



St ANDREWS
BIOINFERENCE
June 10-12, 2026

Book of Abstracts

Posters (A-Z)

Begoña Bolos (University of Edinburgh)

The C-index Multiverse

Clinical prediction models identify individuals at risk of adverse health events (e.g cancer relapse). Time-to-event models are well suited for this task, estimating whether and when an event may occur. Model evaluation and selection often rely on discrimination metrics, which assess whether those who experience events earlier receive higher predicted risk.

The concordance index (C-index) is widely used for this purpose, especially with the rise of deep learning in survival analysis. Particularly, the presence of censoring and ties in survival data make the C-index computation non-trivial. As a result, several C-index implementations exist in R and Python, resulting in a multiverse of C-index estimators which, while conceptually similar, yield different results practically. Additionally, the absence of a standardised approach to summarise risk from survival distributions, result in another source of variation dependent on input types.

This work reviews common C-index estimators and maps them to existing software, highlighting departures from original definitions. Using public breast cancer data, we trained models ranging from Cox proportional hazards to deep model approaches and demonstrated that model selection is influenced by the chosen implementations and input.

We conclude that the C-index multiverse undermines reproducibility and complicates comparisons across models and studies, motivating clearer reporting and providing a practical guideline for analysts navigating existing software.

Sian Dennett (University of Oxford)

Understanding Spatiotemporal Variability in Dengue Transmission Dynamics in Brazil

Recent seasons have seen record levels of dengue case incidence in Brazil, including expansion of the arbovirus to southern regions that have previously not experienced endemicity. Such a severe shift in transmission dynamics was not anticipated so soon and is often attributed to a complex combination of human, environmental and biological factors driving changes in, for example, human movement, climate patterns, and mosquito abundance. The relative influence of each and the overall time- and space-varying mechanisms behind these record seasons remains poorly understood and is difficult to characterise due to the size and

heterogeneity of Brazil and its population. Therefore, in this work, we use wavelet analysis, clustering methods, Granger causality tests and mobility models, to analyse epidemiological, climate, human mobility, and demographic time-series data. Our exploration of the causal relationships driving the magnitude and timing of recent dengue epidemics provides the foundations on which more realistic, descriptive and useful mechanistic models of current dynamics and predictive models of future dynamics are being developed. Crucially, the increased causal understanding developed in this work will supplement forecasts and mechanistic models to generate clearer, evidence-based guidance on when, where, and why interventions should be deployed to reduce the substantial public health burden currently facing Brazil.

Marina Evangelou (Imperial College London)

Adaptive gene and pathway selection through sparse-group SLOPE

Genetics and more broadly OMICS datasets are usually high-dimensional datasets that through their analysis we aim to identify the genes (the variables) that are associated with the response of interest, e.g. with a disease. Popular models for addressing such a question are penalised regression models, like least absolute shrinkage and selection operator (LASSO), that shrink the coefficients of most of the variables equal to zero. A number of extensions of LASSO have been proposed in the literature over the years, where one of them is Sorted L-one Penalised Estimation (SLOPE) that simultaneously to variable selection offers false discovery rate (FDR) control. This is a key property when analysing genetics datasets as a large number of false discoveries can have a tremendous negative impact to subsequent functional experiments.

Genes naturally form into groups where members of a group work together for achieving different biological tasks. Such groups are called biological pathways where each one of the pathways has a specific role and product in the human body. Pathway analysis aims to identify the pathways (groups of genes) that are associated with the response of interest. Over the years, the identification of pathways and their corresponding genes has led to important biological discoveries in the literature.

In this work we present a sparse-group model based on SLOPE that offers both variable (gene) and group (pathway) selection. The proposed sparse-group SLOPE (SGS) model is able to control both the false discovery rate at the gene and pathway levels. Through the analysis of both synthetic and real genetics datasets, the proposed SGS approach is

found to outperform other existing lasso- and SLOPE-based models for both variable and group selection and prediction accuracy.

Chao Gao (University of St Andrews)

Decoding Hidden Signals: A Supervised Learning Framework for Information Estimation

Mutual Information (MI) is a fundamental concept in information theory that measures the reduction in uncertainty of one random variable given knowledge of another. Unlike Pearson correlation, which captures only linear relationships, MI can detect both linear and non-linear dependencies. This makes it particularly valuable in fields such as genomics, where complex relationships exist between high-dimensional gene expression data and clinical outcomes. By quantifying the shared information between genes and disease states, MI enables the identification of informative biomarkers and improves the interpretability of predictive models.

