

ERCIM



NEWS



Special theme:

E-VALUES

Statistical Testing for the 21st Century

Editorial Information

ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 2,000 printed copies and is also available online at <https://ercim-news.ercim.eu>.

ERCIM News is published by ERCIM EEIG
BP 93, F-06902 Sophia Antipolis Cedex, France
+33 4 9238 5010, contact@ercim.eu
Director: Dominique Hazaël-Massieux,
ISSN 0926-4981

Contributions

Contributions should be submitted to the local editor of your country

Copyright notice

All authors, as identified in each article, retain copyright of their work. ERCIM News is licensed under a Creative Commons Attribution 4.0 International License (CC-BY).

Advertising

For current advertising rates and conditions, see <https://ercim-news.ercim.eu/> or contact peter.kunz@ercim.eu

ERCIM News online edition

<https://ercim-news.ercim.eu>

Subscription

Subscribe to ERCIM News by sending an email to en-subscriptions@ercim.eu

Editorial Board:

Central editor: Peter Kunz, ERCIM office (peter.kunz@ercim.eu)

Special Theme guest editors:

- Peter Grünwald (CWI and Leiden University)
- Wouter Koolen (CWI and University of Twente)
- Johanna Ziegel (ETH Zurich)

Local Editors:

- Ferran Argelaguet, Inria, France (ferran.argelaguet@inria.fr)
- Andras Benczur, SZTAKI, Hungary (benczur@info.ilab.sztaki.hu)
- José Borbinha, Univ. of Technology Lisboa, Portugal (jlb@ist.utl.pt)
- Monica Divitini, NTNU, Norway (divitini@ntnu.no)
- Marie-Claire Forgue, ERCIM/W3C (mcf@w3.org)
- Lida Harami, ICS-FORTH, Greece (lida@ics.forth.gr)
- Athanasios Kalogeras, ISI, Greece (kalogeras@isi.gr)
- Georgia Kapitsaki, Univ. of Cyprus, Cyprus (gkapi@cs.ucy.ac.cy)
- Annette Kik, CWI, The Netherlands (Annette.Kik@cwi.nl)
- Alexander Nouak, Fraunhofer-Gesellschaft, Germany (alexander.nouak@iuk.fraunhofer.de)
- Laura Panizo, University of Malaga (laurapanizo@uma.es)
- Erwin Schoitsch, AIT, Austria (erwin.schoitsch@ait.ac.at)
- Thomas Tamisier, LIST, Luxembourg (thomas.tamisier@list.lu)
- Maurice ter Beek, CNR-ISTI, Italy (maurice.terbeek@isti.cnr.it)

Cover image generated by AI.

Contents

KEYNOTE

- 4 Foreword: From P-values to E-values**
by Michael I. Jordan

JOINT ERCIM ACTIONS

- 5 14th International Workshop on Computational Intelligence for Multimedia Understanding**
by Behçet Uğur Töreyn (İTÜ), Maria Trocan (ISEP) and Davide Moroni (CNR-ISTI)
- 5 Spring ERCIM Meetings 2026 Held in Vienna**

SPECIAL THEME

Introduction to the Special Theme:

- 6 E-values – Statistical Testing for the 21st Century**
by the guest editors Peter Grünwald (CWI and Leiden University), Wouter Koolen (CWI and University of Twente) and Johanna Ziegel (ETH Zurich)
- 8 On Testing by Imaginary Betting**
by Glenn Shafer
- 9 Optimal Gambling Strategies when Testing by Betting**
by Martin Larsson (Carnegie Mellon University), Aaditya Ramdas (Carnegie Mellon University), and Johannes Ruf (London School of Economics)
- 10 E-values Are *Just* (Some) P-values**
by Eugenio Clerico (University of Oxford)
- 11 E-values, P-values and Counterfactuals**
by Peter Grünwald (CWI and Leiden University)
- 13 E-processes, Predictive Recursion, and Objective Empirical Probability**
by Vaidehi Dixit (University of Nottingham) and Ryan Martin (North Carolina State University)
- 14 Rethinking Conformal Prediction for Constrained Environments**
by Etienne Gauthier (Inria, Ecole Normale Supérieure, Université PSL)
- 16 More Discoveries and Flexibility in Multiple Testing**
by Rianne de Heide (University of Twente and CWI); based on joint work with Jelle Goeman (Leiden University Medical Center), and Aldo Solari (Ca' Foscari University of Venice), and others
- 17 Combining Rejection Sets with False Discovery Control**
by Zhimei Ren (University of Pennsylvania)

RESEARCH AND INNOVATION

18 E-values for Stable and Robust Variable Selection in Microbiome Studies

by Gianna Serafina Monti (University of Milano-Bicocca), and Peter Filzmoser (TU Wien)

20 Betting on Bets: Comparing the Upsides of Uncertain Prospects over Time with E-values

by Yo Joong Choe (INSEAD) and Sebastian Arnold (CWI)

21 A Map of Sub-Gaussian E-variables

by Wouter M. Koolen (CWI & University of Twente), Shubhada Agrawal (IISc Bangalore) and Martin Larsson (Carnegie Mellon University)

23 E-values for Contingency Tables

by Thorsten Dickhaus (University of Bremen), Francesca Giuffrida (Leiden University and IMT School for Advanced Studies Lucca) and Yongqi Wang (CWI)

25 E-values Instead of P-values in Clinical Trials: What Happens?

by Yongxi Long and Erik van Zwet (Leiden University Medical Center)

26 Racing to the Truth: How E-values Can Speed Up Science

by Sebastian Arias, Alexander Ly, Michele Meziu (CWI) and Angel Reyero Lobo (CWI and Inria)

28 How Small is Negligible? Rethinking Equivalence Testing with E-values

by Stan Koobs and Nick W. Koning (Erasmus University Rotterdam)

29 Safe Preference-Based Political Polls

by Adrienne Tuynman and Timothée Mathieu (Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 – CRISTAL)

31 Assessing Generative AI Systems Using E-scores

by Guneet Singh Dhillon (University of Oxford), Teodora Pandeva (Microsoft Research), and Alicia Curth (Microsoft Research)

32 Winning by Watching: Evaluating Policies Safely for Reinforcement Learning

by Stephan Bongers (CWI)

34 Scrutinizing Banks' Risk Forecasts with E-backtesting

by Ruodu Wang (University of Waterloo)

35 Anytime-Valid Testing in the Age of AI-Assisted Software Development

by Michael Scott Lindon (Netflix)

37 Large Language Models as Design Partners: Automating Graphical Mockups to Refine Requirements

by Giovanna Broccia, Maurice H. ter Beek (CNR-ISTI), and Alessio Ferrari (University College Dublin and CNR-ISTI)

39 Challenges in Small-Scale Medical Data Exchange Platform Development

by Hubert Schölnast, Peter Kieseberg, Patrick Kochberger and Henri Ruotsalainen (University of Applied Sciences St. Pölten)

40 Building Sovereign AI for Europe's Energy Networks

by Antonello Monti (Fraunhofer Institute for Applied Information Technology FIT, Germany)

42 SCIANCE: AI for Scientific Discovery in Europe

by András Benczúr, Edina Nemeth (SZTAKI), Jonas L'Haridon (European Science Foundation) and Magdalena Brus (EGI Foundation)

ANNOUNCEMENTS

43 Horizon Europe: ERCIM Project Support Services

44 Dagstuhl Seminars and Perspectives Workshops

44 ERCIM Forum Beyond Compliance on Digital Ethics in Research

44 Training in Digital Ethics

45 ERCIM Working Group "Inclusive Digital Futures" Held Kick-off Meeting

45 FMICS 2026: 31st International Conference on Formal Methods for Industrial Critical Systems

45 MODELS 2026 in Málaga: Advancing Model-Driven Engineering in the Age of AI

46 ERCIM "Alain Bensoussan" Fellowship Programme

47 Showcase your EU Project in ERCIM News

NEXT ISSUE

ERCIM News 146, October 2026

Special theme:

Quantum Technology

Foreword: From P-values to E-values

by Michael I. Jordan

What is an “e-value” and why has it become an object of intense study in statistics and in the allied fields of machine learning, signal processing, and econometrics? To briefly introduce the basic idea, let us consider one of the core problems in statistics – the “hypothesis testing problem” of deciding whether observed data is consistent with some particular data-generating mechanism (often referred to as a “null hypothesis”) or is better explained by another mechanism (referred to as an “alternative hypothesis”). This problem is addressed by defining some function of the data (a “statistic”) whose distribution is as different as possible under the null and the alternative. Given an observed value of such a statistic, one then makes a choice between the two distributions, doing so in a way that minimizes the probability of errors. Classical statistical theory provides a unifying framework – the “p-value” – by which the choice between the null and alternative hypotheses reduces to a thresholding procedure.

E-values provide an alternative to the classical p-value paradigm. Rather than being tail probabilities under the null hypothesis, e-values are nonnegative random variables whose expectation is less than or equal to one under the null hypothesis. Given that tail probabilities and expectations are related by Markov’s inequality, it may seem that p-values and e-values are not so very different. But the key point is that they arise from different perspectives and in different problem settings, and they accordingly have different strengths and weaknesses.

To elaborate and to set the stage for the current collection, note that the focus of p-values is the analysis of an experiment in which a sample size is chosen, a batch of data having that sample size is collected, a p-value is computed, and a decision is made – with no possibility of revisiting the decision. E-values, on the other hand, focus on an online framework in which a stream of data is observed, with no a priori choice of the number of data points. The statistician is viewed as accruing evidence over time with the possibility of making a decision at any point in time. The decisions are thus tentative and revokable.

This online perspective is turned into mathematics by making use of the machinery of martingale theory. One views the statistician as placing bets over time that aim to reveal whether the null hypothesis or the alternative hypothesis is the source of the data stream. The null hypothesis can be interpreted in terms of a casino in which the odds are not in the statistician’s favor, such that her wealth dwindles over time, no matter how the bets are placed. This dwindling behavior can be modeled mathematically with martingale theory – in particular as a nonnegative supermartingale. On the other hand, the alternative hypothesis corresponds to a casino in which the odds are in the statistician’s favor, so that if she places her bets wisely her wealth will increase over time.



Michael I. Jordan

Directeur de Recherche, Inria and École normale supérieure (ENS), Paris; Distinguished Professor Emeritus, University of California, Berkeley.

Pioneer of modern machine learning, statistics and artificial intelligence.

