'Estadística i Investigació Operativa, Institute for Research and Innovation in Health (IRIS), Universitat Politècnica de Catalunya - BarcelonaTech (UPC)*)

Time-to-event and longitudinal data are common in health studies, and joint models (JMs) provide a way to analyze both while allowing dynamic predictions that are updated over time. Applying JMs to multivariate longitudinal data is computationally challenging. Ensemble methods like super learning (SL) combine algorithms to obtain optimal predictions and offer a solution for these problems. This work explores the use of SL for deriving dynamic predictions in multivariate JMs.

Keywords: Joint models, dynamic predictions, super learning.

88. On the evaluation of time-dependent discrimination ability for survival models

Leire Garmendia Bergés

(*BCAM - Basque Center for Applied Mathematics*)

Irantzu Barrio

(*Department of Mathematics, University of the Basque Country UPV/EHU*)

Guadalupe Gómez Melis

(*Statistics and Operations Research Department and Institute for Research and Innovation in Health (IRIS), Universitat Politècnica de Catalunya - BarcelonaTech (UPC)*)

Survival models are often used to predict the course of future individuals. The time-dependent AUC ($AUC(t)$) is commonly used to quantify the ability of a survival model to correctly predict future events at a fixed time t. This work aims to analyze the asymptotic properties of the Conditional Inverse Probability of Censoring Weighting (CIPCW) estimator and to compare its behavior to other $AUC(t)$ estimators under conditions different from those analyzed in the literature.

Keywords: Survival analysis; Discrimination ability; Time-dependent AUC.

89. A Bayesian Shared-Parameter Joint Model for Multiple (Un)Bounded Longitudinal Markers, Competing Risks, and Recurrent Events Using Patient Registry Data

Pedro Miranda Afonso
(*Department of Epidemiology and Biostatistics, Erasmus University Medical Center*)

Dimitris Rizopoulos
(*Department of Epidemiology and Biostatistics, Erasmus University Medical Center*)

Anushka K. Palipana
(*School of Nursing, Duke University*)

Emrah Gecili
(*Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center*)

Cole Brokamp
(*Department of Pediatrics, University of Cincinnati*)

John P. Clancy
(*Cystic Fibrosis Foundation*)

Rhonda D. Szczesniak
(*Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center*)

Eleni-Rosalina Andrinopoulou
(*Department of Epidemiology and Biostatistics, Erasmus University Medical Center*)

Motivated by a cystic fibrosis study, we developed a Bayesian shared-parameter joint model for multiple continuous (possibly bounded) longitudinal markers, a recurrent event process, and multiple competing terminal events. The model supports various forms of association, discontinuous risk intervals, and both gap and calendar timescales. Our efficient `C++` implementation allows fast fitting even for complex models and large datasets. The model is available in the R package `JMbayes2`.

Keywords: competing risks, multivariate longitudinal data, recurrent events.

Índice de autores