Workshop on the Occasion of the 10th Anniversary of GRBIO
January 30th and 31st, 2025

GRBIO - Organizing Committee

29 de gener de 2025

Índex

Presentation

It is both an honor and a privilege for the GRBIO (*Grup de Recerca en Bioinformàtica i Bioestadística*) to celebrate the 10th anniversary of its creation with this special workshop, held on January 30 and 31, 2025, at the *Sala d'Actes of the Facultat de Matemàtiques i Estadística (FME)* at the *Universitat Politècnica de Catalunya (UPC)*.

This workshop has been made possible thanks to the FME, and in particular, its Dean, Prof. Jordi Guàrdia, the Departments of *Estadística i Investigació Operativa* at *UPC* and *Genètica, Microbiologia i Estadística* at *Universitat de Barcelona (UB)*, and to the efforts of the members of GRBIO.

This event provides a unique forum for academics, researchers, and professionals in Biostatistics and Bioinformatics to exchange ideas and experiences, bringing together leading experts in Biostatistics and Bioinformatics at the Catalan, national, and international levels. Biostatistics and Bioinformatics play a vital role in decision-making across diverse fields including Health Sciences and Medicine, Biology and Life Sciences, Agriculture and Environmental Science, Data Science and Artificial Intelligence, Industry Applications, Education and Research. These fields illustrate the interdisciplinary nature of Biostatistics and Bioinformatics, demonstrating their vital role in addressing complex challenges in science and society, and it highlights the strong momentum of these disciplines and their importance in advancing 21st-century society.

GRBIO is a consolidated and funded research group composed of professors from the *Universitat Politècnica de Catalunya-BarcelonaTECH (UPC)* and the *University of Barcelona (UB)*, along with researchers from the *Institut d'Investigació Germans Trias i Pujol (IGTiP)* and the *Vall d'Hebron Institut de Recerca (VHIR)*. It also includes PhD and Master's students from institutions such as *UPC*, *UB*, *Victoria University of Wellington*, *Basque Center for Applied Mathematics (BCAM)*, *VHIR*, and *IGTiP*. Our collaborators include researchers from *UB*, *Universitat de Vic - Universitat Central de Catalunya (UVic-UCC)*, *Universitat Internacional de Catalunya (UIC)*, *Aarhus Universitet (Denmark)*, *Institut Català d'Oncologia (ICO)*, *Servei d'Epidemiologia i Prevenció del Càncer (HUSJR)*, *Duke Molecular Physiology Institute*, and *Alexion Pharma Spain*.

Since its creation 10 years ago, GRBIO has aimed to promote joint research in Biostatistics and Bioinformatics, advancing both applications and the theoretical and computational development of new methodologies. Over the past decade, the group has grown significantly, establishing itself as a national and international reference.

The opening ceremony, chaired by the Dean of the FME, Prof. Jordi Guàrdia,

will be attended by Vice-Rectors for Research from UPC (Prof. Jordi Llorca) and UB (Prof. Jordi García). The closing ceremony, led by Klaus Langohr (GRBIO-UPC) and Conxita Arenas (GRBIO-UB), will count with the director of the Department of Statistics and Operations Research (UPC), Prof. Xavier Tort, and the director of the Department of Genetics, Microbiology and Statistics (UB), Prof. Bru Cormand. The ceremony will be closed with remarks by the Dean of the FME, Prof. Jordi Guàrdia.

We are honored to host three prominent speakers: M. Luz Calle (UVic-UCC), Roderic Guigó (Centre for Genomic Regulation), and Geert Molenberghs (Faculty of Medicine at KU Leuven and Hasselt University). The workshop will feature 15 plenary talks, 7 oral communications from PhD students, and 10 posters presented by researchers from 13 Spanish universities (including 7 from Catalonia), 3 international universities, 10 research centers, and 3 private companies that have collaborated with GRBIO over the years.

We would like to extend our gratitude to the Dean of the FME, Prof. Jordi Guàrdia, for their support. We are also grateful for the presence of the Vice-Rectors for Research, Prof. Jordi Llorca (UPC) and Prof. Jordi García (UB), as well as Prof. Xavier Tort and Prof. Bru Cormand, directors of the departments involved.

A special thanks goes to all GRBIO members, whose dedication made this workshop possible. We are particularly grateful to Sonia Navarro and Nacho Pérez for their exceptional work and commitment from the very beginning of the planning process.

Finally, we want to sincerely thank all participants for sharing their research during these two days. Your contributions are indispensable to the success of this scientific program. We hope you enjoy this celebration of our first decade together.

Guadalupe Gómez Melis (GRBIO coordinator), **Alex Sánchex Pla** (GRBIO-UB coordinator), **Klaus Langohr** (GRBIO-UPC coordinator) and **Conxita Arenas Sola** (GRBIO Scientifc Outreach Coordinator)

Organizers of the Workshop on the occasion of the 10th Anniversary of GRBIO

Programme Day 1

Morning Sessions

Hostess: Nora Amama Ben Hassun
Host: Daniel Fernández Martínez

08:30 – 09:30: Registration

09:30 – 10:00: Opening Ceremony

- **Guadalupe Gómez Melis** (GRBIO, UPC)
- **Jordi Llorca Piqué** (Vice-rector for Research, UPC)
- **Jordi Guàrdia Rúbies** (Dean Facultat de Matemàtiques i Estadística (FME), UPC)
- **Jordi Garcia Fernández** (Vice-rector for Research, UB)
- **Àlex Sánchez Pla** (GRBIO, UB)

10:00 – 10:15: GRBIO Through the Years: A Journey of Research and Collaboration

Presenter: Carles Serrat Piè

10:30 – 11:25: Keynote Speaker

Speaker: Mª Luz Calle Rosingana
Title: Trends in Biostatistics Over the Last Decade
Chair: Klaus Langohr

11:30 – 11:55: Coffee Break

Location: Sala Q + R

12:00 – 12:45: GRBIO PhD Students' Talks

Chair: Marta Bofill Roig

- **Pavla Krotka:** Bias-corrected treatment effect estimators for group-sequential platform trials with non-concurrent controls
- **Andrea Toloba López-Egea:** Likelihood-based approach for handling interval-censored covariates in generalized linear models
- **Laia Egea Cortés:** Partial Ordered Stereotype Model, a New Model for Ordinal Data

---

12:50 – 13:50: GRBIO Friends' Invited Talks

Chair: Mireia Besalú Mayol

- **Carmen Armero Cervera:** Joint Bayesian models for heart failure survival and longitudinal data and how we learned about these models together with GRBIO colleagues
- **Jacobo de Uña Álvarez:** On goodness-of-fit testing with survival data
- **Rosa Lamarca Casado:** Rare diseases challenge: no or insufficient patients in a control arm

---

13:55 – 14:10: Poster's Elevator Speech

14:10 – 14:25: Group Photo

14:25 – 15:15: Lunch

15:15 – 15:55: Posters

Chair: Klaus Langohr and Ferran Reverter Comes

- **Unveiling the Underlying Severity of Multiple Pandemic Indicators** – Manuela Alcañiz, Marc Estevez, Miguel Santolino
- **Development and validation of prognostic scores in phase I oncology clinical trials** – Maria Lee Alcober, Guillermo Villacampa, Klaus Langohr
- **Exploring the genetic overlap between attention-deficit/hyperactivity disorder and migraine** – Pau Carabí-Gassol, Natàlia Llonga, Uxue Zubizarreta-Arruti, Valeria Macias-Chimborazo, Silvia Alemany, Christian Fadeuilhe, Montse Corrales, Vanesa Richarte, Josep Antoni Ramos-Quiroga, Marta Ribasés, Judit Cabana-Dominguez, María Soler Artigas

- **Proximal Algorithms: ISTA and FISTA for L1-Regularized Regressions** – YingHong Chen, Esteban Vegas Lozano, Ferran Reverter Comes

- **Patterns, predictors of recurrence-free survival and prognosis impact of comprehensive genomic profiling in salivary gland cancers: a Spanish multicenter study** – S. Tous, M. Balsa, A. Izquierdo, A. Alay, E. Purqueras, M. Gomà, A. Marí, B. Cirauqui, A. Quer, X. León, N. Basté, D. Azuara, M. Oliva

- **Goodness-of-fit methods for accelerated failure time models** – Arnau Garcia Fernández, Klaus Langohr, Mireia Besalú, Guadalupe Gómez Melis

- **Can AI Effectively Interpret Omics Data in Biomedical Research? The Development of GANGO, BIOFUNCTIONAL and RAG** – Xavi Tarragó, Alejandro Rodríguez, Antonio Monleón

- **Forecasting models for COVID-19: Omicron period** – Nere Lerrea, Dae-Jin Lee, Irantzu Barrio, Eduardo Millán, José M. Quintana, Inmaculada Arostegui

- **Proportionality Index of Parts (PIP) measures the association between taxa in microbiome data** – Juan Jose Egozcue and Vera Pawlowsky-Glahn

- **Early-detection of high-risk patient profiles admitted to hospital with respiratory infections using a multistate mode** – João Carmezim, Cristian Tebé, Natàlia Pallarès, Roger Paredes, Cavan Reilly

---

Afternoon Sessions

Host: Joao Pedro Carmezim Correia
Hostess: Marta Bofill Roig

16:00 – 16:55: Keynote Speaker

Speaker: Roderic Guigó Serra
Title: Bioinformatics in the last decade: between quantitative biology and sequence analysis
Chair: Jordi Ocaña Rebull

17:00 – 18:00: GRBIO Friends' Invited Talks

Chair: Àlex Sánchez Pla

- **Natalia Vilor Tejedor:** Precision Genetic Neurodepidemiology: from risk factors to statistical prediction, prevention and clinical translation

- **Xavier de la Cruz Montserrat:** Breaking the Bottleneck in Genetic Variant Interpretation for Precision Medicine
- **Itziar Irigoien Garbizu:** Functional data analysis and fuzzy classification. Independent concepts or a successful combination?