Thus, under the alternative the wealth is not a martingale; indeed, it can be made to increase exponentially. Statistical hypothesis testing is thereby viewed as the discrimination between a losing betting process that dwindles to zero and a winning process that grows exponentially.

E-values are then defined as the value of the wealth process at a stopping time of the process. It is a consequence of the optional stopping theorem that e-values are bounded in expectation by one under the null hypothesis. Decisions can be made by thresholding e-values.

This online perspective on hypothesis testing arose slowly over many decades in statistics, in parallel with the rise of p-value-based statistics. Its early development came stochastic processes (Ville, Doob), information theory (Kelly, Breiman), and sequential analysis (Wald, Robbins, Darling). Recent years have seen a flowering of research that not only picks up on classical themes, but also revisits a wide range of other problems in statistics that have traditionally been solved via p-values, tackles challenges that p-value-based statistics have not solved convincingly, and, perhaps most importantly, develops e-value-based solutions to modern problems that have arisen in high-dimensional statistics and machine learning. The articles in the current collection provide an appealing point of entry into this fast-moving and impactful literature.

Workshop Report

14th International Workshop on Computational Intelligence for Multimedia Understanding

by Behçet Uğur Töreyn (İTÜ), Maria Trocan (ISEP) and Davide Moroni (CNR-ISTI)

The 14th International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM 2026), organised by the ERCIM Working Group Multimedia Understanding through Semantics, Computation and Learning (MUSCLE), was held as a special session of IEEE ISCAS 2026 in Shanghai, China, on 26 May 2026.

The workshop brought together researchers working on multimedia understanding, multimodal learning and semantic analysis of complex data. Presentations addressed topics including multimodal data processing, machine learning for multimedia applications, intelligent sensing systems, and emerging approaches for extracting semantic information from heterogeneous data sources.



Harald Rietdijk, PhD student at ISEP – Paris Institute of Digital Technology presenting at IWCIM 2026.

Discussions highlighted the growing importance of multimodal and cross-modal analysis for enabling intelligent applications in domains ranging from industry and services to scientific research.

The workshop featured four peer-reviewed papers covering topics such as multimodal learning, semantic multimedia analysis, intelligent sensing systems and machine learning for multimedia applications. Discussions highlighted the growing importance of multimodal and cross-modal approaches for enabling intelligent applications in domains ranging from industry and services to scientific research.

Through IWCIM, the MUSCLE Working Group continues to foster col-

laboration between researchers and practitioners from academia and industry working on multimedia understanding and computational intelligence.

ERCIM Working Groups are open to researchers from ERCIM member institutions as well as interested researchers from outside ERCIM who wish to contribute to their activities.

The 15th edition of IWCIM is planned to take place in conjunction with IEEE ISCAS 2027 in Bordeaux, France, on 6–9 June 2027.

More information:
<https://iwcim.itu.edu.tr>

Spring ERCIM Meetings 2026 Held in Vienna

ERCIM warmly thanks SBA Research for hosting the ERCIM Days on 20–21 May 2026 at its offices in Vienna. Representatives from leading European research institutions gathered for two days of strategic discussions, collaborative planning, and networking focused on the future of informatics and mathematics research in Europe.

As ERCIM's Austrian member, SBA Research presented its research activities and strategic priorities to the ERCIM community. Discussions highlighted the importance of cross-border collaboration in addressing emerging challenges and opportunities in the rapidly evolving ICT landscape.

The ERCIM Board reviewed several strategic initiatives aimed at strengthening the consortium's impact and long-term positioning, including membership development, enhanced collaboration among member institutes, new oppor-



ERCIM Spring meetings participants. Photo: SBA.

tunities for industry engagement, and cooperation with ACM Europe. Particular attention was given to positioning ERCIM around key themes such as Digital Sovereignty and Security of Research.

Preparations are now underway for the ERCIM Fall Meetings in Porto (29–30 October 2026), while the Board also reaffirmed the importance of ERCIM News as a flagship communication channel for showcasing the achievements and activities of the ERCIM community.



Introduction to the Special Theme

E-values

Statistical Testing for the 21st Century

by the guest editors Peter Grünwald (CWI and Leiden University), Wouter Koolen (CWI and University of Twente) and Johanna Ziegel (ETH Zurich)

As new measurements become available over time, we face the classic problem of updating our information state. In science, this typically means refining our view of hypotheses based on experimental outcomes – either determining if the data allow us to reject a null hypothesis, or estimating which parameter values remain statistically plausible. Anytime-valid methods allow us to reliably refine these assessments sequentially while guaranteeing at most a controlled fraction of mistakes.

How do we quantify and sequentially combine this incoming evidence? E-values provide a concrete answer. In their most basic form, we use them to quantify evidence against a null hypothesis. As is standard in statistics, the null hypothesis is a probabilistic formalization of “nothing special is going on”: the medication under consideration has no effect; the pairs of outcomes we see are independent; the coin we study is unbiased, and so on. An e-variable for such a null hypothesis is a nonnegative random variable with an expectation of at most 1 under the null. When batches of data come in sequentially, we can calculate e-values sequentially for each of them and multiply them, forming what mathematicians call a supermartin-

gale, which now quantifies the total evidence in all data seen so far. A comprehensive statistical theory can be built from these ingredients. One cornerstone of this framework is Ville’s inequality, which strictly bounds the probability of this supermartingale ever exceeding a large threshold under the null. Conversely, ideas from learning theory guide us in designing powerful e-values and supermartingales with optimal growth rates under alternative hypotheses. For example, based on such e-values and supermartingales, one can construct anytime-valid confidence intervals which quantify uncertainty of a quantity of interest such as the strength of an effect (in clinical trials) or the difference in click-through rate for different web page layouts (in A/B testing).

E-values, named for both evidence and expectation, are a recent label for ideas with a much longer history. Closely related objects have appeared repeatedly across several disciplines: as nonnegative martingales and sequential tests in probability and statistics (Ville, Doob, Wald, Robbins); as likelihood ratios in statistics, universal codes in information theory, and Martin-Löf randomness tests in computability theory; and as betting capital in the game-theoretic probability framework of

Shafer and Vovk. Parallel developments in concentration inequalities, self-normalized processes, online learning, and adaptive experimentation supplied increasingly powerful methods for constructing and analyzing such processes. The modern theory of e-values may be viewed as a synthesis of these traditions, revealing a common mathematical structure underlying sequential inference, evidence accumulation, and prediction.

The justification for e-values is most sharply revealed in general decision-making problems, where feasibility and optimality considerations force their emergence. We next highlight two remarkable instances of this phenomenon. The first characterization arises in post-hoc decision making. If evidence is to remain valid when reused for arbitrary future analyses and decisions while retaining frequentist guarantees, then it must be an e-variable. In this sense, e-values provide the unique representation of reusable evidence. The second characterization arises in active sequential testing, where learning and experimentation form a feedback loop: current evidence determines the next experiment, whose outcome updates the evidence, and so on. If the goal is to reach a reliable conclusion after as few experiments as possible, then matching lower and upper bounds show that the minimal expected sample size is characterized by e-value growth rates. Thus, e-values also arise as the solution to optimal sequential design problems.

However, e-values offer more than anytime-validity. Any weighted average of e-values is again an e-value. This simple closure property provides a principled way to combine evidence from different sources, even when the data are not generated sequentially in time or are partially dependent, making e-values a fundamental tool in multiple testing. As such, they are considerably more flexible than the classical notion of statistical evidence based on p-values, for which combination over time or under dependence is significantly more delicate.

This flexibility is also reflected in the fact that, unlike classical methods, e-values can provide risk guarantees even when the Type-I loss (roughly, the cost of rejecting a true null hypothesis; strong Type-I guarantees correspond to small significance levels) is not fixed in advance but may itself depend on the data or the decision process.

Of course, flexibility may come at a price. Analyses based on e-values may sometimes require more data than classical fixed-sample procedures. Moreover, for a given statistical problem, there are typically many valid e-values, and it is not always easy to compute the optimal one in terms of the growth rates discussed above. More generally, different tasks lead to different notions of optimality, and no single choice is uniformly best. Which notion is most relevant therefore depends on the context.

Research into e-variables is rapidly advancing the core ideas underlying anytime-valid inference and, more generally, principled combination of evidence. New goals have been introduced, such as e-detectors and notions of asymptotic safety. New optimality criteria have been proposed, including those based on Rényi divergence and concave utility functions. Admissible and optimal e-variables are being characterized more precisely. In addition, connections have been established

to other areas, including probability and measure theory, group invariance, nonparametric statistics, game theory and convex duality, information theory, and online learning. This special issue brings together recent developments in the field, highlighting advances across these methodological, conceptual, and technical dimensions.

From foundations to applications

More specifically, this special issue covers both foundational and applied aspects of e-values. On the foundational side, contributions address testing by betting (**Shafer, p. 8**), the search for optimal e-values in a general setting (**Larsson, Ramdas, and Ruf, p. 9**), the relationship between e-values and p-values (**Clerico, p. 10**), the role of e-values in resolving counterfactuals (**Grünwald, p. 11**), objective probability through sequential mixture learning (**Dixit and Martin, p. 13**), and the combination of e-values with conformal prediction to obtain prediction rather than confidence intervals (**Gauthier, p. 14**).

Several articles focus on methodology, including safe procedures for multiple testing (**de Heide, p. 16; Ren, p. 17**), the derandomization of such procedures (**Monti and Filzmoser, p. 18**), tests for stochastic dominance (**Choe and Arnold, p. 20**), tests for sub-Gaussianity (**Koolen, Larsson and Agrawal, p. 21**), and near-optimal e-values for testing association (**Dickhaus, Giuffrida and Wang, p. 23**).

Applications range from clinical trials (**Long and Van Zwet, p. 25**) and replication meta-analyses in psychology (**Arias, Ly, Meziu and Reyero-Lobo, p. 27**) to the specification of clinically relevant effect sizes (**Koobs and Koning, p. 28**), the analysis of preference data in elections (**Tuynman and Mathieu, p. 29**), the assessment of AI systems (**Dhillon, Pandeva and Curth, p. 31**) and reinforcement-learning agents (**Bongers, p. 32**), and the backtesting of financial risk assessments (**Wang, p. 34**). Finally, **Michael Lindon, Netflix (p. 35)** discusses the use of e-values in the technology industry.

Together, these developments point toward a unified view of statistical evidence in which e-values play the central role. They provide new ways to address old problems, and inspire new challenges to be tackled head-on. We are excited by this prospect, and hope you will be too. Please enjoy this special issue.