Despite its advantages, estimating MI between a discrete variable and a high-dimensional continuous variable remains challenging. Classical estimators, including K-nearest neighbors (KNN), kernel density estimation (KDE), SLEMI, and neural estimators such as MINE, are often hindered by the curse of dimensionality or require prohibitively large sample sizes to achieve reliable performance. This limits their applicability in many real-world scenarios where data are high-dimensional but sample sizes are moderate.

To overcome these limitations, we propose a novel, supervised learning-based approach. Rather than estimating MI directly in the original high-dimensional space, we compute MI between the discrete variable and the class probability outputs of a trained supervised model. These probabilities provide a low-dimensional, informative representation of the continuous variable, yielding a computationally efficient lower bound on the true MI. Empirical comparisons demonstrate that our non-parametric, data-driven method outperforms traditional estimators and remains robust in high-dimensional settings.

Alicia Gill (University of Oxford)

Estimating Lassa Fever Vaccine Efficacy Under Diagnostic Uncertainty

Evaluation of Lassa fever vaccine efficacy in endemic settings is complicated by imperfect diagnostic tests, frequent malaria coinfection,

and heterogeneity in clinical severity. We present the design and planned analysis of a phase 3 randomised trial of a Lassa fever vaccine candidate that explicitly addresses these challenges. Rather than relying on a binary case definition, we propose an ordinal outcome capturing increasing levels of disease severity. Diagnostic uncertainty is incorporated by modeling test sensitivity and specificity directly, while malaria coinfection and other baseline covariates are included as adjustment variables. We develop a Bayesian model using data augmentation to represent latent true infection and severity states, allowing propagation of uncertainty through all stages of inference. The proposed approach improves efficiency, reduces bias due to misclassification, and yields interpretable posterior estimates of vaccine efficacy across severity levels. Simulation results demonstrate advantages over conventional analyses that ignore diagnostic error or collapse outcomes to binary endpoints. This framework offers a flexible strategy for late-phase vaccine trials in resource-limited, high-burden settings.

Charis Hanna (University of St Andrews)

IoMB: Evaluating Object Detectors on Occluded and Imbalanced Seabird populations

Computer vision techniques are seeing increasing use in wildlife monitoring surveys but challenges remain when applying standard techniques to complex, real-world ecological data. To evaluate the state of the art in this domain, we introduce IoMB (Isle of May Birds), a dataset specifically designed to reflect the challenges of monitoring cliff-nesting seabird species. Unlike widely used benchmarks such as CUB (Caltech-UCSD Birds-200-2011) and NABirds, IoMB captures a wide range of relative bird sizes, high class imbalance, and frequent occlusion. We benchmark several of state-of-the-art object detectors YOLO, RetinaNet, Faster R-CNN and DINO on this dataset and show that these standard detectors struggle with occlusion, limited species representation, and large size variation. Although pre-training on larger datasets provides some improvement, it cannot fully address these challenges. The abundance of unannotated data (12,000 images) relative to our 200 annotated samples motivates exploration of self- and semi-supervised approaches, such as pseudo-labelling, to leverage unlabelled data. Such methods could help models learn robust features from diverse visual conditions without requiring exhaustive manual annotation. Overall, our study highlights the need for techniques tailored to wildlife monitoring, capable of combining limited labelled data with abundant unlabelled imagery to improve detection and classification in challenging ecological settings.

Grant Henderson (Biomathematics and Statistics Scotland)

Quantifying Spread and Control of UK HPAI Outbreaks

Highly Pathogenic Avian Influenza (HPAI) outbreaks represent a significant threat to UK poultry and kept-bird sectors and raise concerns of spillover and zoonotic risk. It is known that premises can be infected by wild birds, but questions remain about the risk of subsequent local disease incursion around already infected premises, and the relative importance of these medium- and long-range infection routes.

We use advanced Markov chain Monte Carlo approaches embedded in BICI, our open-source Bayesian individual-based compartmental inference software tool [1] to build, parameterise and compare spatio-temporal models describing UK poultry premises infections, their subsequent detection, and control.

We employ data on farm breakdowns between 2021 and 2023 and show a mixture of significant local transmission and more limited long-range seeding is consistent with the observed outbreak. Furthermore, these tools enable quantification of the efficacy of the local and national restriction zones used as control mechanisms during the outbreak.

This analysis has significant capacity to quantitatively inform response to ongoing outbreaks, and we are pursuing engagement with policy colleagues through our participation in EPIC, the Scottish Government's Centre of Expertise on Animal Disease Outbreaks.