18:00 – 18:25: Berenar Break

Location: Sala Q + R

18:30 – 19:30: GRBIO Friends' Invited Talks

Chair: Esteban Vegas Lozano

- **Ramón Díaz Uriarte:** An overview of cancer progression and evolutionary accumulation models
- **Sonia Tarazona Campos:** Decoding multi-omic regulatory networks: a regression-based approach
- **Pol Castellano Escuder:** Interpretable multi-omics integration with UMAP embeddings and density-based clustering

---

19:30 – 20:30: GRBINGO

Facilitators: Daniel Fernández Martínez, Maria Lee, and Arnau García

---

Programme Day 2

Morning Sessions

Hostess: Andrea Toloba López-Egea
Host: Cristian Tebé Cordomí

9:00 - 10:00: GRBIO Friends' Invited Talks

Chair: Santiago Perez Hoyos

- **María Durban Reguera**
  Coherent cause-specific mortality forecasting via constrained penalized regression models

- **David Conesa Guillén**
  A computationally efficient procedure for combining ecological datasets by means of sequential consensus inference

- **Martí Casals Toquero**
  The Rise of Sport Analytics: New Opportunities in Research

10:05 - 11:05: GRBIO PhD Students' Talks

Chair: Jordi Cortés Martínez

- **Nora Amama Ben Hassun**
  Development and Evaluation of Metrics for Assessing Synthetic Tabular Data Quality

- **Leire Garmendia Bergés**
  Study of the global AUC(t) for a multi-state model

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

- **Pablo Flores Muñoz**
  An equivalence test to detect functional similarity between feature lists based on the joint enrichment of gene ontology terms

---

11:05 - 11:30: Coffee Break

Location: Sala Q + R

11:30 - 12:30: GRBIO Friends' Invited Talks

Chair: Antoni Miñarro Alonso

- **Josu Najera-Zuloaga**
  Modelling Patient-Reported Outcomes: A case-study of COPD patients
- **María Xosé Rodríguez Álvarez**
  Evaluating the Accuracy of Prognostic Biomarkers in the Presence of External Information
- **Pere Puig Casado**
  Estimating the population size in capture-recapture experiments with right censored data

---

12:35 - 13:30: Keynote Speaker

Speaker: Geert Molenberghs

Title: The applied statistical (data) scientist in a high-profile and societal environment: Past, present, and future

Chair: Guadalupe Gómez Melis

13:30 - 13:45: "Tirant daus, descobrim el GRBIO divulga!"

Presenter: Núria Pérez Álvarez

13:45 - 14:10: Closing Ceremony

- **Klaus Langohr** (GRBIO, UPC)
- **Xavier Tort Martorell** (Director Dpt. Estadística i Investigació Operativa, UPC)
- **Jordi Guàrdia Rúbies** (Dean Facultat de Matemàtiques i Estadística (FME), UPC)
- **Bru Cormand Rifà** (Director Dpt. Genètica, Microbiología i Estadística, UB)
- **Conxita Arenas Sola** (GRBIO, UB)

Abstracts

Trends in Biostatistics Over the Last Decade

30th Jan 2025
10:30h

Malu Calle Rosingana[1]

[1]University of Vic – Central University of Catalunya; Faculty of Sciences, Technology and Engineering; Bioscience Department, 08500 Vic, Spain

Over the past decade, biostatistics has undergone significant advancements, driven by the increasing availability of complex data, the emergence of novel analytical methods, and the growing demand for robust and reproducible results in health sciences. This communication explores key developments in the field and how biostatistics has evolved to meet modern scientific challenges such as, high-dimensionality and data heterogeneity, related to the explosion of big data and the increasing complexity of biomedical research.

Keywords: Biostatistics, Biomedical Research, Data Analysis.

30th Jan 2025
15:45h

Bioinformatics in the last decade: between quantitative biology and sequence analysis

Roderic Guigó Serra[1]

[1] Center for Genomic Regulation, Universitat Pompeu Fabra

Bioinformatics oscillates between sequence analyses, which is largely discrete, and quantitative biology. The initial success of bioinformatics built on the relevant biological information encoded in the sequence of genomes and proteins. Linguistic-related methods, mostly based on string comparisons, become powerful tools to mine this information. As high throughput, highly automated methods to monitor genome activity quantitatively (i.e transcription) became increasingly sophisticated, biostatistical methods became dominant within the field. With the recent advent of Artificial Intelligence methods based on Large Language Models, string based methods are gaining in popularity again, emphasizing both the similarities and the differences between the human and the genome languages.

Keywords: bioinformatics, biostatistics, sequence analysis, quantiative biology.

The applied statistical (data) scientist in a high-profile and societal environment: Past, present, and future

31st Jan 2025
12:40h

Geert Molenberghs[1]

[1]Universiteit Hasselt & KU Leuven, Belgium, I-BioStat

A perspective will be offered on the profession of the biometrician, the biostatistician, and more generally the applied statistical scientist, in a continually and rapidly changing environment. The specifics of working in a multi-disciplinary environment will be discussed, referring to collaboration with agronomists, biologists, epidemiologists, medical professionals, etc. At the same time, interactions with other semi- or fully quantitative fields will be touched upon, such as computational biologists, computer scientists, engineers, etc. The current-day (r)evolution towards data science will be placed against a historical timeline of our field, which saw, over a relatively brief period of just one century, the coming of epidemiology and observational studies, (statistical) genetics, bioinformatics, the omics, big data, data science, data analytics, artificial intelligence, etc. Historical notes related to the international evolution of our field, with particular emphasis on the Hispanic world, will be offered.

Keywords: Applied Statistics; Biometry; Biostatistics; Data Science.

30st Jan 2025
12:00h

Bias-corrected treatment effect estimators for group-sequential platform trials with non-concurrent controls

Pavla Krotka[1], Martin Posch[2], Marta Bofill Roig[1]

[1]Universitat Politècnica de Catalunya/Department of Statistics and Operations Research; [2]Medical University of Vienna/Center for Medical Data Science

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

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

In this work, we examine the performance of the currently available model in group-sequential platform trials and show that it leads to a loss of the type I error rate control and bias in the effect estimators. In addition, we describe how the weight of the non-concurrent controls in the treatment effect estimator is stochastically dependent on the outcome in the non-concurrent controls. Moreover, we will investigate adjusted treatment effect estimators that aim to eliminate or reduce the potential bias and resulting type I error rate inflation. Focusing on a simple platform trial with two experimental treatment arms and a continuous endpoint, we will present results from a simulation study, where we evaluate the performance of the considered approaches and compare them to current methods.

Keywords: Platform trials, Interim analysis, Non-concurrent controls, Statistical inference, Statistical modeling.

Likelihood-based approach for handling interval-censored covariates in generalized linear models

Andrea Toloba[1], Guadalupe Gómez Melis[1], Klaus Langohr[1]

[1]Universitat Politècnica de Catalunya·BarcelonaTech
Department of Statistics and Operations Research

30st Jan 2025
12:00h

The development of methods to address censored covariates has gained significant attention in recent years. Although the problem itself is not new, its presence in real-world data has often been overlooked. Interval-censored covariate data, in particular, is frequently replaced by a single imputed value, which is known to introduce bias and underestimate variance. While recent methods have emerged for handling discrete time-to-event covariates, these approaches are often limited to survival analysis contexts, leaving other applications unaddressed.