Please contact:

Peter Grünwald

CWI and Leiden University, The Netherlands

Peter.Grunwald@cwi.nl

Wouter Koolen

CWI and University of Twente, The Netherlands

wmkoolen@cwi.nl

Johanna Ziegel

ETH Zurich, Switzerland

ziegel@stat.math.ethz.ch

On Testing by Imaginary Betting

by Glenn Shafer

Data analysis requires principles as well as mathematics. Traditionally, we have relied on Cournot's principle when we use probability theory for data analysis. But when we test by betting instead of relying on small probabilities, we can formulate principles that dig deeper into statistical practice and apply more broadly.

"An event with very small probability will not happen." "An event with probability close to one will happen." These aphorisms have gone by various names, including Cournot's principle, the principle of improbability, and the principle of practical certainty.

Cournot's principle does a lot of work. We predict events that have very high probability. We use the happening of events with very small probabilities to discredit theories that attribute the small probabilities to them. (Here the small probabilities are called p-values.) Augustin Cournot, depicted in Figure 1, told us that it is the only way probability theory makes contact with the world of phenomena.

But something is wrong. Events of small probability happen all the time. One of the tickets in a lottery will win. When a probability distribution is continuous, all the possible outcomes have zero probability. How do we discipline Cournot's principle? Are there further principles for its use?

Testing by betting provides a framework for formulating such principles. Instead of testing probabilities, we test a forecaster. The forecaster might be a theory, a person, or an artificial intelligence. We interpret his forecasts as betting offers. Perhaps he states probabilities and offers to sell random variables for their expected values. Perhaps he makes more limited betting offers. You bet against him by deciding which offers to accept. But the bets are imaginary. The factor by which you multiply the capital you risk is your betting score or e-value. You can make successive bets, using the capital from one bet to make the next. Your cumulative capital after multiple rounds is your cumulative betting score.

In this framework, we can refine and expand Cournot's principle into five fundamental principles.

1. Principle for testing by betting

If you make successive bets against a forecaster, beginning with a unit stake and never risking additional capital, and you obtain a large cumulative betting score, then you discredit the forecaster.

A large betting score has a small probability of happening. Our principle for testing by betting generalizes the use of p-values to discredit probabilities, because the forecaster might not state a joint probability distribution for the successive outcomes. But it also disciplines the practice, because you must bet before you see the outcomes.

How large does a cumulative betting score need to be in order to discredit a forecaster? It depends. The degree of discredit depends in part on the reputation of the players. Cautious betting based on expertise and extending over many rounds of betting is more persuasive than careless betting based merely on magical thinking.

2. Principle for multiple discredit

If you simultaneously test multiple forecasters who are forecasting the same outcome, obtaining a large betting score against each forecaster you test, then you have discredited all the forecasters.

Why? Because if you multiply each unit stake by a large factor, you have multiplied your total stake by a large factor. This principle reduces the problem of testing statistical models to the problem of testing a single forecaster.

3. Cournot's principle, disciplined

You may predict that a forecaster who has consistently withstood certain test strategies in the past will similarly withstand a similar test strategy in the future.

A strategy for successive betting determines how your evolving capital depends on forecasts and outcomes. This evolving capital as a function of forecasts and outcomes is called a supermartingale. If the strategy does not risk the capital becoming negative, the strategy is called a test strategy, and the supermartingale is nonnegative. When a nonnegative supermartingale has a large value unless an event happens, the event has small (upper) probability.

Our version of Cournot's principle reduces the aphorism that events of small probability will not happen to a type of philosophical induction. The future will be like the past. The principle is disciplined, because it applies only to test strategies that have been previously used against the forecaster.



Figure 1: Antoine Augustin Cournot (1801-1877). He is celebrated by economists for his work on duopoly, but his contribution to statistics is underrated. As the philosopher Fernand Faure wrote in 1905, it was too philosophical for the statisticians, too statistical for the philosophers. Photo source: Wikimedia.

4. Principle for point prediction

If you can replicate a payoff with a betting strategy using bets that the forecaster has withstood in the past, you may use the cost of the replication as a prediction of the payoff.

This principle generalizes and disciplines the practice of using expected values as predictions. We call predictions based on Principles 3 or 4 warranted.

5. Principle for statistical estimation

A warranted prediction about outcomes that are not observed may further warrant the inference that related unknowns are consistent with the prediction.

Suppose, for example, that you make a prediction about the errors of a measuring instrument based on experience testing the instrument to measure known quantities. If your prediction is about the errors in measuring an unknown quantity, then you will not observe the errors you are predicting, but your prediction of the average error to be within certain bounds will warrant the inference that the unknown quantity you are measuring is within certain bounds.

Probability theory was originally a theory about betting, and it is still driven by intuitions about betting. Trying to appear more objective, statistical theory has often tried to hide this underlying logic of betting. The five principles listed here show that by bringing betting back to the surface, we are able to articulate aspects of data analysis that are often hidden from students of statistics and from the public.

References:

- [1] L. Mazliak and G. Shafer, “The Splendors and Miseries of Martingales: Their History from the Casino to Mathematics”, Cham, Switzerland: Birkhäuser, 2022.
- [2] G. Shafer, “Testing by betting: A strategy for statistical and scientific communication”, *Journal of the Royal Statistical Society: Series A*, vol. 184, no. 2, pp. 407–478, 2021.

Please contact:

Glenn Shafer
Professor Emeritus, Rutgers University, USA
gshafer@business.rutgers.edu

Optimal Gambling Strategies when Testing by Betting

by Martin Larsson (Carnegie Mellon University), Aaditya Ramdas (Carnegie Mellon University), and Johannes Ruf (London School of Economics)

Under what conditions do optimal bets against a given probabilistic hypothesis exist? Answer: they always do!

John Kelly Jr., a researcher at Bell Labs, asked a simple question in 1954 [1]. Suppose we are invited to make bets on successive coin tosses at even odds. Suppose also that the coin is not fair, and we know that the probability of heads equals some value Q (eg: 0.7). Kelly asked: what gambling strategy will maximize our wealth in the long run?

Let us briefly elaborate on the game rules. Before each toss, we may bet some money on whether it will land heads or tails. If we are right, we earn double the amount that we bet, else we lose the amount that we bet. We start the game with one dollar (without loss of generality), and importantly, we can never bet more than we currently have.

Kelly did not use the language of hypothesis testing, but the connection to testing is simple. One can interpret the null hypothesis as being that of a fair coin, the alternative hypothesis as being that of a coin with bias Q . All allowed bets (double-or-nothing bets of different amounts) would not be expected to make money under the null, but they may under the alternative.

Kelly realized that if one bets smartly, the wealth can be made to grow exponentially fast in the number of rounds, and suggested maximizing that exponent. This is equivalent to maximizing the expected logarithmic wealth (i.e., the growth rate), now aptly called the Kelly criterion. A simple calculation shows that this log-optimal strategy corresponds to betting a $2Q-1$ fraction of your current wealth on heads in every round (equivalently, betting a fraction Q on heads and $1-Q$ on tails, a form of hedging).

This strategy is intuitive: if Q equals half, there is no point betting, and if Q equals one, you should go all-in in every round, and the optimal strategy linearly interpolates between these extremes. Kelly also proved that the optimal exponent (rate of wealth growth) equals the Kullback-Leibler divergence, or relative entropy, between a coin with bias Q and a fair coin.

This justified the title of his paper, which linked gambling to what was then the newly developed information theory of Claude Shannon. In the following decades, Kelly’s ideas were generalized by Leo Breiman to settings with more outcomes and varied odds, by Krizhevsky and Trofimov to handle unknown Q , and by Thomas Cover in his famous work on universal portfolios.

Recently, we asked a natural question that fully generalized Kelly’s. Suppose we are constrained to make a bet that is fair under a general class of distributions P . Then, does there still exist a log-optimal bet (against some alternative Q)?

To elaborate, suppose we start with a dollar, and our wealth after one bet is called B . Then B must have two properties: it must be nonnegative (since we cannot lose more than that starting dollar), and its expected value under any distribution P in \mathcal{P} must be at most one (we should not make money under the null hypothesis). Such a bet B is nowadays called an e-value.

Remarkably, with absolutely no assumptions on \mathcal{P} , we showed that a log-optimal bet against Q always exists [2]. Moreover, this bet, that we called the numeraire, is Q -almost surely unique. There also exists a strong duality between the expected logarithm of the numeraire under Q , and the Kullback-Leibler divergence of Q to a special sub-distribution P^* , which is the reverse information projection of Q onto \mathcal{P} (the closest element of \mathcal{P} to Q , in the information divergence). In fact, the numeraire is a likelihood ratio of Q to P^* , the latter object lying in the effective null hypothesis of \mathcal{P} (an enlargement of the convex hull).

Remarkably though, if one plays a repeated game of \mathcal{P} against Q , the optimal gambling strategy does not play the above single-round numeraire bet repeatedly. Instead, we recently showed [3] that after betting the numeraire, one should then play the numeraire bet based on the next two observations observed together, then on the next four observations together, and so on. This ensures that the wealth eventually grows at the optimal rate (calculated over all possible games that one could construct) which is larger than the Kullback-Leibler divergence between Q and P^* , but in general smaller than the infimum Kullback-Leibler divergence between Q and \mathcal{P} .

When Q is unknown, one can still design betting wealth processes (called e-processes) that are asymptotically relatively growth rate optimal. For example, the latter authors showed that in a Polish space, if \mathcal{P} is compact in the usual weak topology (a very weak assumption, not even requiring convexity), then one can design a single betting strategy that can obtain the asymptotically optimal growth rate for every Q in the complement of \mathcal{P} . This in turn implies that there exist power-one sequential tests for weakly compact \mathcal{P} against its complement, a remarkably general fact. While the authors of [2] have recently completely characterized when nontrivial fixed-sample tests exist for arbitrary \mathcal{P} versus Q , such a complete characterization for sequential testing remains open at the time of writing this article.

Martin Larsson acknowledges support from NSF grant DMS-2510965.

References:

- [1] J. L. Kelly Jr., “A new interpretation of information rate,” *Bell System Technical Journal*, vol. 35, no. 4, pp. 917–926, Jul. 1956.
- [2] M. Larsson, A. Ramdas, and J. Ruf, “The numeraire e-variable and reverse information projection,” *Annals of Statistics*, vol. 53, no. 3, pp. 1467–1493, 2025.
- [3] A. Ram and A. Ramdas, “The optimal betting wealth growth rate,” arXiv preprint arXiv:2604.25280, 2026.