References:

[1] C. M. Pooley et al., <https://github.com/theITEAM/BICI> (2026).

Ishtiaq Hussain (University of Strathclyde)

Mathematical Modelling for Using Insect-Specific Viruses to Curb Dengue Epidemics: A Case Study in Pakistan

Mosquito-borne viruses such as Dengue, Zika, and Chikungunya viruses remain a major global health problem. Dengue alone caused more than 14.4 million reported cases and over 11,000 deaths worldwide in 2024, according to the World Health Organization (WHO). Recent work has suggested that insect-specific viruses (ISVs), which infect mosquitoes but not humans, may reduce the ability of mosquitoes to transmit pathogenic viruses. This mechanism is superinfection exclusion (SIE), where infection with one virus prevents or limits infection by another.

In this work, we explore whether ISVs could potentially be used to help reduce dengue outbreaks. We develop a weather-driven SEIR–SEI compartmental model that describes dengue transmission between humans and mosquitoes, while also including ISV infection in the mosquito population. The model is parameterised using dengue case data and meteorological data from 11 cities in Punjab, Pakistan, covering the period 2013–2025.

We then perform sensitivity analyses to understand which parameters are most important for ISV establishment and for reducing dengue transmission. In particular, we explore the conditions under which ISV-mediated SIE could meaningfully suppress dengue epidemics under different climatic scenarios.

Shilji Nihar Kadavath madathin padayantavida (University of Lancashire)

Country level variations in time-to-recurrence after curative resection of colorectal liver metastasis (CRLM) patients: A multinational retrospective cohort analysis

This study aimed to examine country-level heterogeneity in time to recurrence after curative resection of colorectal liver metastases (CRLM) and to identify predictive factors of recurrence.

We analysed 28,365 patients with CRLM from 55 countries using a Cox proportional hazard model with a shared frailty term at the country level.

During a median follow-up of 22 months, recurrence occurred in 23% of patients, with most events arising within the first three postoperative years. The shared frailty model demonstrated substantial variations at the country level. In the multivariable frailty-adjusted analysis, a higher tumour burden score, advanced metastatic stage, bilobar liver involvement, and a greater number of metastases were associated with a shorter time to recurrence, whereas wider resection margins were protective.

The results showed that recurrence after CRLM resection is driven by both substantial country-level heterogeneity and consistent clinical predictors. This reiterates the need for global standardisation of treatment strategies to achieve optimal patient outcomes.

Cameron Kerr (University of Edinburgh)

Decoding Clonal Dynamics in Blood Cancer Through An Adapted CHIP Model

An estimated 330 billion blood cells are produced daily by approximately 100,000 haematopoietic stem and progenitor cells through haematopoiesis. During this process, somatic mutations accumulate with age, occasionally conferring fitness advantages that enable clonal expansion. This can lead to pre-malignant clonal haematopoiesis of indeterminate potential (CHIP) or malignancies like myelodysplastic syndromes (MDS) and leukemia.

Mathematical models successfully infer mutant fitness in CHIP, but their applicability to disease progression remains unclear. We focus on low-risk MDS as a realistic progression step from CHIP. A critical challenge is loss of heterozygosity (LOH), where cells lose one chromosome copy, doubling the variant allele frequency (VAF) without changing actual clone size—a phenomenon current models cannot accommodate.

Since longitudinal patient data lack zygosity information, our model must infer this parameter alongside clonal fitness. Using synthetic data, we assessed whether clonal fitness and zygosity can be jointly inferred for single and multiple competing clones under various realistic sampling conditions.

Our results demonstrate the importance of incorporating additional biological complexity into clonal dynamics models and provide a framework for improved mutation risk stratification in MDS. This approach supports more robust and clinically meaningful inference from longitudinal sequencing data, potentially enabling better patient outcome predictions.

Richard Kettle (University of Edinburgh)

Statror Reveals Context Dependent Cytokine Responses in CD8+ T Cells

The immune system is responsible for protecting the body against infection and malignant tissue through the coordination of gene programs acting across diverse cell types, subtypes and transient cellular states. Standard approaches at characterizing cell subpopulations with single-cell RNA sequencing (scRNA-seq) data rely on clustering in aggregate RNA expression space, potentially overlooking non-local structure.