In this talk, we shift our focus to analytical chemistry, specifically to data related to the quantification of compounds in mixtures. Compounds are often defined by multiple analytes, each measured via liquid chromatography and subject to analyte-specific detection and quantification limits. This chemical technique results in interval-censored data for the overall quantity of a compound. Our motivating example originates from metabolomics, exploring the association between circulating carotenoids—molecules present in the bloodstream and cardiometabolic health. Advancing research in this area requires fitting generalized linear models for cardiometabolic biomarkers while incorporating interval-censored circulating carotenoid levels as a covariate.

Building on this example, we present an extension of the GEL algorithm[1], which was originally developed for time-to-event interval-censored covariates in linear models. The GEL algorithm is an EM-type method that alternates between estimating the distribution of the censored covariate and maximizing the model's likelihood function. However, like other recent approaches in the literature, it relies heavily on the assumption that the censored covariate has a discrete support, which limits its applicability. Our extension overcomes this limitation by handling interval-censored covariates nonparametrically and regardless of the distribution's support, broadening its usability to a wide range of applications.

[1] Gómez, Espinal and Lagakos (2003) Inference for a linear regression model with an interval-censored covariate. Stat in Med 22(3):409-25

Keywords: regression modeling; censored covariates; interval censoring

30th Jan 2025
12:00h

Partial Ordered Stereotype Model, a New Model for Ordinal Data

Laia Egea-Cortés[1], Daniel Fernández[2], Ivy Liu[1], Richard Arnold[1]

[1]School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand; [1]Centre for Data Science and Artificial Intelligence, Victoria University of Wellington, Wellington, New Zealand; [2]Department of Statistics and Operations Research (DEIO), Universitat Politècnica de Catalunya - BarcelonaTech (UPC)

Ordinal response variables are prevalent in many fields and require specific methods that properly respect the natural ordering of their categories. However, many researchers and practitioners still apply techniques designed for nominal or continuous variables to analyse ordinal data, often treating the response categories as equally spaced when they may not be. This approach can lead to misleading results.

My talk presents the Partial Ordered Stereotype Model (POSM), an extension of the Ordered Stereotype Model (OSM) for ordinal response variables. The OSM does not assume equal-spaced response categories by incorporating score parameters, which specify the potentially unequal distances between adjacent response categories. These parameters reflect the discriminant capability of the covariates, indicating how effectively they can distinguish between response categories. However, different covariates may exhibit distinct discriminant capabilities. The POSM addresses this by allowing different sets of score parameters within the same model, thus capturing the characteristics of each covariate in a single framework. An application of the model using a real-world dataset in aquaculture is included to show the utility and interpretation of the method. Our objective is to identify variables impacting salmon health and assess how these variables differentiate between health levels.

Keywords: Ordinal data, Partial Ordered Stereotype Model, Ordered stereo-type model, Uneven spacing, Aquaculture.

Joint Bayesian models for heart failure survival and longitudinal data and how we learned about these models together with GRBIO colleagues

30th Jan 2025
12:45h

Carmen Armero[1], Jesús Gutierrez[2], Thomas Kneib[3], Javier García Seara[4]

[1]Universitat de València; [2]Universidade de Santiago de Compostela; [3]Georg-August-Universität Göttingen; [4]Hospital Clínico Universitario de Santiago de Compostela

Joint modeling of longitudinal and survival (JM-LS) data allow the inclusion of longitudinal information in survival models, as well as the addition of missing data processes in longitudinal studies. These models are very attractive from a methodological point of view and very valuable in biomedical studies. They were also a point of union between a group of researchers from VABAR (València Bayesian Research Group) and GRBIO, who started a joint task of studying these models: we learned a lot and had more fun. We present a Bayesian JM-LS which accounts for longitudinal continuous information in the unit interval and ordinal longitudinal covariates to learn about competing risk models and discuss their application to a Heart Failure (HF) study where patients underwent cardiac resynchronization therapy.

Keywords: -

30th Jan 2025
12:45h

On goodness-of-fit testing with survival data

Jacobo de Uña-Álvarez[1,2]

[1]Universidade de Vigo, Department of Statistics and Operations Research & SiDOR research group; [2]CITMAga

In this talk I will present a new general strategy for goodness-of-fit testing with survival data. The setting is that of testing for a parametric family of distribution functions when the data are deteriorated due to random censoring and/or random truncation. A key step is the characterization of the null hypothesis through a moment equation which involves the estimation of the observable distribution under both the null and the alternative. A new omnibus test will be proposed, and its theoretical properties will be presented. Particular applications include, but are not limited to, right-censored data, left-, right- or doubly-truncated data, or interval censored data. Advantages with respect to existing methods will be discussed. The finite sample performance of the test will be investigated through simulations. Illustrative real data analyses will be given. This is joint work with Juan Carlos Escanciano.

Keywords: Censoring, Nonparametric Statistics, Specification Tests, Survival Analysis, Truncation.

Rare diseases challenge: no or insufficient patients in a control arm

30th Jan 2025
12:45h

Rosa Lamarca[1]

[1]Alexion Pharmaceuticals/AstraZeneca Rare Disease business unit/Quantitative Sciences

In rare diseases, single-arm, non-randomised, open-label trials are frequently conducted, mainly due to ethical reasons or the study being unfeasible as patients reject to participate. However, there are some inherent limitations in this type of designs, for example, time-to-event endpoints and patient reported outcomes are not interpretable without a control arm in the study. There are other circumstances, where a randomised control trial is doable but the number of subjects in the control arm are insufficient. The use of external data (clinical trial data or real-world data) appears as a way to overcome these limitations and improve the efficiency of clinical trials.

A critical step in bringing external data is to ensure that the external data is comparable to the study population in terms of study entry criteria, in particular to measured baseline prognostic/ confounding variables. Ideally, both external data and study population should be exchangeable with each other. There are several frequentist methodologies to adjust for differences in baseline prognostic/ confounding factors, such as, the propensity scores (Rosenbaum and Rubin, 1983) based on matching, stratification, inverse probability of treatment weights, or covariate adjustment on propensity score methods. These methods balance the prognostic factors, then the comparison of outcomes between the treatment groups yields an unbiased treatment effect estimate, as long as all the confounding variables are included in the propensity score model.

Also, Bayesian methods have been developed to borrow information from external data by creating an informative prior distribution. The prior can be derived based on different approaches such as the meta analytic predictive method.

It is important to note that the type I error may be inflated by incorporating external data as a nonrandomised comparison may introduce bias due to unmeasured confounding covariates. Therefore, simulations should be carried out to evaluate the operating characteristics when including external data.

Regulatory agencies have not ignored this situation and have taken some initiatives and released corresponding guidance with recommendations when designing externally controlled clinical trials. However, the use of external controls is not mature enough yet and interactions with regulatory agencies are advisable at the time of the study design.

Keywords: Rare disease, single-arm clinical trial, external data, propensity scores, Bayesian methods.

30th Jan 2025
16:45h

Precision Genetic Neurodepidemiology: from risk factors to statistical prediction, prevention and clinical translation

Natàlia Vilor Tejedor[1]

[1]Genetic Neuroepidemiology and Biostatistics group, BarcelonaBeta Brain Research Center, Barcelona, Spain; Deparment of Human Genetics, Radboud University Medical Center, Nijmegen, Netherlands; Computational Biology and Health Genomics group, Center for Genomic Regulation, Barcelona, Spain, Spain

This talk will delve into biostatistical strategies advancing the field of genetic neuropidemiology. Recent advancements have enabled more precise identification of genetic and environmental factors, significantly enhancing brain health, risk stratification, disease prediction, and prevention strategies. Key highlights include applying multivariate models to extensive genomic, environmental, and brain imaging datasets, and the assessment and implementation of statistical tools designed for data integration. These methods further emphasize incorporating diversity and sex-specific mechanisms into study populations, bolstering the applicability and accuracy of our findings. The implications of these advances extend beyond improved diagnostic accuracy, paving the way for potential biological pathways that support personalized medicine, prevention, and targeted therapeutic interventions.

Keywords: Genetic neuroepidemiology; multivariate models; data integration; personalized medicine; disease prediction.

Breaking the Bottleneck in Genetic Variant Interpretation for Precision Medicine

30th Jan 2025
16:45h

Xavier de la Cruz[1]
[1]Vall d'Hebron University Hospital, Barcelona

Personalized medicine, a promising branch of modern healthcare, has been made possible by the rapid development of next-generation sequencing (NGS), which has revolutionized genetic diagnostics and provided unprecedented opportunities for tailored treatments. However, the clinical utility of NGS remains constrained by the challenge of interpreting the impact of the genetic variants it uncovers. A significant portion of these variants remains classified as Variants of Uncertain Significance (VUS), undermining their clinical utility and creating anxiety for patients and their families. This situation has driven the development of computational pathogenicity predictors, machine learning tools trained to produce binary classifications—benign or pathogenic—of variants. While these methods have been integrated into clinical workflows, their accuracy and interpretability still fall short of meeting the stringent requirements of medical applications.