Please contact:

Aaditya Ramdas, Carnegie Mellon University, USA
aramdas@stat.cmu.edu

E-values Are *Just* (Some) P-values

by Eugenio Clerico (University of Oxford)

While e-values and p-values are often presented as competitors, they share deep structural connections. We highlight a functional perspective linking the two, and suggest how it might lead to new ways of translating p-value methods into the e-value framework.

P-values are probably the most widespread tool for hypothesis testing. A p-value is a number that summarises the data: a small value indicates that the observed data are unlikely under the null hypothesis. This provides a simple way to proceed: the null is rejected if the p-value is small enough. Yet, p-values are fragile objects that need to be handled with care. Their validity comes with prescriptions... For instance, the experimental setup (dataset size, significance level) must be fixed in advance, and cannot be changed on the fly as data arrive.

E-values are a more flexible counterpart to p-values, designed to overcome these practical limitations. An e-value is again a number summarising the data, but chosen so that its expectation under the null is at most one. A valid test is obtained by rejecting the null hypothesis if the e-value exceeds a given threshold. Intuitively, e-values directly represent evidence against the null hypothesis. E-values are typically easier to merge across different experiments than p-values, and they can be updated adaptively as data arrive, allowing the experiment to be stopped at any time.

The relationship between e-values and p-values is often presented as a rivalry, framed mostly in terms of contrast. Yet, once the testing problem is fixed, they are structurally similar objects. In both cases, they represent realisations on the observed data of random variables valued in $[0, \infty]$, typically called p-variables and e-variables. (Although p-values are usually defined in $[0, 1]$, extending them to $[0, \infty]$ does not affect the test.) An α -level test based on a p-variable P rejects whenever the observed p-value is at most α , while an α -level test based on an e-variable E rejects whenever the observed e-value is at least $1/\alpha$. Choosing $P = 1/E$ (which can be shown to define a valid p-variable) yields a p-value procedure with exactly the same rejection rule as the e-value-based test.

So, are e-values *just* p-values? Well... In a way, e-values are *some* p-values. For a fixed data space and null hypothesis, we can take any e-variable E and associate with it its reciprocal $1/E$, a p-variable with the same rejection rule as the original e-variable. Thus, there is a canonical e-to-p map. Yet, this map is not surjective: its image is a strict subclass of p-variables (see Figure), consisting only of those P for which $1/P$ is valid as an e-variable. [1] refers to the members of this subset as post-hoc p-variables, emphasising that their significance level can meaningfully be chosen after observing the data. In summary, an e-value test can be canonically identified with a p-value test, while the converse is in general not true, which is the price paid for the extra flexibility of e-value testing.

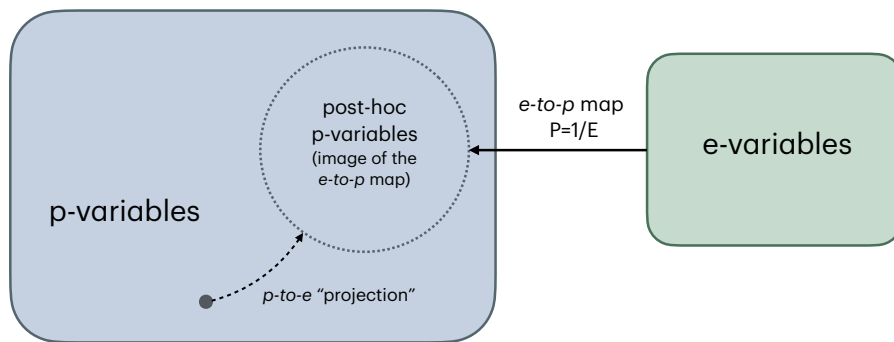


Figure 1: Representation of the canonical correspondence between e-variables and a strict subset of p-variables. Constructing a meaningful way to project any p-variable onto this subset might lead to new insights.

The connection between p-values and e-values is of course well known. However, it is usually formulated at the level of observed values, through calibrators: maps from $[0, \infty]$ to $[0, \infty]$ turning e-values into p-values (or conversely p-values into e-values) in a universal way that works uniformly for all possible null hypotheses [2]. Here we take a different viewpoint. Once the data space and the null hypothesis are fixed, p-variables and e-variables are functions on the same space, and one can ask how these classes of functions relate to each other. In the e-to-p direction, the distinction disappears: applying the reciprocal transformation to an observed e-value is the same as applying it pointwise to the whole e-variable. However, the p-to-e direction is different: an arbitrary p-variable cannot simply be inverted to obtain an e-variable. P-to-e calibrators provide universal numerical transformations from p-values to e-values. Yet, their universality comes with a loss of power, since they cannot exploit the specific structure of the given null hypothesis.

The functional viewpoint suggests a different direction, which (to my knowledge) remains largely unexplored. Since e-variables canonically correspond to a strict subclass of p-variables, one may ask whether any p-variable can be somehow “projected” onto this subclass. Such functional p-to-e conversion would be local (rather than universal), depending on the given data space and null hypothesis. Ideally, it should leave post-hoc p-variables unchanged (so that no power is lost when the p-variable already comes from an e-variable) and preserve the ordering of evidence (smaller p-values map to larger e-values). Finding a canonical principled way to define such “projections” might be challenging, yet these transformations could provide a useful bridge between p-values and e-values, helping translate existing p-value methods into the e-value framework without relying only on universal calibration.

E-values are *just* p-values is surely not an appealing slogan and would not do justice to the practical usefulness of e-values. The usual presentation remains the most natural one: defining e-values through their expected value is simple, direct, and well suited to the flexible procedures for which they were introduced. Yet, my aim is to recall a complementary interpretation and stress that working with e-values does not require leaving the p-values world behind. Rather, it means restricting attention to a subclass of p-values, characterised by stronger validity requirements, which make optional stopping, merging of evidence, and post-hoc interpretation possible. Hence, e-value theory should not be seen as discarding p-values, but as specialising p-value theory to make stronger forms of flexibility valid.

References:

- [1] N. W. Koning, “Post-hoc α hypothesis testing and the post-hoc p-value,” arXiv preprint, 2023.
- [2] V. Vovk and R. Wang, “E-values: Calibration, combinations and applications,” *The Annals of Statistics*, vol. 49, no. 3, pp. 1736–1754, 2021.

Please contact:

Eugenio Clerico
 Department of Statistics, University of Oxford, UK
eugenio.clerico@gmail.com

E-values, P-values and Counterfactuals

by Peter Grünwald (CWI and Leiden University)

A major criticism of p-values and standard confidence intervals, first coined around 1960, is their sensitivity to counterfactuals: their validity depends on how data would have been collected in situations that never occurred, which is often unknown or even unknowable. The fact that e-based methods remain valid under optional continuation implies that they do not suffer from this problem...or does it?

Suppose that a randomized clinical trial to test a new medical treatment is performed on 50 patients represented by 50-dimensional data vector $X_{(1)}$. The result turns out to be promising but not conclusive: the researchers observed a p-value $p_1 = 0.1$ while they had a significance level of 0.05 in mind. But their boss is optimistic at the news and agrees to supply the resources to test another 30 patients, resulting in data $X_{(2)}$.

Is this good news? Not if one measures evidence in the second trial by another p-value, say p_2 . For some realizations of $X_{(1)}$, it may be decided not to gather $X_{(2)}$. As a result, standard combination methods for p-values like Fisher’s cannot be employed, since they invariably require that, no matter what is observed in each sample, we always combine both. Similarly, joining the two data sets and recalculating the p-value leads to a wrong answer as well: take a method that stops for some values of $X_{(1)}$ and in that case outputs $p^* = p_1$, a sharp, nonconservative p-value for $X_{(1)}$, yet continues for other values of $X_{(1)}$ and in that case, after observing $X_{(2)}$ outputs $p^* = p'$ with p' a number strictly smaller than 1. It is easily shown that for any

such method, the resulting p^* is not a p-value – it tends to exaggerate the evidence that an effect is present.

Now, it is often countered that p-values were never meant to be used for such optional continuation. So we simply shouldn't use them here! But this is a subtle matter: in a variation of the example, suppose the researchers told their boss merely that p_1 was small enough for the result to be “promising but not conclusive,” and they did not tell its actual value. The boss, once again, feels optimistic in when he hears “promising” and suggests they continue the trial on a second batch of 30 patients.

The researchers, who know their statistics, are now worried about invalidating the results. But then news reaches the boss that the p-value is 0.1. Disappointed, he now decides to stop the trial after all – he had thought “promising” really meant something closer to 0.05 than 0.1.

Should the researchers be relieved? Perhaps surprisingly, the answer is no! [1,2] A simple calculation shows that the mere fact that the sample would have been 80 patients (thus different from the originally planned 50) in some counterfactual situation (i.e., if the first 50 data points had been different than they actually were) already makes the p-value invalid. That is, counterfactuals can ruin the validity of the p-value even if the sample plan was not in fact changed for the data which were actually observed.

As we show in [1], using e- instead of p-values avoids the above problem. This suggests that e-based methods solve the general conundrum about counterfactuals. But, it turns out, they only do this if the counterfactuals involved are related to time. For counterfactuals related to censoring, they don't:

Pratt's (1962) voltmeter

Suppose we observe X_1, \dots, X_n where n is fixed and the X_i represent voltages of electron tubes, measured with an accurate voltmeter. A statistician examines the X_i assuming they are normally distributed with fixed variance and some mean μ . He aims to use a p-value to measure the evidence against the null hypothesis $\mu = 4$. Later he visits the engineer's laboratory, and notices that the voltmeter reads only as far as 6 (Figure 1). Even though none of the X_i were ≥ 6 this makes the standard p-value invalid; it necessitates a new calculation that takes into account the (potential, counterfactual!) censoring [L1,1,2,3].

However, the engineer then says she also has a super-high-range-meter, equally accurate, which she would have used if any of the measurements had turned out 6. This is a relief to the statistician, because it means the original p-value is correct after all. But the next day the engineer telephones and says “I just discovered my high-range voltmeter was not working the day I did the experiment”. The statistician then informs her that a new analysis will be required after all!

The engineer is astounded. She says, “But the experiment turned out just the same as if the high-range meter had been working. I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope!”