Here, we apply Stator, a framework that identifies meaningful cell subpopulations – termed ‘States’ – by calculating high-order gene interactions, and labels cells based on the expression of combinatorial RNA markers. We analyse a cytokine-perturbation scRNA-seq immune cell dataset of a single cell type – specifically CD8+ T Cells – to investigate how cytokines regulate downstream transcriptional programs and cell state occupancy. We demonstrate how Stator uncovers distinct gene programs and states that exhibit context-dependent responses to cytokine perturbations. Our results map relationships between cytokine signals, transcriptional programs and cellular states, providing a higher fidelity perspective of immune cell population biology than is possible through conventional clustering approaches.

Janique Krasnowska (University of Warwick)

Coalescence in Multi-Type Branching Processes with Applications to Dormancy

Branching processes constitute a broad class of stochastic models with applications across many problems in biology. Understanding the ancestry of a sample of individuals is a fundamental question in evolutionary biology and helps describe phenomena such as genetic drift or speciation. The genealogy of branching processes has been studied extensively in the setting where all individuals in a population share a common offspring distribution. A natural generalisation is to consider populations consisting of multiple types of individuals with distinct offspring distributions, representing features such as genetic types, mutations, or dormant and active states.

In this work, we study a supercritical multi-type Galton–Watson process with a possibly countably infinite number of types. We derive an expression for the distribution function of the generation of the most recent common ancestor for a sample of k individuals drawn from a large generation. Under moment conditions on the offspring distributions, we also establish upper and lower bounds, with explicit constants, for this distribution function in terms of moments of the inverse population size. Finally, through simulation studies, we demonstrate that our approach is computationally more efficient than direct simulation of the genealogy of a sample.

This is joint work with P. Jenkins and A. Johansen.

Mohammed Moinuddin (University of Lancashire) and Halima Kabir Tonni (University of Lancashire)

Identifying the factors that moderate the association between resection margin and survival outcomes for colorectal liver metastasis patients in Europe

This study investigated the clinicopathological factors that moderate the association between surgical resection margins (R0 or R1) and survival outcomes in European patients with colorectal liver metastases (CRLM).

Thirty thousand patients from an international CRLM database in selected European countries were included in the analysis. Survival analysis techniques were applied to investigate the data.

R1 resection margins and older age were consistently associated with poorer overall survival (OS) and disease-free survival (DFS). Advanced TNM stage, right-sided and rectal tumours, elevated CEA, multiple and unresectable tumours, extra-hepatic metastases, and local ablative therapies predicted worse outcomes, whereas left-sided tumours were protective for both outcomes (OS and DFS). For overall survival, the key moderators were right-sided tumours, rectal tumours, higher tumour count, and initial resectability, whereas right-sided and initial resectability moderated the association between disease-free survival and resection margin.

Moderators should be considered when developing new interventions and multimodal treatment approaches for CRLM to improve survival outcomes.

Francesco Morandin (University of Pisa)

From Genes Correlation to Differential Expression: COTAN scRNA-seq Analysis

Single-cell RNA-sequencing (scRNA-seq) now routinely profiles thousands of individual transcriptomes, yet no consensus pipeline exists for either laboratory protocol or downstream analysis. Droplet-based datasets are especially challenging: their count matrices are extremely sparse and many biologically important genes are expressed at very low levels. Standard workflows, borrowed from bulk RNA-seq, apply normalization, log-transformation, and feature filtering, steps that can distort true zero counts and systematically discard low-expression regulators. A method that natively handles sparsity, avoids imputation, and retains informative low-abundance genes is therefore desirable. We

present COTAN, a framework based co-expression of genes, estimated via a generalized Phi correlation coefficient. COTAN models zero counts directly, requires no log-normalization or scaling, and is tailored to UMI droplet data.

Akihiro Mori (University of Aberdeen)

Hybrid stochastic modelling of protocell growth and homochirality emergence

Compartmentalized vesicles are widely considered a key step in the origin of life, providing confined environments where early biochemical reactions could occur. However, the coupling between stochastic biochemical activity within protocells and the resulting morphological dynamics of the compartment remains poorly understood.

Here we introduce a hybrid stochastic modelling framework to study vesicle growth and internal polymerization dynamics. As a minimal system, the model describes the diffusion of two enantiomeric monomers (L and D) across the vesicle membrane through volume-dependent passive transport in a two-dimensional vesicle model. Once inside the compartment, these monomers initiate and elongate homopolymer chains via cooperative Hills functions. To capture multiple stochastic timescales, the system combines an exact Gillespie stochastic simulation algorithm for low-frequency reaction events with tau-leaping approximations for frequent diffusion processes.