In this context, recent years have witnessed a paradigm shift toward continuous prediction models, which aim to provide more precise quantitative assessments of variant impacts on protein function. These approaches leverage a combination of technologies that include data from deep mutational scanning experiments and machine learning techniques. By moving beyond binary labels, continuous predictors hold the promise of elucidating critical aspects of variant effects, such as disease severity and therapeutic response, thereby enhancing their relevance for clinical decision-making in precision medicine.

This talk will explore the current state of methodologies to estimate the impact of protein variants, focusing on an original approach developed in our group to address the problem of using a small amount of protein-specific datasets to generate predictions for any protein, combining regression models and an ensemble-based approach. I will discuss, among other things, the results obtained both in rigorous validation experiments as well as in our participation in the CAGI5 and CAGI6 challenges, comparing our performance with that of other methods in the field.

Keywords: pathogenicity prediction, machine learning, precision medicine, clinical variant annotation, bioinformatics, predictive models in healthcare, genomic data analysis.

30th Jan 2025
16:45h

Functional data analysis and fuzzy classification. Independent concepts or a successful combination?

Itziar Irigoien[1], Concepción Arenas[2]

[1]University of the Basque Country (UPV/EHU), Department of Computation Science and Artificial Intelligence;
[2]University of Barcelona (UB), Statistics Section of the Department of Genetics, Microbiology and Statistics

Nowadays we are increasingly able to collect more complex data, and many challenges in data analysis stem from that complexity. The progression from a single numerical value as the unit of study, to a multivariate vector, then to a functional curve, or even to a skeletal shape representation, illustrates this evolution. In other words, there has been a shift from using large sample sizes in low-dimensional spaces to using relatively small ones in high-dimensional spaces. The perspective offered by functional data analysis (FDA) often provides a framework that allows the analysis of curves, images, or functions in high dimensions overcoming the problem of high dimensionality. For this reason, FDA has started to appear in the computational and bioinformatics literature over the last years. On the other hand, fuzzy classification assigns a degree of membership to each unit, often used in disease diagnosis to classify patients based on medical data and with artificial intelligence techniques to address uncertainty in diagnosis. Using the COVID-19 Raman spectroscopy data set we show the usefulness of combining functional data analysis and the distance-based fuzzy classifier FC-DF highlighting their strengths and limitations.

Keywords: Functional Data, Fuzzy Classification, Depth Function, Distance- based approach.

An overview of cancer progression and evolutionary accumulation models

30th Jan 2025
18:15h

Ramon Diaz-Uriarte[1]

[1]Department of Biochemistry, School of Medicine, Universidad Autónoma de Madrid, and Instituto de Investigaciones Biomédicas Sols-Morreale (IIBM), CSIC-UAM, Madrid, Spain.

Cancer progression and evolutionary accumulation models have been developed to discover dependencies in the irreversible acquisition of binary traits (e.g., mutations) from cross-sectional data. They have been used in computational oncology and virology but also in widely different problems such as malaria progression. Some of these methods have been applied to data with phylogenetic and longitudinal dependencies in questions including tool acquisition in animals and antimicrobial resistance in tuberculosis. Because of their interest, new methods continue to be developed.

These tools have been used to make predictions about future and unobserved states of the system, identify different routes of, and dependencies in, feature acquisition in subsets of the data, and improve patient stratification and survival prediction based on the evolutionary trajectories and denoising of the data. The rich variety of available models increases their utility as markedly different dependency structures can be compared on the same data. These methods also hold promise to help identify therapeutic targets and improve evolutionary-based treatment approaches.

I will first give an overview of the available methods. Then, using fitness landscapes, and discussing the conflation of lines of descent, path of the maximum, and mutational profiles, I will focus on how and why inferences might not be about the processes we intend, in particular under bulk sequencing.

I will comment on major research opportunities, including translational uses, identifying dependencies that derive from frequency-dependent selection, and the relationship of these methods with phylogenetic comparative methods.

Keywords: cancer progression model, fitness landscape, bulk sequencing, evolutionary accumulation model, epistasis.

30th Jan 2025
18:15h

Decoding multi-omic regulatory networks: a regression-based approach.

Sonia Tarazona[1]

[1]Universitat Politècnica de València, Department of Applied Statistics and Operations Research, and Quality

Multi-omic experiments offer an unprecedented opportunity to explore gene expression regulation, providing deep insights into the intricate regulatory mechanisms of biological systems. However, the high dimensionality, heterogeneity, and multicollinearity of multi-omic datasets present significant challenges for statistical modeling and variable selection when inferring regulatory networks. Additionally, most existing tools for multi-omic regulatory network inference either fail to accommodate diverse omic modalities or lack the ability to generate and compare phenotype-specific networks.

To address these limitations, we developed MORE (Multi-Omics Regulation), a novel methodology that leverages regression-based frameworks and advanced variable selection strategies to construct phenotype-specific regulatory networks across any number or type of omic data. MORE integrates prior regulatory knowledge and offers functionalities for systematic comparison of the resulting networks.

We benchmarked MORE against other state-of-the-art tools using simulated datasets and applied it to an ovarian cancer case study. Our results demonstrate the robustness and versatility of MORE in unraveling regulatory mechanisms in complex biological systems, underscoring its potential as a valuable resource for multi-omic data analysis.

Keywords: multi-omics, regression models, variable selection, regulatory networks.

Interpretable multi-omics integration with UMAP embeddings and density-based clustering

30th Jan 2025
18:15h

Pol Castellano-Escuder[1], Derek K. Zachman[1,2], Kevin Han[1], Matthey D. Hirschey[1,3,4]

[1]Duke University School of Medicine, Duke Molecular Physiology Institute; [2]Duke University School of Medicine, Duke Department of Pediatrics/Division of Hematology-Oncology; [3]Duke University School of Medicine, Department of Medicine; [4]Duke University School of Medicine, Department of Pharmacology and Cancer Biology

Integrating high-dimensional multi-omics data is essential for understanding the different layers of biological control. Single-omics methods offer useful insights but often miss the complex relationships between genes, proteins, and metabolites. In this talk, I will present GAUDI (Group Aggregation via UMAP Data Integration), a non-linear, unsupervised method that uses independent UMAP embeddings to analyze multiple data types together. GAUDI reveals relationships across omics layers better than several current methods. It not only clusters samples by their multi-omics profiles but also identifies key features contributing to each cluster, providing clear and interpretable visualizations. I will discuss how GAUDI enables researchers to identify meaningful patterns and potential biomarkers across diverse omics types.

Keywords: Multi-omics, Data integration, UMAP, Clustering, Feature selection.

31st Jan 2025
09:00h

Coherent cause-specific mortality forecasting via constrained penalized regression models

María Durbán[1], Carlos G. Camarda[2]
[1]Universidad Carlos III de Madrid, Department of Statistics
[2]Institut National d'Etudes Demographiques, France

A perspective will be offered on the profession of the biometrician, the biostatistician, and more generally the applied statistical scientist, in a continually and rapidly changing environment. The specifics of working in a multi-disciplinary environment will be discussed, referring to collaboration with agronomists, biologists, epidemiologists, medical professionals, etc. At the same time, interactions with other semi- or fully quantitative fields will be touched upon, such as computational biologists, computer scientists, engineers, etc. The current-day (r)evolution towards data science will be placed against a historical timeline of our field, which saw, over a relatively brief period of just one century, the coming of epidemiology and observational studies, (statistical) genetics, bioinformatics, the omics, big data, data science, data analytics, artificial intelligence, etc. Historical notes related to the international evolution of our field, with particular emphasis on the Hispanic world, will be offered.

Keywords: Applied Statistics; Biometry; Biostatistics; Data Science.

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

31st Jan 2025
09:00h

David Conesa[1], Mario Figueira[1], Antonio López Quílez[1], Iosu Paradinas[2]

[1]Universitat de València/Valencia Bayesian Research Group;
[2]AZTI - Centro de Investigación Marina y Alimentaria

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 three different examples—one simulated and two with real ecological data—highlighting its strengths and limitations in practical ecological and environmental applications.

Keywords: Geostatistics, INLA, Preferential sampling, sequential inference, SPDE.

31st Jan 2025
09:00h

The Rise of Sport Analytics: New Opportunities in Research

Martí Casals Toquero[1]

[1]National Institute of Physical Education of Catalonia (INEFC-UB), Spain

Sports Analytics have rapidly grown, offering new avenues for research and applications in performance improvement, injury prevention, and game strategy. This talk explores the evolution and current impact of Sports Analytics while emphasizing the untapped potential of interdisciplinary networks among statisticians and researchers. Such collaborations offer unique opportunities to tackle complex challenges, innovate methodologies, and uncover new insights. Through practical examples, we will illustrate how data-driven approaches are transforming sports science and highlight the future possibilities this growing field holds for research and application.