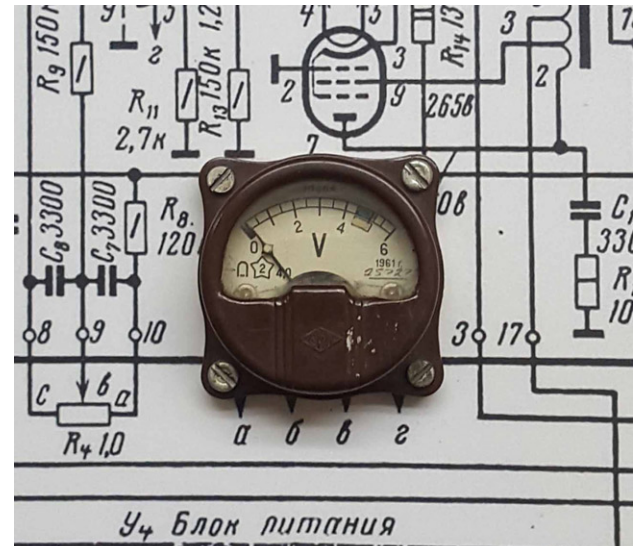


Figure 1: A voltmeter prone to censoring, from the era in which Pratt came up with the example. Photo by Andrey Shtanko CC BY 4.0.

Unknown Unknowns

As we show in [1], variations of this problem do affect e-methods: they might also have to be re-defined in terms of counterfactual censoring. Nevertheless, both the voltmeter and the optional continuation story have a similar flavour: one first obtains less precise information (50 data points; or censored measurements). Then, depending on the value of this initial data, one may or may not decide to get more precise information (additional data points; uncensored measurements). This suggests that extensions of e-values that deal with such more general counterfactuals are possible – this is one of the projects that I suggested to tackle in my honoured ERC Advanced Grant. For this, we need to formalize the intriguing notion of observing the outcome of a random variable without knowing the definition of that random variable, which we managed to do in [1]. Yet, much work still needs to be done; specifically, one needs to be able to employ, within the same mathematical formula, random variables that have known definitions and random variables that have unknown definitions. This is still a major challenge – which may require tools from epistemic logic rather than probability and statistics!

Links:

[L1] <https://www2.stat.duke.edu/~st118/sta732/PrincHO.pdf>

References:

- [1] B. Chugg, A. Ramdas, and P. D. Grünwald, “E-values as statistical evidence: A comparison to Bayes factors, likelihoods, and p-values,” arXiv preprint arXiv:2603.24421, 2026.
- [2] E. J. Wagenmakers, “A practical solution to the pervasive problem of p-values,” *Psychonomic Bulletin & Review*, vol. 14, no. 5, pp. 779–804, 2007.
- [3] L. J. Savage, G. Barnard, J. Cornfield, et al., “On the foundations of statistical inference: Discussion,” *Journal of the American Statistical Association*, vol. 57, no. 298, pp. 307–326, 1962.

Please contact:

Peter Grünwald, CWI and Leiden University The Netherlands pdg@cwi.nl

E-processes, Predictive Recursion, and Objective Empirical Probability

by Vaidehi Dixit (University of Nottingham) and Ryan Martin (North Carolina State University)

Good e-processes can be constructed for testing a specific null hypothesis against a specific alternative, but general inference need not have a specific alternative in mind. In such cases, one might seek an alternative hypothesis-agnostic e-process with fast growth rate under a wide range of alternatives. Our predictive recursion-based e-process construction offers just that, along with some deeper insights related to “objective” empirical probability.

E-processes are a hot topic in the modern statistics, machine learning, and data science landscape, thanks to their strong frequentist reliability properties, namely, anytime validity. But e-processes actually have a rather long history. Indeed, Neyman & Pearson constructed e-processes in their landmark paper on hypothesis testing and, likewise, Jeffreys advocated for the use of Bayes factors which, in certain cases, are e-processes. What these two seemingly different approaches have in common is that their respective e-processes correspond to likelihood ratios, and this connection between e-processes and likelihood ratios has been established in increasing generality in recent years [1, 2]. Roughly, this likelihood ratio has a numerator and denominator corresponding to the “best representatives” of the evidence of the alternative and null hypothesis, respectively. For example, in the classical Neyman–Pearson case where the null and alternative hypotheses are simple, these “best representatives” are trivial, namely the corresponding probability distributions. More formally, for a general composite null versus a composite alternative hypothesis, constructing these “best representatives” is more challenging; see the above references for details.

The e-process constructions just described are tailored to a particular testing problem, i.e., given null and alternative hypotheses. In the Fisherian spirit of probing the data for “significant” explanations that warrant further exploration, as op-

posed to formally testing a null against a specified alternative, one might ask if there is a generic or “alternative-hypothesis agnostic” e-process. Like with the universal inference framework of Wasserman, Ramdas, and Balakrishnan, we can assume that candidate explanations determine the denominator of the e-process likelihood ratio, hence only the numerator likelihood needs to be specified. Since a likelihood can be factored as a product of successive predictive distributions, i.e., the distribution of data point X_k , given X_1, \dots, X_{k-1} , for $k=1, \dots, n$, this boils down to specification of a forecasting system. Many forecasting systems are possible, but an essential requirement is that it be flexible enough to adapt to data produced by any one of a wide collection of possible “true distributions”. A secondary requirement is that the predictive distribution can be updated efficiently given a previous predictive distribution and new data.

We highlight one such predictive scheme here, the so-called Predictive Recursion (PR) algorithm that is a fast, recursive and nonparametric algorithm that estimates the underlying mixing distribution of a general mixture distribution. This is important here on two accounts. First, since the algorithm fits a general mixture density, e.g., a location–scale mixture of Gaussian distributions, it is sufficiently flexible to capture a wide collection of true distributions. Second, PR is a recursive algorithm which means that it efficiently learns the mixing distribution and hence updates the mixture density estimate and corresponding likelihood as new data arrive. With these two crucial points, we construct a PR-based e-process called a PRe-process for testing a composite null family of probability distributions versus a composite alternative family. This uses the PR-based marginal likelihood in the numerator of the e-process and a supremum under the null in the denominator. This construction satisfies the anytime-valid properties of e-processes and the value of the PRe-process grows optimally if the null is indeed not true as the sample size approaches infinity; see [3]. For a practical illustration, we consider the earthquake locations in the state of Oklahoma from 2000–2011 and model them as a spatial Poisson point process. There is suspicion of fracking activities in this region since 2009 and we investigate this by testing for whether the post-fracking intensity function is proportional to the pre-fracking intensity function via the PRe-process. Figure 1 shows that the log PRe-process shows an increasing trend: so, there is an accumulation of ev-

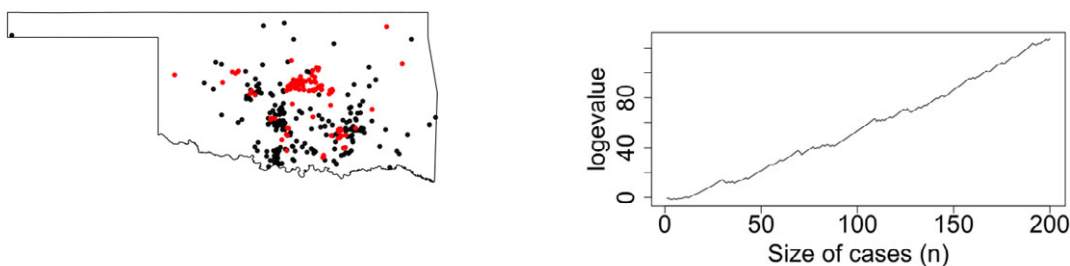


Figure 1:
(a, left) A random subset of the locations of earthquakes used from the earthquake locations in the state of Oklahoma from 2000–2011. The black points denote earthquakes up to 2008, and the red points denote earthquakes from 2009–2011.
(b, right) Plot of $\log(\text{PRe-process})$ versus size of cases (n) for testing whether intensity of earthquakes post-fracking (after 2009) is proportional to intensity of earthquakes pre-fracking (before 2009)

idence against the null hypothesis and hence a significant effect of fracking.

Beyond their use in the construction of e-processes as outlined above, predictive schemes apparently have a fundamental role to play in the broader context of uncertainty quantification. The specific question we have in mind here, going back at least to Keynes and Carnap, is if one can define an “objective” empirical probability associated with an observed data sequence. Building on the classic work of von Mises, Phil Dawid set out to define a notion of objective probability through what he called calibrated forecasting systems – PR defines a calibrated forecasting system under relatively mild conditions. An interesting conclusion Dawid drew from the famous Blackwell-Dubins theorem is that, while calibrated forecasting systems are incredibly diverse, the actual probabilities they assign to the observed data sequence must agree asymptotically; hence one can define a genuinely “objective” empirical probability as the common forecast value assigned by all calibrated forecasting systems asymptotically. The efforts by Vladimir Vovk and others to generalize Dawid’s results to finite sequences, etc., were surely influential to the development of Shafer and Vovk’s game-theoretic probability, which, coming full circle, provides a deeper foundation for the latest e-process advancements. This forecasting system perspective and “objective probability” insight opens new opportunities for PR as a genuine “model” for observations, meaningful and flexible, rather than as a method for estimating unknown mixing distributions. Suffice it to say, PR and the corresponding PRe-process have yet to realize their full potential.

Links:

[L1] <https://kwz.me/hIy>

[L2] <https://www.ou.edu/ogs/research/earthquakes/catalogs>

References:

- [1] P. Grünwald, R. de Heide, and W. M. Koolen, “Safe testing,” *Journal of the Royal Statistical Society: Series B*, 2024.
- [2] M. Larsson, A. Ramdas, and J. Ruf, “The numeraire e-variable and reverse information projection,” *The Annals of Statistics*, 2025.
- [3] V. Dixit and R. Martin, “Anytime valid and asymptotically optimal inference driven by predictive recursion,” *Biometrika*, 2025.

Please contact:

Vaidehi Dixit

University of Nottingham, United Kingdom

Vaidehi.Dixit@nottingham.ac.uk

Rethinking Conformal Prediction for Constrained Environments

by Etienne Gauthier (Inria, Ecole Normale Supérieure, Université PSL)

How can we trust a model’s predictions in the presence of uncertainty? Conformal prediction provides a principled framework for attaching reliable confidence guarantees to machine learning outputs. By incorporating e-values, this framework moves beyond rigid, pre-specified guarantees. It enables finer control over the predictions in a dynamic way, adapting the reliability of AI systems to the constraints of real-world applications.

As machine learning models are increasingly deployed in high-stakes environments, such as autonomous navigation, finance, and healthcare, quantifying their uncertainty is just as critical as the predictions themselves. Conformal Prediction (CP) has become a gold standard for this task. Practically, CP acts as a rigorous mathematical wrapper applicable to any predictive model. It relies on a dedicated “calibration set” of held-out data. By evaluating a non-conformity score across both the calibration and test data, and relying on the foundational assumption that these data points are exchangeable, we can easily obtain a prediction set by calculating the associated quantile.

