Discussion: Hotly Debated Topics in Biostatistics

STRATOS Session

Malka Gorfine
Tel Aviv University, Israel

IBC Seoul, July 2026

My main take-home message

Across the four talks, I see one common theme:

Modern biostatistics should not be judged only by methodological sophistication, but by whether it supports reliable scientific conclusions.

This requires:

- clear estimands and scientific targets;
- transparent analysis protocols;
- careful model building and tuning;
- honest uncertainty quantification;
- reporting that supports reproducibility.

Federico Ambrogi: ML/AI and traditional modelling

Main message.

- ML/AI is promising, especially for high-dimensional, multimodal and unstructured data.
- But ML is not automatically superior to classical statistical models.
- Model performance depends on data structure, sample size, validation, calibration and tuning.
- Poor tuning and poor validation can create artificial ML superiority.

My comment.

- Prediction accuracy is not enough in biomedical research.
- AI/ML methods should also provide valid uncertainty quantification.
- Data-driven tuning should be part of the uncertainty assessment.

Federico: suggested discussion question

Point 1

Uncertainty estimation should become a standard requirement for clinical ML/AI models, in addition to discrimination, calibration and external validation.

Point 2

Tuning is not a technical detail. Architecture choice, penalty parameters, early stopping, variable screening and model selection are all part of the statistical procedure.

Inference should acknowledge the full data-driven pipeline.

Carsten Schmidt: SAPI protocol

Main message.

- SAPI combines a statistical analysis plan with initial data analysis.
- It aims to document data quality checks, preprocessing, missingness, deviations from the original plan and analysis decisions.
- The goal is to reduce hidden flexibility and improve transparency.

How can this improve research?

- It forces researchers to define the target population, index date, variables, estimands and main analysis before final results are known.
- It makes unavoidable data-driven decisions explicit and documented.
- It can reduce selective reporting and irreproducibility.

Carsten: suggested discussion questions

Question 1

Do you believe journals will eventually require SAPI for observational studies, at least for major applied papers?

Question 2

How should we balance pre-specification with flexibility? Observational data often contain surprises. When is an amendment legitimate, and when does it become post-hoc analysis?

My concern: without incentives from journals, reviewers and funders, protocols may remain useful but optional documents.

Michal Abrahamowicz: hazards of hazard ratios

Main message.

- The talk re-assesses the criticism that hazard ratios suffer from built-in selection bias.
- Unmeasured frailty can induce bias and artificial time-varying HRs.
- But simulations suggest this is serious mainly under strong frailty and/or high event incidence.
- In the hormone therapy example, non-adherence or biological time-varying effects may be more plausible explanations.

My comment.

We should avoid two extremes:

- HRs are always safe;
- HRs are always misleading.

Michal: from “hazards of HRs” to useful hazard thinking

What I take from Michal’s talk

Time-varying HRs should not be dismissed automatically as artifacts. They may reflect real biological or treatment-process dynamics.

Hazards are identifiable because they condition on being at risk. This is exactly why they are useful for censored event-time data, but also why their causal interpretation requires care.

My addition: Censoring, competing events and multistate structure may affect what a hazard-based summary is actually summarizing.

So my message is: keep hazards, model them flexibly, but translate the conclusion into the scientific estimand.

James Carpenter: p-values beyond polemics

Main message.

- The problem is not the p-value itself, but careless use and reporting.
- P-values, confidence intervals and point estimates are all useful when interpreted correctly.
- P-values are not reproducible statistics; design, power and error control are the reproducible components.
- P-values can also be used as model-complexity tuning parameters.

My comment.

- I like the distinction between p-values as inferential summaries and p-values as tuning parameters.
- The latter connects naturally to ML tuning.
- In both cases, we need transparent reporting and sensitivity analysis.

James: suggested discussion question

Question

When p-values are used as tuning parameters in model building, how should we report uncertainty after model selection?

Related point:

- Whether the tuning parameter is a p-value threshold, a penalty parameter, a number of trees, or a neural-network architecture, it controls model complexity.
- Therefore, tuning choices should be reported and examined through stability or sensitivity analyses.

Closing

The future is not simply more complex models, nor simply stricter rules.

It is better analytical thinking.

- ML/AI needs uncertainty, not only prediction.
- Protocols need incentives and adoption.
- HRs need careful interpretation and complementary estimands.
- P-values need principled use, not polemics.