Keywords: Sports Analytics, Sports Statistics, Data Science, Injury Prevention, Performance Optimization, Interdisciplinary Research, Statistical Education.

Development and Evaluation of Metrics for Assessing Synthetic Tabular Data Quality

31st Jan 2025
10:00h

Nora Amama Ben Hassun[1], Jordi Cortés Martínez[1], Daniel Fernández[1]

[1]Department of Statistics and Operations Research(DEIO). Universitat Politècnica de Catalunya · BarcelonaTech(UPC), Spain

The growing reluctance to share original datasets and the increasing demand to comply with privacy regulations have motivated the adoption of synthetic data. Synthetic data replicates the statistical properties of the original datasets while ensuring that individual-level information or sensitive variables are not disclosed. However, to effectively evaluate the quality of synthetic data, the development and refinement of validation metrics based is required. This assessment ensures the usability and reliability of synthetic datasets.

This research aims to introduce some existing validation metrics implemented in tools such as the `synthpop` package. The focus is on synthetic tabular data, with an emphasis on showcasing a comprehensive list of validation metrics that hold statistical significance and serve as a foundation for the development of new metrics. To address the challenges of validating synthetic data, the research highlights tailored methodologies for specific domains, such as energy, where there are unique challenges. Synthetic data offers opportunities to accelerate model training while ensuring compliance with privacy regulations. By developing robust metrics, the goal is to provide a practical framework for validating high-quality synthetic datasets that meet the needs of sensitive fields. All these metrics will be illustrated through a case study to highlight their applicability and relevance, ultimately filling a considerable gap in the literature concerning synthetic data validation in the energy sector.

Validation metrics are examined on three key dimensions: resemblance, utility, and privacy. Resemblance metrics evaluate the similarity in the statistical distributions between the synthetic and original datasets. Utility assesses the suitability of synthetic data for specific analytical tasks, such as machine learning or statistical modeling. Privacy, meanwhile, ensure that sensitive information from the original data cannot be reconstructed or identified.

Keywords: Synthetic Tabular Data, Validation Metrics, Statistics, Resemblance, Utility, Privacy.

31st Jan 2025
10:00h

Study of the global AUC(t) for a multi-state model

Leire Garmendia Bergés[1,2], Irantzu Barrio[1,2], Guadalupe Gómez Melis[3]

[1]BCAM, Basque Center for Applied Mathematics; [2]Department of Mathematics, University of the Basque Country UPV/EHU; [3]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya

The motivation for my PhD arises from clinical data from the DIVINE project where patients hospitalized due to COVID-19 are followed through several states. One of the aims of this project was to analyze the evolution of those patients and for that a complex multi-state model (MSM) was designed. This MSM allows us to analyze the risk factors for the different events of interest (e.g. non-invasive mechanical ventilation (*NIMV*), invasive mechanical ventilation (*IMV*), or death) as well as to predict the course of the disease for new patients, but we realized that we didn't know how to analyze its predictive capacity. Therefore, the main objective of my PhD is to evaluate the discriminative ability for MSM, and for that, the area under the time-dependent ROC curve ($AUC(t)$) can be used.

In this work we focus initially in those patients with severe pneumonia who can transition to two competing events: the need for *NIMV* or *IMV*; and we propose an estimator for the global $AUC(t)$ for a competing risk model. Under competing risk models, different estimators can be used to estimate the (partial) $AUC(t)$ of each transition ($AUC_k(t), k = 1, 2$). In this work, we propose an estimator $\widehat{AUC}_{CR}(t)$ for the global AUC(t) ($AUC_{CR}(t)$) for a competing risk model as a weighted sum of $\widehat{AUC}_k(t), k = 1, 2$ with each $AUC_k(t)$ being weighted by the probability of experiencing that event k before time t. We have proved that $\widehat{AUC}_{CR}(t)$ is consistent and asymptotically normal.

Keywords: Multi-state models, discriminative ability, time-dependent AUC.

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

31st Jan 2025
10:00h

Natàlia Pallarès[1], Cristian Tebé[1], Jordi Carratalà[2], Sebastià Videla[3]

[1]Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Badalona, Spain; [2]Department of Infectious Diseases, Bellvitge University Hospital, Barcelona, Spain; [3]Clinical Research Support Area, Department of Clinical Pharmacology, Germans Trias i Pujol University Hospital, Badalona, Spain

Background and objective: From March 2020 to July 2022, 6 waves of the COVID-19 pandemic were registered in Spain. There are several studies comparing different COVID-19 waves but, as far as we know, none of them uses a matching procedure to make patients comparable or accounts for ceiling of care. Our aim is to compare in-hospital mortality across waves in patients with and without ceiling of care at hospital admission.

Methods: Data come from an observational study conducted during four waves of COVID-19 (March 2020-August 2021) in 5 hospitals in Catalonia. Three models were constructed to compare in-hospital mortality by wave: 1) a raw logistic model with only wave as a covariate; 2) a fully clinical adjusted logistic regression model with wave and patient baseline information as covariates and 3) a logistic model with weights obtained from an inverse probability weighting procedure to account for differences in baseline profile between waves. Models were presented stratified by ceiling of care. All analyses were conducted using R software version 4.3.0.

Results: A total of 3982 patients without ceiling of care and 1831 patients with ceiling of care were included. Patients with ceiling of care were, in median, 20 years older than patients without ceiling of care and in-hospital mortality ranged from 5% to 45%. The adjusted odds ratio (OR) of in-hospital mortality in the second wave were 0.57 (95%CI 0.40 to 0.80), in the third 0.56 (95%CI 0.37 to 0.84) and in the fourth 0.34 (95%CI 0.21 to 0.56) compared with the first wave in subjects without ceiling of care. The adjusted odds ratio were significantly lower in the fourth (0.38 95%CI 0.25 to 0.58) wave compared to the first wave in subjects with ceiling of care.

Discussion: The likely impact of the wave on in-hospital mortality differs between patients with and without ceiling of care. In patients without ceiling of care, mortality decreased over time which may be explained by better disease knowledge and management. In ceiling of care, only fourth-wave patients were less likely to die than first-wave patients. In a future infectious disease pandemic, it will be a challenge to improve the management of patients with ceiling of care.

Keywords: Inverse probability weighting, Ceiling of care, COVID-19.

An equivalence test to detect functional similarity between feature lists based on the joint enrichment of gene ontology terms

31st Jan 2025
10:00h

Pablo Flores Muñoz[1,2]

[1]Universitat Politècnica de Catalunya/Faculty of Mathematics and Statistics/Department of Statistics and Operational Research
[2]Escuela Superior Politécnica de Chimborazo (ESPOCH)/Faculty of Sciences

In the current era, omics technologies such as high-throughput experiments have significantly transformed the fields of biology and medicine. These advances enable the generation of large volumes of biological data, such as gene lists, proteins, and other biological features, under different experimental conditions. Although getting this large amount of information represents a breakthrough, it is crucial to develop appropriate statistical methods to analyze and extract knowledge from these data.

In this context, the present study proposes a statistical method based on an equivalence hypothesis test to evaluate biological similarity between feature lists. The central idea is that two or more feature lists can be considered biologically similar if they share a significant proportion of enriched GO terms.

First, the choice of the Sorensen index is justified as an appropriate metric for assessing the dissimilarity of joint enrichment between the lists under comparison. Next, the sampling distribution of this measure is studied both theoretically and through approximation using the Bootstrap method, which proves to be particularly effective when the enrichment level is low. Based on these distributions, an equivalence hypothesis test is developed, along with its corresponding irrelevance threshold, which is less arbitrary than the thresholds commonly used in equivalence approaches.

Furthermore, the R package goSorensen has been developed, published, and is available on the Bioconductor platform. This informatics tool allows for the efficient application of the proposed methodology.

Additionally, a dissimilarity matrix is constructed based on the irrelevance threshold, which defines when two lists are significantly equivalent. This matrix provides an inferential measure of how close or distant the compared lists are from each other. The graphical representation and interpretation of this matrix, such as in an MDS-Biplot, is useful for identifying the GO terms associated with the formation of equivalence between lists.

Finally, it is important to note that the proposed methodology has been rigorously evaluated and applied to real gene lists, with an exhaustive comparison of the results obtained against other similar comparison methods.

Keywords: Equivalence test, feature lists, functional similarity, Gene Ontology, high-throughput experiments.