Instead of generating a single, potentially overconfident guess, CP provides a guaranteed set of possibilities. For instance, rather than simply predicting “dog,” an image classifier using CP with a 99% confidence level might output {dog, cat}, guaranteeing that the true label is within this set 99% of the time.

While classical conformal prediction is highly effective, it enforces a fixed confidence level by allowing the size of the prediction set to adapt to the model’s underlying uncertainty. This adaptive behaviour is essential to its guarantees, but in some applications it may lead to unpredictable or undesirable variation in prediction set size.

Consider a motivating example in the medical field: a doctor using an AI diagnostic tool. If the model is highly uncertain about a complex case, classical conformal prediction might output a set of 15 possible diseases to maintain a strict 99% coverage guarantee. Yet in practice, the doctor faces physical, financial, and temporal constraints that make such a broad set difficult to act upon. For example, slightly relaxing the target coverage level (for instance, to 98%) would already lead to a much smaller, more actionable prediction set of around five candidates (see Figure 1). This illustrates the need for more flexible approaches that explicitly trade off coverage and set size in a controlled way, while remaining useful in real-world decision-making.

To address this, we turn to a more flexible statistical framework based on e-values, which enables post-hoc control of uncertainty. This flexibility allows us to move beyond fixed confidence levels and adapt inference after observing the data.

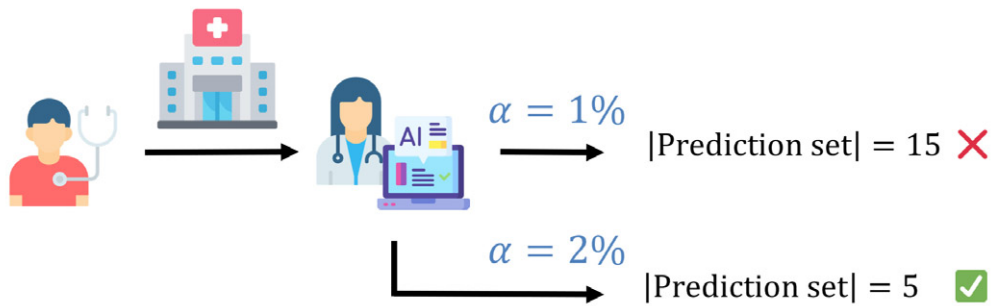


Figure 1: Illustration of the trade-off between coverage and prediction set size in conformal prediction. Relaxing the target miscoverage level from $\alpha=1\%$ to $\alpha=2\%$ can substantially reduce the size of the prediction set, yielding more actionable predictions in constrained medical settings.

Unlike traditional p-values, which lose their statistical validity if you modify your parameters after observing the data, e-values allow us to audit and adjust our statistical guarantees retrospectively. Importantly, e-values allow us to derive adaptive coverage levels on the fly, without necessitating additional data or complex data splitting. We can seamlessly achieve this simply by using the exact same calibration set already available in classical CP. They give us the mathematical right to look at our predictions first, apply constraints, and then modify our confidence levels post-hoc while maintaining perfect theoretical validity.

Leveraging this post-hoc power, our project has introduced an approach called Backward Conformal Prediction (BCP) [1]. This method allows practitioners to impose a strict “hard constraint”: a maximum allowed size for the prediction set.

If our doctor can only reasonably manage a maximum of 3 clinical exams, BCP ensures the AI will never suggest more than 3 diseases. Because e-values permit post-hoc significance testing without penalty, we can first limit the set to match our physical constraints, and then retrospectively calculate the highest mathematical confidence level we can safely attribute to that specific, constrained prediction.

In broader operational scenarios, a strict limit on every single prediction might be unnecessarily rigid. This is where adaptive coverage policies [2] come in. Using the same theoretical foundation, this approach applies a “soft constraint,” which instead provides a guarantee on the expected (or average) set size at test time.

For example, a hospital network might require the AI to recommend an average of 3 tests per patient to align with laboratory capacity over a fiscal quarter. E-values allow us to dynamically adjust the post-hoc certainty so that the predictions

seamlessly adapt, ensuring the average size constraint is rigorously met across the entire patient population.

We initiated this research in 2025 at Inria, Ecole Normale Supérieure, PSL University in collaboration with my PhD supervisors Francis Bach and Michael I. Jordan. This work represents a step forward from purely theoretical safety to actionable, real-world reliability. Classical conformal prediction provides statistical guarantees, but this may sometimes result in prediction sets that are too large to be practically actionable. Our methods invert this paradigm, starting from actual physical constraints and using the post-hoc flexibility of e-values to extract the best possible statistical guarantee for those specific limits.

Moving forward, we are eager to explore practical deployments of this method across various industrial and clinical pipelines. By bridging the gap between rigorous statistics and operational realities, our ultimate goal is to improve what people do in practice, making trustworthy AI accessible, scalable, and genuinely effective for heavily constrained environments.

References:

- [1] E. Gauthier, et al., “Backward Conformal Prediction”, in *Advances in Neural Information Processing Systems* (2025).
- [2] E. Gauthier, F. Bach, and M. I. Jordan, “Adaptive coverage policies in conformal prediction”, in *International Conference on Artificial Intelligence and Statistics*, 2026.

Please contact:

Etienne Gauthier
Inria, Ecole Normale Supérieure, Université PSL, France
etienne.gauthier@inria.fr

More Discoveries and Flexibility in Multiple Testing

by Rianne de Heide (University of Twente and CWI)

Modern data analysis can test thousands of scientific questions at once, from genes in cancer studies to voxels in brain scans. A new general principle called e-closure gives researchers more freedom to explore these results after seeing the data, while keeping false discoveries under control.

In many areas of science, a single experiment no longer asks a single question. A cancer researcher may test tens of thousands of genes for a difference between tumour types. A neuroscientist may test activity in more than one hundred thousand brain locations. A climate scientist or computer scientist may scan huge collections of variables, models or configurations. Each individual test may look innocent, but together they create a simple danger: many discoveries will be false positives.

Statistics has developed safeguards for this situation, known as multiple testing methods. One classic safeguard is very strict: try to avoid even one false alarm (Family-Wise Error Rate control – FWER). This is useful in confirmatory studies, but often too conservative for discovery science. A more practical rule, now standard in genomics and many other fields, controls the false discovery rate (FDR). Informally, this means that among the reported discoveries, only a controlled fraction is expected to be false. It is the statistical reason why a long list of genes can be taken seriously enough for follow-up experiments.

But there is a catch. The guarantee is usually attached to the list produced by a pre-specified method. In practice, scientists rarely stop there. They look at the list, compare the discoveries with biological knowledge, study effect sizes, remove uninteresting items, split results into clusters or pathways, and choose a smaller set for expensive validation. This is sensible scientific behaviour, but it falls outside the mathematical guarantee.

A familiar example is the volcano plot, widely used in molecular biology and oncology (Figure 1). Each point represents a gene or other molecular feature. One axis shows how statistically surprising the result is; the other shows how large the estimated effect is. The most attractive discoveries appear in the top left and top right corners: both statistically convincing and practically large. Researchers often first apply an FDR-method and then, using the volcano plot, keep only the most extreme effects. The visual intuition is compelling. Unfortunately, this second filtering step can quietly yet severely break the original error guarantee. A subset of a well-con-

trolled discovery list is not automatically well controlled. Ebrahimipour and Goeman showed that this can substantially inflate the false discovery rate in realistic genomic settings [1].

Our work asks whether researchers can have both things at once: rigorous error control and the freedom to interact with the results. The answer is yes, through a new unifying principle called e-closure.

The idea builds on two lines of work. The first is the classical closure principle for multiple testing. For several decades, closure has provided a complete recipe for FWER controlling methods. Later work extended this type of thinking to other error measures (but not FDR). The second ingredient is the e-value, a modern measure of evidence that behaves particularly well when evidence is combined, accumulated over time, or inspected adaptively. Unlike the p-value, which asks how surprising the data would be under a null hypothesis, an e-value can be read as a kind of betting score: large values are evidence against the null, while the rules of the game prevent systematically exaggerated evidence when the null is true.

E-closure combines these ideas. On top of designing one test for each separate hypothesis, we design evidence summaries for groups of hypotheses. The principle then turns these summaries into a menu of discovery sets that are all valid simultaneously. The researcher is not forced to report exactly one pre-ordained list. After seeing the data, they may choose from the menu, for example to focus on genes with larger effects, to report a biologically meaningful pathway, or to separate a brain-imaging result into anatomical regions.

This is the key conceptual shift. Traditional FDR procedures give: “Here is the list of discoveries; do not modify it if you want to keep the guarantee.” E-closure gives: “Here is a menu of lists; you may choose from this menu after looking, and the guarantee still holds.” In the volcano plot setting, this means

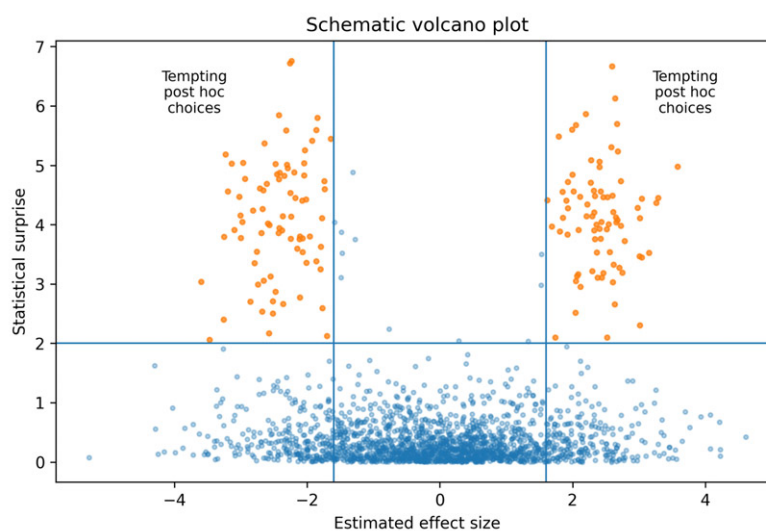


Figure 1: Schematic volcano plot illustrating double filtering (synthetic data). Each point represents one feature, such as a gene. The horizontal axis shows the estimated effect size, while the vertical axis shows statistical surprise. The tempting top-left and top-right corners contain results that look both statistically significant and practically large. Although these points are often treated as the most interesting findings, selecting them after applying a multiple-testing procedure can invalidate the original FDR guarantee. E-closure aims to make such post hoc choices part of the valid statistical analysis.

that post hoc filtering can be made part of the valid procedure rather than an invalid step after it.