Vesicle volume is dynamically regulated by the length of the longest active homopolymer, generating nonlinear feedback between internal biochemical activity and vesicle morphology. Simulations show that this feedback, together with intrinsic molecular noise, can drive symmetry breaking between L and D polymers, suggesting a stochastic mechanism for the emergence of homochirality in protocellular systems.

We further extend the framework to spatially resolved 3D protocell simulations incorporating spatial and temporal gradients in the concentrations of nucleotides and amino acids. This modelling approach provides a quantitative platform for exploring how stochastic reaction dynamics and compartment morphology jointly shape protocellular evolution and offers a foundation for integrating theoretical predictions with experimental studies of vesicle-based biochemical systems.

Scott Pirrie (Optima Partners Ltd)

Integrating genetics and multi omics to distinguish risk and progression targets for drug repurposing in Parkinson's disease

Aims

While GWAS studies have identified numerous risk loci common, complex diseases, causal genes and pathways remain difficult to pinpoint. Using Parkinson's disease (PD) as an example, we aimed to build an integrative multi omic aetiology map linking genetic risk to molecular mechanisms, cell types, pathways and clinical phenotypes, to help prioritise therapeutic targets and drug repurposing candidates.

Methods

Using summary statistics from a large European PD GWAS meta analysis (Nalls et al., 2019), and over 5,000 molecular QTL datasets across blood and brain, we applied mathematically and computationally optimised genetic fine mapping and multi trait colocalisation algorithms to identify molecular traits with likely causal roles in disease.

Results

Our layered approach identified clusters of molecular signals with strong causal evidence. Embedding these into a PD-focused knowledge graph enabled construction of a phenotype network highlighting key pathways and distinguishing risk-associated genes from those linked to disease onset. Incorporating clinical traits and drug information further allowed prioritisation of mechanistically informed targets (e.g. RAB29, EFNA1) and repurposing candidates (e.g. CD38).

Conclusions

Our framework moves beyond association signals to nominate biologically grounded therapeutic targets, offering a scalable strategy to accelerate PD drug discovery and improve clinical trial design.

Andrey Shternshis (Uppsala University)

Learning dynamics of microbiome abundance by logit-normal processes

Longitudinal studies involving human data often contain missing values, typically arising from missed appointments or insufficient read depth. Such incomplete data reduce the effective sample size, which is already

limited in longitudinal settings. I aim is to impute missing microbiome samples with a primary focus on imputation accuracy, defined as the distance between predicted and true samples during cross-validation. The data consist of relative species abundances, with each sample represented as a compositional vector. I implemented and trained a Gaussian random field on a two-dimensional representation of the data, where samples are positioned at the intersections of time points and study participants. The Gaussian field consists of a mean function dependent on participant-level covariates and time, combined with a Gaussian process whose covariance structure depends on time. The approach involves a moderate number of parameters and is therefore suitable for relatively small datasets. To generate compositional vectors, I use either a Dirichlet distribution or a softmax (ramp) transformation. The latter corresponds to a modified logit-normal distribution and outperforms both the Dirichlet-based model with a latent Gaussian field and a baseline method that copies the closest available sample. However, more comparison analysis with other compositional distributions can be done to improve down-streaming prediction accuracy.

Alexander Zarebski (University of Cambridge)

GENIE: A Spatio-Temporal Neural Network for Probabilistic Disease Forecasting using Agent-Based Model simulations

The SARS-CoV-2 pandemic highlighted the burden that pathogens can place on society. The compartmental models, used to track the pandemic, struggled to capture the complex transmission dynamics, leading to biased and overconfident forecasts. They also lacked the necessary geographical resolution for actionable decision-making.

To address these limitations, we introduce GENIE, a machine learning model for spatio-temporal disease forecasting. GENIE predicts hospitalisations, deaths, infections, and reproduction numbers at a fine geographic scale. By training on data generated by an agent-based model, GENIE learns from realistic simulated epidemic trajectories, capturing patterns that are oversimplified by compartmental models. GENIE's architecture uses Graph Neural Networks to learn the spread of infection across space while drawing upon location-specific socio-demographic information. Autoregressive prediction is used to generate probabilistic forecasts.

GENIE can rapidly generate approximate samples from the posterior predictive distribution on consumer hardware. Through simulation studies, we demonstrate GENIE's ability to generalise beyond its training data and the importance of accounting for spatial structure. We also compare its performance to existing models, including Mantis, a foundation model for

mechanistic disease forecasting, and hhh4, an endemic-epidemic multivariate time-series model.