Modelling Patient-Reported Outcomes: A case-study of COPD patients

31st Jan 2025
11:30h

J. Najera-Zuloaga[1], C. Galán-Arcicollar[2], I. Barrio[1,2], D.-J. Lee[3], I. Arostegui[1,2]

[1]Department of Mathematics, University of the Basque Country UPV/EHU; [2]Basque Center for Applied Mathematics - BCAM; [3]School of Science and Technology, IE University

The World Health Organization defines health as a complete physical, mental, and social well-being and not merely the absence of disease or infirmity. In this sense, patient-reported outcomes (PRO) are becoming primary outcome measurements in observational and experimental studies, as they capture evidence of patients' status that is difficult to evaluate physically, such as pain, quality of life or, satisfaction with care. PRO are usually obtained using item-based questionnaires, assigning scores to each item response and summing the scores across a group of items to create overall scores, usually called dimensions, which decompose the health aspect they are evaluating.

The binomial distribution is the most common candidate when modeling discrete and bounded outcomes, such as PRO dimensions. However, the fact that questionnaire items are answered by the same individuals sets up a correlation structure in the ordinal responses that constitute the final score, which increases the variability beyond the mean-variance structure of the binomial distribution, a property called overdispersion. In fact, PRO scores tend to have skewed distributions, often showing U, J or J-inverse shapes.

In this talk, we are going to present the main contributions of our research group in the field of PRO modeling, from the proposal of an optimal probability distribution to a joint model for the analysis of longitudinal PRO and survival data. Additionally, we will present the most clinically significant results obtained from applying the developed models to a health-related quality of life study in patients with Chronic Obstructive Pulmonary Disease (COPD).

Keywords: PRO, COPD, beta-binomial distribution, longitudinal data, survival, joint modeling.

31st Jan 2025
11:30h

Evaluating the Accuracy of Prognostic Biomarkers in the Presence of External Information

María Xosé Rodríguez Álvarez [1,2], Vanda Inácio[3]

[1]Department of Statistics and Operations Research, Universidade de Vigo, Vigo, Spain; [2]CITMAga, The Galician Centre for Mathematical Research and Technology, Spain; [3] School of Mathematics, University of Edinburgh, Scotland, UK

The receiver operating characteristic (ROC) curve is widely used to assess the accuracy of continuous biomarkers for binary outcomes (e.g., healthy and diseased). However, evaluating the impact of additional patient or environmental information on diagnostic accuracy is also important. Furthermore, studies often focus on prognosis rather than diagnosis, especially in survival analysis, where outcomes evolve over time (e.g., alive and death). To assess the accuracy of continuous prognostic biomarkers for time-varying outcomes, time-dependent extensions of the ROC curve have been proposed.

This work introduces a novel penalised-based estimator of the cumulative-dynamic time-dependent ROC curve, which accounts for the potential modifying effects of covariates on biomarker accuracy. Building on previous approaches, we adopt a modelling framework that considers flexible models for the conditional hazard function and the biomarker, allowing for the accommodation of non-proportional hazards and nonlinear effects through penalised splines, thus addressing the limitations of earlier methods. We apply our method to evaluate the ability of the Global Registry of Acute Coronary Events (GRACE) risk score to predict mortality after discharge in patients who have experienced acute coronary syndrome, and how this ability may vary with left ventricular ejection fraction.

Keywords: acute coronary syndrome, location-scale regression model, piecewise exponential additive model, penalised splines, predictive accuracy, survival analysis.

Estimating the population size in capture-recapture experiments with right censored data

31st Jan 2025
11:30h

Pere Puig Casado[1]

[1]Departament de Matemàtiques, Servei d'Estadística Aplicada (SEA), Universitat Autònoma de Barcelona (UAB), Barcelona, Spain

Capture-recapture methods are commonly used in ecology to estimate animal population sizes and species richness. These methods have become popular, not only in ecology but also in social and medical sciences, to estimate the size of elusive populations such as illegal immigrants, illicit drug users, or people having a drinking problem. The talk will address a new non-parametric approach for estimating the population size when we only know how many animals or individuals were observed once, twice, ... , as well as how many animals or individuals were observed r or more times (right censoring pattern). Similar to the Chao estimator, the method provides a lower bound on population size as well as bootstrap confidence intervals. The particular case of censoring at r=2 will be studied in detail, along with several applications in ecological and social sciences.

Keywords: -

30th Jan 2025
14:45h

Exploring the genetic overlap between attention-deficit/hyperactivity disorder and migraine

Pau Carabí-Gassol[1,2,3], Natàlia Llonga[1,2,3], Uxue Zubizarreta-Arruti[1,2,3], Valeria Macias-Chimborazo[1,3], Silvia Alemany[1,3], Christian Fadeuilhe[1,3], Montse Corrales[1,3], Vanesa Richarte[1,3], Josep Antoni Ramos-Quiroga[1,3], Marta Ribasés[1,2,3], Judit Cabana-Dominguez[1,2,3,4], María Soler Artigas[1,2,3,4]

[1]Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addiction, Vall d'Hebron Research Institute, Universitat Autònoma de Barcelona, Spain; [2]Department of Genetics, Microbiology, and Statistics, Faculty of Biology, Universitat de Barcelona, Spain; [3]Biomedical Network Research Centre on Mental Health (CIBERSAM), Spain; [4]These authors jointly supervised this work

Attention-deficit/hyperactivity disorder (ADHD) is a prevalent neurodevelopmental disorder that emerges in childhood and often persists into adulthood. Migraine is one of the most common neurological disorders, associated with a high morbidity and disability. Previous studies have reported an association between ADHD and migraine, revealing that individuals with both disorders have lower quality of life. Given the high heritability of both disorders, we aim to study their genetic overlap. Using data from the largest genome-wide association meta-analyses to date of ADHD (38,691 cases and 186,843 controls) and migraine (102,084 cases and 771,257 controls), we: (i) estimated their genetic overlap using LDscore and MiXer, (ii) identified genetic variants influencing both traits through PolarMorphism, and (iii) assessed the combined effect of migraine genetic background (in the form of polygenic risk scores, PRS) on headache phenotypes in an ADHD cohort (n=930), with and without including information from variants with effect on both traits and their direction of effect. We confirmed a robust positive genetic correlation (LDSC: rg=0.205; SE= 0.032; P=2.38E-10). Besides, we found that ADHD is much more polygenic than migraine (7729 variants, SE=363 and 1731 variants, SE=91, respectively), and both traits share 951 variants, most of them with concordant direction of effect (84%). Cross-trait analysis identified 29 genetic loci, 20 of them newly related to ADHD and migraine. Variants were separated in those with a consistent and inconsistent direction of effect between ADHD and migraine (concordant and discordant, respectively). Each set mapped to 83 and 48 genes, respectively, with no overlap between them. Concordant variants were enriched for traits related to the immune system and inflammatory diseases. Whereas, genes mapped by discordant variants were enriched for cardiovascular and blood pressure traits, neurological and sensory conditions or mental health disorders. In ADHD cases, the PRS for migraine including all genetic variants (n=1,030,039) was associated with childhood headache (OR=1.246,95%CI=1.073,1.447; p=3.88E-03). When considering the PRS constructed only with concordant variants (n=268,823) the association remained (OR=1.236,95%CI=1.083,1.411; p=1.68E-03). However when considering only those with discordant effects (n=257,858) there was no association (OR=1.019,95%CI=0.894,1.163; p=0.773). Our findings suggest that although ADHD and migraine differ in their genetic architecture, they share a substantial genetic background. Furthermore, separating variants according to their direction of effect may be a valuable strategy to identify specific mechanisms and to refine the selection of variants for PRS in the context of comorbid disorders.

Keywords: ADHD, migraine, GWAS, cross-trait analysis, polygenic risk score.

30th Jan 2025
14:45h

Development and validation of prognostic scores in phase I oncology clinical trials

Maria Lee Alcober[1,2], Guillermo Villacampa[1], Klaus Langohr[2]

[1]Statistics Unit, Vall d'Hebron Institute of Oncology (VHIO), Barcelona; [2]Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya (UPC), Barcelona

Accurate patient selection for phase I oncology trials is critical for improving trial outcomes and advancing drug development. This study aims to develop and validate prognostic models and associated risk scores to better identify oncology patients who may benefit from early-phase clinical trials.

A total of 921 patients treated at the Vall d'Hebron Institute of Oncology were included in this study (799 in the training cohort and 122 in the validation cohort). To develop prognostic models, we applied: i) Cox proportional hazards models enhanced with restricted cubic splines to address non-linearity, and ii) machine learning techniques such as decision trees and random survival forests to capture complex interactions. Risk scores derived from these models provide interpretable summaries of patient risk profiles, facilitating practical clinical use.