The principle is not merely a new method among many. Our technical result shows that e-closure is necessary and sufficient for a broad class of multiple-testing guarantees based on expected error. In particular, it covers FDR control, and it recovers the classical closure principle for FWER as a special case [2, 3]. This gives the field a common theory for methods that previously looked unrelated.

It also gives practical benefits. When existing procedures such as Benjamini-Hochberg or Benjamini-Yekutieli are expressed through e-closure, they can be uniformly improved: the new version never makes fewer discoveries and sometimes makes more, while preserving the same error guarantee [2, 3]. In real-data examples in our technical paper, closed versions of established procedures find – sometimes many more – additional discoveries at the same target error level.

There are computational challenges, because in principle all groups of hypotheses could be considered: exponentially many in the number of hypotheses. Yet the same was true for classical closure, and decades of work have produced shortcuts for important cases. Our paper gives polynomial-time algorithms for several useful procedures and points to further algorithmic work.

For applied scientists, the promise is that exploratory analysis need not be statistically fragile. For theoreticians, e-closure offers an elegant general theory. For users of volcano plots, pathway analyses and brain maps, it suggests a future in which interactive scientific judgement and formal error control can be used together, rather than treated as opposing goals.

The work described in this article is based on joint research with Jelle Goeman (Leiden University Medical Center), Aldo Solari (Ca' Foscari University of Venice), Ziyu Xu (Carnegie Mellon University), Lasse Fischer (University of Bremen), and Aaditya Ramdas (Carnegie Mellon University).

Links:

- [L1] <https://arxiv.org/abs/2509.02517>,
<https://github.com/neilz xu/eclosure> (python)
<https://kwz.me/hId> (R package by Jelle Goeman)
- [L2] Active Volcano Plot software:
<https://github.com/mitra-ep/ActiveVolcanoPlot>

References:

- [1] M. Ebrahimpour and J. J. Goeman, “Inflated false discovery rate due to volcano plots: problem and solutions,” *Briefings in Bioinformatics*, 22(5), 2021.
- [2] Z. Xu et al., “Bringing Closure to False Discovery Rate Control: A General Principle for Multiple Testing,” *arXiv:2509.02517*, 2026.
- [3] J. Goeman, “A Uniform Improvement of the Benjamini-Hochberg Procedure using e-Closure”, *arXiv preprint arXiv:2606.01854*, 2026.

Please contact:

Rianne de Heide
University of Twente and Centrum Wiskunde & Informatica,
The Netherlands, r.deheide@utwente.nl

Combining Rejection Sets with False Discovery Control

by Zhimei Ren (University of Pennsylvania)

How can we combine discoveries from multiple FDR-controlling rejection sets without losing statistical validity? In joint work with Rina Foygel Barber, we show that the knockoff procedure can be represented through e-values and e-BH, allowing rejection sets from multiple randomized runs to be aggregated by averaging e-values while preserving FDR control. More broadly, this e-value perspective provides a general framework for merging FDR-controlling procedures and opens new directions for understanding, improving, and aggregating large-scale testing methods.

Modern statistical applications often require testing multiple hypotheses simultaneously – for example, testing which features from a large pool are associated with an outcome of interest, or testing which outcomes are significantly affected by an exposure. In such large-scale testing problems, it is desirable to control the errors among the rejected hypotheses to improve reliability and replicability. A widely used error metric is the false discovery rate (FDR), defined as the expected proportion of the number of false rejections among all rejections.

Since the introduction of FDR, many multiple testing methods have been developed to produce rejection sets with FDR control guarantees. However, it remains unclear how to combine separately obtained FDR-controlling rejection sets while preserving FDR control for the aggregated set. This problem arises in many practical settings: for instance, a randomized multiple testing procedure may produce different rejection sets under different random seeds, or multiple research teams may test the same hypotheses and report different discoveries. How can these discoveries be aggregated in a statistically valid way?

A natural approach is to combine the test statistics underlying the individual rejection sets and then apply a multiple testing procedure to the combined statistics. The challenge is that these test statistics across different runs or teams may exhibit complex dependence structures – for example, when they arise from multiple runs of a randomized procedure or when different teams use overlapping samples in their studies. As a result, the distribution of the combined statistics is generally unclear, making FDR control difficult to establish.

In a work joint with Rina Foygel Barber, we study this problem in the context of feature selection with FDR control. Specifically, we consider the knockoff procedure [1], a randomized multiple testing procedure that produces rejection sets with FDR control. To combine rejection sets from different random knockoff runs, we first show that the knockoff procedure can be equivalently represented as the e-BH procedure [3] applied to a set of special e-values. We then average these e-values across runs for each hypothesis and finally apply the e-BH procedure to the resulting aggregated e-values.

Since the e-values are defined as nonnegative random variables with expectation at most one under the null hypothesis, an average of valid e-values remains an e-value. In addition, the e-BH procedure is guaranteed to produce rejection sets with FDR control whenever the individual e-values are valid. Therefore, since averaged knockoff e-values are still valid e-values, the final rejection set obtained by applying e-BH also has FDR control guarantees.

This workflow turns out to extend well beyond the knockoff procedure: it can be shown that any FDR-controlling multiple testing procedure can be equivalently represented as e-BH applied to a suitable collection of e-values. This perspective provides a flexible framework for merging rejection sets without sacrificing FDR control. It also opens several research directions, including the design of more powerful e-value representations for existing procedures, principled methods for efficiently aggregating discoveries across randomized or distributed analyses, and new ways to understand the structure and limitations of FDR-controlling procedures through the lens of e-values.

References:

- [1] Z. Ren and R. F. Barber, “Derandomised knockoffs: Leveraging e-values for false discovery rate control”, *J. Roy. Stat. Soc. Ser. B (Stat. Methodol.)*, vol. 86, no. 1, pp. 122–154, 2024.
- [2] E. Candès, et al., “Panning for gold: ‘Model-X’ knockoffs for high-dimensional controlled variable selection”, *J. Roy. Stat. Soc. Ser. B (Stat. Methodol.)*, vol. 80, no. 3, pp. 551–577, 2018.
- [3] R. Wang and A. Ramdas, “False discovery rate control with e-values”, *J. Roy. Stat. Soc. Ser. B (Stat. Methodol.)*, vol. 84, no. 3, pp. 822–852, 2022.

Please contact:

Zhimei Ren
University of Pennsylvania, United States of America
zren@wharton.upenn.edu

E-Values for Stable and Robust Variable Selection in Microbiome Studies

by Gianna Serafina Monti (University of Milano-Bicocca), and Peter Filzmoser (TU Wien)

E-values provide a principled foundation for false discovery rate control in high-dimensional microbiome data analysis. We show how their aggregation properties enable a derandomized, robust knockoff filter that outperforms classical approaches in stability and reproducibility.

Identifying which bacterial species in the gut microbiome are genuinely associated with clinical outcomes, such as obesity, liver disease, or cancer, is a central challenge in modern computational biology. The statistical hurdles are formidable: microbiome data are compositional (only relative abundances are observed), high-dimensional (far more species than subjects), sparse (few species are truly related to outcomes), and frequently contaminated by outliers. Clinical studies also routinely measure several outcomes simultaneously, calling for multivariate methods rather than repeated univariate analyses.

A recurring challenge in this setting is controlling the false discovery rate (FDR), the proportion of spurious associations among those reported, while maintaining adequate power to detect true signals. P-value-based FDR control often relies on restrictive assumptions about the data’s dependence structure – assumptions that are frequently violated in microbiome studies – whereas e-values provide a more flexible framework. Their natural closure property under averaging enables robust error control even in high-dimensional settings with complex, unknown dependencies. An e-value is a non-negative number that quantifies evidence against a null hypothesis: a value of 1 indicates no evidence, while large values indicate strong evidence. Unlike p-values, e-values can be freely averaged and combined across independent experiments without compromising their validity as evidence measures. This closure property under averaging is the key to a new generation of derandomized procedures: by aggregating e-values across repeated knockoff constructions, we achieve a level of selection stability and reproducibility previously unattainable with p-value-based filters.

The knockoff framework and its randomness problem

The knockoff filter is a powerful approach to variable selection with FDR control. It works by augmenting the dataset with synthetic copies of the predictors, the “knockoffs”, that mimic the correlation structure of the originals but carry no signal. Comparing the regression coefficients of the original and knockoff features yields evidence scores, and a data-driven threshold selects features above a level that provably controls the FDR. The Multi-Response Knockoff Filter (MRKF) of Srinivasan et al. (2023) extends this idea to multivariate response settings, making it applicable to studies with multiple clinical outcomes.

However, the classical knockoff filter has a well-known Achilles’ heel: its inherent randomness. Because the knockoff

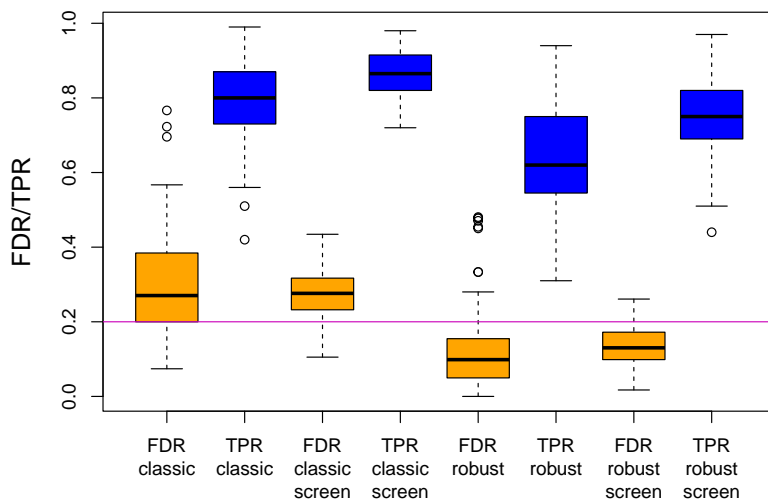


Figure 1: Simulation results for the FDR and TPR (true positive rate) when comparing the non-robust (classical) and robust procedures, with and without an initial screening step.

matrix is inherently stochastic, repeated analyses of the same dataset yield different sets of selected variables – an undesirable property in scientific practice, where reproducibility is paramount. This instability is not a minor inconvenience; it can undermine trust in reported findings and complicate downstream biological interpretation.