Internal validation employed bootstrapping and cross-validation to ensure model robustness, while external validation assessed generaliability using an independent dataset. Model performance was evaluated through discrimination (using the C-statistic), calibration (calibration plots and the Hosmer-Lemeshow test), and clinical utility (decision curve analysis). Initial results indicate that internal validation consistently outperformed external validation across all performance metrics, particularly in calibration. Among the models, random survival forests achieved the highest C-statistic, demonstrating superior discrimination. Conversely, incorporating restricted cubic splines into the Cox proportional hazards model did not notably improve its C-statistic.

This work offers a replicable framework for deriving and validating risk scores that enhance precision in patient selection for phase I trials. Future efforts will focus on formalising calibration methods and comparing these models and scores with other published prognostic tools using external validation.

Keywords: Prognostic models, survival analysis, Cox proportional hazards model, restricted cubic splines, random survival forests, decision curve analysis.

Proximal Algorithms: ISTA and FISTA for L1-Regularized Regressions

30th Jan 2025
14:45h

YingHong Chen[1], Esteban Vegas Lozano[1], Ferran Reverter Comes[1]

[1]Department of Genetics, Microbiology and Statistics, Faculty of Biology, Universitat de Barcelona, Barcelona, Spain

The subgradient method and the proximal operator method are techniques primarily used for non differentiable optimization functions which are widely found in the machine learning area. A notable example is predictive models with non-differentiable regularization terms, such as L1 penalties, which are essential for variable selection in models. Within the proximal operator framework, we have developed a R package ProxReg that enables the fitting of regression and classification models for both binary and multiclass responses with L1 regularization. Among the proximal algorithms for L1 regularized regression, ISTA (iterative shrinkage-thresholding algorithm) is well regarded for its simplicity and is suitable for solving large-scale problems. Meanwhile FISTA (Fast shrinkage thresholding algorithm) is an accelerated version of ISTA with a significantly better convergence rate while maintaining the ISTA's computational simplicity. Our package implements both algorithms that allows users to choose the most efficient option depending on their needs. There will be pre- sented a comprehensive comparison of ISTA and FISTA in terms of convergence rate and efficiency. Additionally, the package is benchmarked against the widely used glmnet package, which employs a coordinate gradient descent algorithm for regularized regression. Application examples, including Lasso regression in computer vision, biology, and finance, are used to illustrate the comparative analysis.

Keywords: Proximal Algorithms, Machine Learning, Iterative Shrinkage-Thresholding Algorithm, Lasso Regression.

Patterns, predictors of recurrence-free survival and prognosis impact of comprehensive genomic profiling in salivary gland cancers: a Spanish multicenter study

30th Jan 2025
14:45h

S. Tous[1], M. Balsa[2], A. Izquierdo[3], A. Alay[4], E. Purqueras[5], M. Gomà[5], A. Marí[6], B. Cirauqui[7], A. Quer[8], X. León[9], N. Basté[10], D. Azuara[11], M. Oliva[1]

[1]Unit of Molecular Epidemiology and Genetics/Cancer Epidemiology Research Program/Institut Català d'Oncologia (ICO)-Hospitalet (H); [2]Department of Medical Oncology/ICO-Hospitalet; [3]Department of Maxillofacial Surgery/Hospital Germans Tries i Pujol (HTP); [4]Unit of Bioinformatics for Precision Oncology/ICO-H; [5]Department of Pathology/Hospital Universitari de Bellvitge (HUB); [6]Department of Maxillofacial Surgery/HUB; [7]Department of Medical Oncology/ICO-Badalona; [8]Department of Pathology/HTP; [9]Department of Otorrhinolaringology/Hospital Santa Creu i Sant Pau; [10]Department of Medical Oncology/Hospital Clínic; [11]Molecular Diagnostics Laboratory/Heredetary Cancer Program/ICO-H

Salivary gland cancers (SGC) include multiple histologies with variable prognosis and up to 40% will recur despite curative treatment. Lukovic's score (LS), based on clinicopathological variables, predicts the risk (high vs low) of distant metastasis (DM). This study analyzed recurrence patterns, predictive factors for recurrence-free survival (RFS), and the prognostic impact of clinical data, and comprehensive genomic profiling (CGP) in a Spanish multicenter cohort of SGC patients (pts).

Clinical data from 142 newly diagnosed SGC pts treated with curative intent between 2000 and 2020 at five Head and Neck Cancer Spanish institutions was analyzed. Tumor samples at diagnosis underwent CGP using the Roche AVENIO Tumor Tissue CGP Kit detecting small variants in 324 genes, copy number alterations, fusions, and tumor mutational burden to identify actionable alterations (AA+). Recurrence rates (RR) by histology and median RFS (mRFS) were calculated using the Kaplan-Meier method. Adjusted sub-hazard ratios (aSHR) for local (LM) and distant (DM) recurrences were estimated using competing risk analysis, including stage, histology, LS, and AA+.

Patients's median age was 59 years, being 52% male. Histologies included adenoid cystic (AC, 30%) , mucoepidermoid (MEC, 25%), ductal (DC, 23%), and acinic cell (ACC, 23%). High LS risk was assigned to 63% pts. Overall, RR was 45% (local 28%, distant 54%, both 18%), varying by histology (AC/DC vs MEC/ACC, $p<0.01$), mostly due to DM. Median follow-up was 4y (1-6). mRFS was significantly higher in ACC (13y) and MEC (11y) compared to ACC (8y) and DC (5y) ($p<0.01$). High LS and advanced stage significantly reduced DM-RFS (aSHR: 4.8 [1.3,17] and 3.17 [1.4,7.4], respectively). Among 105 CGP-analyzed samples, 43% had AA+, being more common in DC (83%), ACC (41%) and MEC (36%) compared to AC (8%) ($p<0.001$). High LS risk correlated with higher AA+ rates (54% vs 20%, $p<0.01$).

This study validates LS as a major predictor for DM and highlights the prognostic value of AA+ as the potential for genotype-matched therapies in this setting.

Keywords: Competing Risk Analysis; Sub-Hazard Ratio; Sequencing; Salivary Gland Cancer, Recurrence.

Goodness-of-fit methods for accelerated failure time models

30st Jan 2025
14:45h

Arnau Garcia Fernández[1], Klaus Langohr [1], Mireia Besalú [1], Guadalupe Gómez Melis[1]

[1] Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya

Accelerated Failure Time Models (AFTMs) are a class of survival models widely used to analyze time-to-event data in various fields, including biomedical research, engineering, and economics. Unlike proportional hazards models, AFTMs directly model the effect of covariates on the survival time, assuming that these covariates accelerate or decelerate the life expectancy of subjects. A critical aspect of AFTMs is the choice of the underlying distribution for the survival times, which impacts model accuracy, interpretation, and predictive power.

However, selecting the most appropriate distribution is not straightforward and depends heavily on the characteristics of the observed data. Therefore, assessing the goodness of fit (GOF) is crucial in determining which distribution best represents the survival times for a given dataset. GOF methods provide a framework to validate the distributional assumptions of AFTMs, ensuring reliable parameter estimates and robust model predictions. Despite their importance, GOF techniques for AFTMs are less developed compared to those for other survival models, leading to potential model misspecifications in practical applications.

In this work we propose the usage of standardized residuals to develop goodness-of-fit tools for AFTMs. Consider the AFTM $Y = \log(T) = \mu + \boldsymbol{\beta}'\boldsymbol{X} + \sigma W$, where T is a possibly right-censored survival time, $\boldsymbol{\beta}$ and $\boldsymbol{X}$ are the parameter and covariate vectors, σ is the scale parameter, and W is the error term distribution, which is determined by the parametric choice for T. For example, if T follows a log-normal distribution, W is the standard normal distribution. The validity of the model-based inference relies on this parametric assumption. To check the parametric assumption based on the sample $(y_i = \log(t_i), \boldsymbol{x}_i)$, $i = 1, \ldots, n$, the following residuals can be used: $r_i = (y_i - (\hat{\mu} + \hat{\boldsymbol{\beta}}'\boldsymbol{x}_i))/\hat{\sigma}$. Residuals r_i are right censored whenever t_i is right-censored. Thus, in order to check the parametric assumption above, goodness of fit methods for right-censored data must be used.

To implement these methods we have used as a benchmark the GofCens R package. Flexible and user-friendly functions have been programmed with the objective of allowing the user to pass an AFTM model with its respective data, and to be returned graphs or pvalues to help the user make a decision for the distribution that best fits the data.

Keywords: Survival analysis, goodness-of-fit methods, accelerated failure time models.

30th Jan 2025
14:45h

Unveiling the Underlying Severity of Multiple Pandemic Indicators

Manuela Alcañiz[1], Marc Estevez[1], Miguel Santolino[1]

[1] RISKcenter, Institut de Recerca en Economia Aplicada (IREA)
Department of Econometrics, Statistics and A.E.
Universitat de Barcelona

Background: Multiple interconnected key metrics are frequently available to track the pandemic progression. One of the difficulties health planners face is determining which provides the best description of the status of the health challenge.
Study design: A longitudinal study.

Methods: The aim of this study is to capture the information provided by multiple pandemic magnitudes in a single metric. Drawing on official Spanish data, we apply techniques of dimension reduction of time series to construct a synthetic pandemic indicator that, based on the multivariate information, captures the evolution of disease severity over time. Three metrics of the evolution of the COVID-19 pandemic are used to construct the composite severity indicator: the daily hospitalizations, ICU admissions and deaths attributable to the coronavirus. The time-varying relationship between the severity indicator and the number of positive cases is investigated.

Results: A single indicator adequately explained the variability of the three time series during the analyzed period (May 2020–March 2022). The severity indicator was stable until mid-March 2021, then fell sharply until October 2021, before stabilizing again. The period of decline coincided with mass vaccination. By age group, the association between underlying severity and positive cases in those aged 80+ was almost 20 times higher than in those aged 20-49.

Conclusions: Our methodology can be applied to other infectious diseases to monitor their severity evolution with a single metric. The synthetic indicator may be useful in assessing the impact of public health interventions on reducing disease severity.

Keywords: Time series; Dimensionality reduction; Dynamic factor models; Pandemics; Vaccination.

Can AI Effectively Interpret Omics Data in Biomedical Research? The Development of GANGO, BIOFUNCTIONAL and RAG

30th Jan 2025
14:45h

Xavi Tarragó[1], Alejandro Rodríguez[1], Antonio Monleón[1,2]

[1]Section of Statistics. Department of Genetics, Microbiology and Statistics. Edifici Aulari (2nd Floor) Faculty of Biology. University of Barcelona Avda Diagonal 645, 08028 Barcelona (Spain);
[2]GRBIO (Research Group in Biostatistics and Bioinformatics)

In biomedical research, interpreting omics data is crucial for understanding the complex molecular mechanisms underlying diseases. To address this challenge, we developed several computational tools. GANGO (Monleon-Getino et al., 2020) retrieves Gene Ontology (GO) and KEGG metabolic pathways from genes and taxa. BIOFUNCTIONAL (Rodriguez and Monleon Getino, 2024) visualizes these ontologies using acyclic graphs for representing multilevel experiments. Our latest tool, RAG (Retrieval-Augmented Generation), leverages AI to facilitate interactive interpretation of GO and KEGG through a chat interface. RAG analyzes input data and generates informative responses, enhancing understanding of biological datasets.

However, a critical question remains: Can AI truly replicate the depth and nuance of human interpretation, particularly in complex biological and pathological contexts? While AI has made significant strides, integrating multiple variables and drawing insightful conclusions remains challenging. The success of AI-powered tools hinges on their ability to accurately and meaningfully interpret data, requiring sophisticated algorithms and a deep understanding of biological processes.

Keywords: Omics data interpretation, Artificial intelligence, Gene Ontology, KEGG pathways, Chat interface bioinformatics tools.

30th Jan 2025
14:45h

Forecasting models for COVID-19: Omicron period

Nere Lerrea[1,2,3], Dae-Jin Lee[3,4,5], Irantzu Barrio[3,4,6], Eduardo Millán[7], José M. Quintana[1,3], Inmaculada Arostegui[3,4,6]

[1]Research Unit Hospital Galdakao-Usansolo; [2]Biosistemak Institute for Health Systems Research; [3]Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS); [4]BCAM, Basque Center for Applied Mathematics; [5]IE University - School of Science and Technology; [6]Department of Mathematics, University of the Basque Country UPV/EHU; [7]Healthcare Services Sub-directorate, Osakidetza-Basque Health Service

Introduction and Objective: One of the objectives during the pandemic was to provide clinicians and healthcare managers with information on predictions about the evolution of the pandemic, for instance, the short-term forecast of the daily number of individuals infected with SARS CoV-2, those admitted to the ward, and ICU admissions. The aim of this study was to evaluate three different forecasting modelling approaches used to provide short-term forecasts and evaluate their accuracy.

Methods: The period of study was between December 15, 2021 and January 25, 2022 (Omicron variant higher peak period) in the Basque Country, Spain. We considered penalized regression splines to evaluate daily counts of SARS-CoV-2 positive cases, hospitalizations and ICU admissions. In order to deal with overdispersion, we used the Negative Binomial distribution for the counts. A generalized additive model (GAM) was considered to account for patterns in the data; three types of penalties have been used for the 1st and 2nd derivatives of this model. The prediction errors were evaluated through the RMSE (root median square error) and the relative error of the 2-day and 5-day forecast.

Results: The RMSE ranged between 3.3-6.6 and 3.6-8.7 for the 2-day and 5-day prediction respectively, with the relative error between 0.15-0.32 and 0.16-0.61 for ICU admissions. For hospital admissions, these values increased to 22.4-25.6 and 23.3-30.2 for the RMSE and 0.24-0.29, 0.23-0.35 for the relative error. Finally, the largest errors were obtained in the prediction of SARS CoV-2 positives, as the errors took values of 2523.5-4363.0 and 3090.7-6787.3 for the RMSE and 0.35-0.55 and 0.45-0.87 for the relative error, respectively. The greatest errors were found when there were sudden trend changes, both for increasing or decreasing trends, and mainly with the model with higher penalty.

Conclusions: Our study on short-term forecasting models during the peak of the Omicron period revealed varying accuracy across different COVID-19 outcomes. While the models performed reasonably well in predicting daily ICU admission counts, their efficacy decreased for hospital admissions and was notably challenged in forecasting positive cases. The highest errors were observed during sudden trend changes, emphasizing the need for more robust models, especially during dynamic shifts in pandemic. These findings underscore the importance of continuous refinement and adaptation of forecasting approaches to enhance their reliability in guiding healthcare responses during pandemics.

Keywords: COVID-19, Pandemic, smoothing methods, short-term forecasting.

Proportionality Index of Parts (PIP) measures the association between taxa in microbiome data

30th Jan 2025
14:45h

Juan Jose Egozcue and Vera Pawlowsky-Glahn

The frequency of taxa in microbiome data sets is compositional. As is well known, correlations (Pearson, Spearman, Kendall) are spurious as they change depending on the normalization adopted. Also, the centred log-ratio (clr) representation produces spurious correlation, thus confronting researchers when analysing these associations between taxa. The Proportionality Index of Parts (PIP) was introduced to measure the degree of association between two taxa (parts in the compositional jargon). Given two taxa, X_1 and X_2, the variance of $\log(X_1/X_2)$ (an element of the variation matrix) does not change with the normalization of the whole composition. The PIP is a normalization of such variance to range $[0, 1]$. Compared with other measures of co-variation, it has the relevant property of being invariant under subcompositions including the involved parts. The PIP was designed to resemble the (bivariate) correlation matrices used to measure the co-variation of real variables. Negative co-variation between taxa is nonsensical. Accordingly, the PIP is a positive index, unlike most correlation indexes. The main characteristics of the PIP are presented and illustrated with a real dataset.

Keywords: -

30th Jan 2025
14:45h

Early-detection of high-risk patient profiles admitted to hospital with respiratory infections using a multistate model

João Carmezim[1], Cristian Tebé[1], Natàlia Pallarès[1], Roger Paredes[1], Cavan Reilly[2]

[1]Germans Trias i Pujol Research Institute and Hospital; [2]University of Minnesota

Background: This study aims to identify clinically relevant prognostic factors associated with oxygen support, death or hospital discharge in a global cohort of adult patients with Influenza or COVID-19 using a multistate model.

Methods: Data was drawn from a cohort of adult patients diagnosed with respiratory infections admitted to a hospital of the Strategies and Treatments for Respiratory Infections and Viral Emergencies (STRIVE) research group. The study evaluates socio-demographic factors, medical history, comorbidities, vaccination status, virus type and clinical symptoms as prognostic factors. The multistate model was defined with the following states: hospital admission, noninvasive ventilation, invasive ventilation, oxygen support discharge, hospital discharge and death. The model estimates cause-specific hazard ratios, cumulative hazards and transition probabilities.

Results: A total of 4968 patients were included where the median age was 62.1 and the percentage of females was 47.9%. The number of patients that needed noninvasive ventilation was 1906 (38.4%), 277 (5.6%) required invasive ventilation, and 275 (5.5%) died. Demographic and clinical risk profiles revealed distinct progression pathways, and visualization using trajectory plots highlighted how risk factors influenced movement through disease states.

Discussion: This study highlights the utility of a multistate model in mapping the progression of respiratory infections, offering critical insights into high-risk patient profiles. Transition probability trajectories provide actionable data for clinicians to predict outcomes and optimize resource allocation for patients with Influenza or COVID-19.

Keywords: Multistate model, Respiratory infections, Risk profiles.

Author Index