E-values as the bridge to derandomization

We show that e-values provide an elegant and principled solution to this instability [2]. The key idea is to run the knockoff filter M times, each with an independently sampled knockoff matrix, and to convert each run’s selection into e-values for each feature. Specifically, for each feature and each knockoff copy, an e-value is defined that is large when the feature is selected and small otherwise. Because these e-values are independent across runs, their average is also a valid e-value – a direct consequence of the martingale structure underlying e-value theory. The averaged e-values are then fed into the e-BH procedure [3], which controls the FDR under arbitrary dependence among hypotheses. The result is a single, stable selection set that aggregates evidence across multiple randomizations without inflating the FDR.

This derandomized procedure, which we call RobMRKF-Derand, inherits all the finite-sample FDR guarantees of the knockoff framework while dramatically reducing variability across runs. The e-value aggregation step is not merely a computational trick. It reflects a genuine accumulation of statistical evidence, consistent with the broader e-value literature on sequential and anytime-valid inference.

Robustness and compositionality

We embed this derandomized knockoff filter within a robust multivariate regression framework for compositional covariates. As the microbiome predictors are compositions and thus live in the simplex sample space, they are first mapped to the Euclidean space via the additive log-ratio (alr) transformation. Regression and covariance estimation are then performed jointly using an adaptation of the robust multivariate lasso with covariance estimation from Chang and Welsh (2023) [1], replacing the standard squared loss with Tukey’s biweight loss to down-weight outlying observations. The combination of ro-

bust estimation and e-value-based derandomization yields a method that is both resistant to contamination and stable across repeated analyses.

Results

Simulation studies confirm that RobMRKF-Derand controls the FDR at the nominal level across a range of contamination scenarios while achieving higher power and substantially lower selection variability than the non-robust MRKF. In real microbiome data applications, the method identifies biologically interpretable associations between microbial species and clinical parameters, findings that are reproducible across runs, unlike those of the classical knockoff filter.

The figure presents simulation results for the FDR and TPR (true positive rate) when comparing the non-robust (classical) and robust procedures, with and without an initial screening step. In contrast to the classical method, the robust procedure keeps the nominal FDR level of 0.2 in most cases (for details, see [2]).

Links:

[L1] R code, synthetic data illustrations, and documentation: <https://github.com/giannamonti/RobMReg>.

References:

- [1] L. Chang L, A. H. Welsh “Robust multivariate lasso regression with covariance estimation”. *J Comput Graph Stat*, 32, 2023
- [2] G.S. Monti, M. Pujolassos, M. Calle Rosingana, P. Filzmoser, “Robust multivariate regression controlling false discoveries for microbiome data”, *Bioinformatics*, 41, 9, 2025
- [3] Z. Ren, R. Barber, “Derandomised knockoffs: leveraging e-values for false discovery rate control,” *J. R. Stat. Soc. B*, 86, 2024

Please contact:

Gianna Serafina Monti
University of Milano-Bicocca
gianna.monti@unimib.it

Betting on Bets: Comparing the Upsides of Uncertain Prospects over Time with E-values

by Yo Joong Choe (INSEAD) and Sebastian Arnold (CWI)

New research develops novel statistical methodology, based on e-values, for monitoring whether one uncertain prospect (say, a new investment option) has an upside over another (say, the current investment). These e-values can flexibly and effectively test for multiple notions of upside over time, as defined by the decision-maker's preferences, and they come with a direct monetary interpretation that guides the decision (say, whether to invest in the new option).

Imagine you are presented with two recurring “bets” whose outcomes are uncertain. One of these bets is the financial asset that you currently invest in, say a government bond; the other is a riskier asset that you were keeping your eye on, say a lesser-known stock that you think may have an upside. Quick research tells you that the average historical returns of these two assets have been nearly identical. Unsure what to do, you decide to monitor these prospects over time to answer the question: is there enough evidence that there is an upside to this new “bet” over the old one?

This upside question, over a pair of uncertain prospects, is not just relevant to financial decision-making. We could also compare the performance of a new football player over a comparable player on the squad; the carbon offset levels of a new policy over an existing one for a business; or even the efficacy of a newly approved drug over an existing one across different subpopulations. While the alternatives in the latter scenarios

may not initially look like “bets,” they are, in reality, uncertain prospects whose relative upsides need to be monitored carefully before we commit to a decision.

In a recent work [1], together with Marco Scarsini and Ilia Tsetlin, we found that various forms of this upside question can be answered effectively using e-values [2] as our measure of statistical evidence. The first key benefit of using e-values is that, instead of a binary decision at a fixed sample size, we can compute a sequentially accumulating evidence measure for the relative upside of a prospect. The resulting test permits continuous monitoring – a procedure that invalidates classical p-values – allowing the decision-maker (DM) to make informed decisions at any time point.

Another benefit of e-values is that it naturally comes with an intuitive monetary interpretation. Consider the simplest form of the upside question: is the new financial asset more likely to yield a positive return than the old one? Now, suppose that a bookmaker offers us a hypothetical betting game. For every euro we invest, we win an extra euro if the new prospect nets a positive return while the old prospect yields a negative return; we lose the euro if the opposite happens. (It is a “push” if both go positive or both go negative.) The rules of this “bet on bets” hint at the bookmaker’s implicit hypothesis: the new prospect is not more likely to yield a positive return than the old one.

This betting game reveals a key insight underlying e-values. If we were to start with a euro and repeatedly bet a fraction of our money on this hypothetical gamble, then we can only expect to grow our wealth if the bookmaker’s hypothesis is wrong and the new prospect is indeed more likely to yield a positive return than the old one. In the testing-by-betting framework [2], the amount of money we make from this gamble is our evidence against the bookmaker’s claim, and we refer to this wealth as the e-value. It would be very surprising to make, say, 100 euros in this betting game, regardless of how



Figure 1: Given a pair of uncertain prospects, the e-value developed in new research quantifies statistical evidence for the relative upside of one prospect over another, and it grows exponentially fast when there is such an upside. The left panel shows histograms for the returns of two hypothetical financial assets (say, A=Bond and B=Stock). The right panel shows e-values for testing the relative upside of either asset as more observations arrive (in pairs). Note that the vertical axis is logarithmic. Given the high upside of asset B (higher probability of achieving a 2% return), the e-value for B grows exponentially fast, while the e-value for A does not. The dashed lines indicate the rejection thresholds for statistical testing at significance levels 1%, 5%, and 10%.

cleverly we make our bets, if the new prospect in fact had no upside over the old one. If needed, we can connect this notion back to classical statistics: rejecting the null hypothesis at a significance level of 1% corresponds to rejecting it after making 100 times the initial wealth.

In our work, we derive an optimal betting strategy for this game that maximizes the expected growth rate of the e-value whenever, in reality, the bookmaker's hypothesis is incorrect. This means that, when using this strategy, we can expect to detect an upside as quickly as possible – if there is one.

What about the broader question of whether the new prospect has an upside at some threshold, say the probability of a return above X% for an unspecified X? To answer this question, we leverage the third benefit of e-values: we can flexibly combine multiple e-values for simpler questions to answer a more complex one. Now imagine playing multiple betting games across different return thresholds, but with the same initial budget of one euro. At each round, we can split our money across all these games, and we make our bet in each game with the money allocated to that game. Then, after any given round, our combined wealth quantifies accumulating evidence for the broader upside question.

The main result of our work shows that, when using optimal bets for each game, and any reasonable strategy to split money across games, our combined wealth will grow exponentially fast, eventually to infinity, whenever there is a relative upside to the new prospect at some threshold. Figure 1 illustrates this on simulated financial returns: whenever an asset has a clear upside at an unspecified threshold, its associated e-value grows exponentially fast over time.

The research further makes advances on two related problems. First, it extends the methodology to other notions of upside [3], such as upsides according to risk-averse DMs. Second, it further extends the methodology to a more challenging question: can we monitor whether a new prospect has definite upside over the old one? A promising future direction is to apply these tools to monitor for upside, flexibly and in real time, across applications like online experimentation, public policy, and sports analytics.

References:

- [1] S. Arnold, Y. J. Choe, M. Scarsini, and I. Tsetlin, "Betting on bets: Anytime-valid tests for stochastic dominance," arXiv preprint arXiv:2604.21851, 2026.
- [2] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer, "Game-theoretic statistics and safe anytime-valid inference," *Statistical Science*, vol. 38, no. 4, pp. 576–601, 2023.
- [3] A. Müller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*. Hoboken, NJ, USA: Wiley, 2002.

Please contact:

Yo Joong Choe, INSEAD, Singapore
yojoong.choe@insead.edu

Sebastian Arnold, CWI, the Netherlands
sebastian.arnold@cwi.nl

A Map of Sub-Gaussian E-variables

by Wouter M. Koolen (CWI & University of Twente),
Shubhada Agrawal (IISc Bangalore) and Martin Larsson
(Carnegie Mellon University)

Could our data be sub-Gaussian noise? We explore rejecting that null hypothesis with the help of e-variables. We map the landscape of optimal e-variables against two-point alternatives.

In hypothesis testing, our goal is to reject a null hypothesis by finding evidence that contradicts it. In the framework of e-values, this evidence is measured by an e-variable – a non-negative random variable whose expectation under the null is at most one. To maximise our testing power, we seek the e-variable with the largest expected logarithmic growth under a chosen alternative distribution.

While the optimal e-variable against a simple point-null is just the likelihood ratio, composite null hypotheses are far more challenging. For convex sets of distributions, we understand the growth-optimal e-variables for Bernoulli trials, for distributions defined by finitely many moment constraints, and even the one-sided sub-phi class [1,2]. However, the case of sub-Gaussian distributions has remained a glaring omission.

Sub-Gaussianity is the workhorse of modern noise modelling. It defines a non-parametric, convex set of distributions – including the standard Gaussian and the discrete Rademacher distribution – characterised by a continuous family of moment constraints. This raises a fundamental question: How do we refute sub-Gaussianity using e-values?

One might expect to straightforwardly extend the solution for finitely many moment constraints. But one cannot. The reason lies in a mathematical "Easter egg" found in the definition of sub-Gaussianity itself. Indeed, any sub-Gaussian distribution has mean zero and variance bounded by one. Hence the outcome squared is an e-variable. Yet crucially, that square cannot be expressed as a mixture of the exponentials that define sub-Gaussianity. It emerges only through limits.

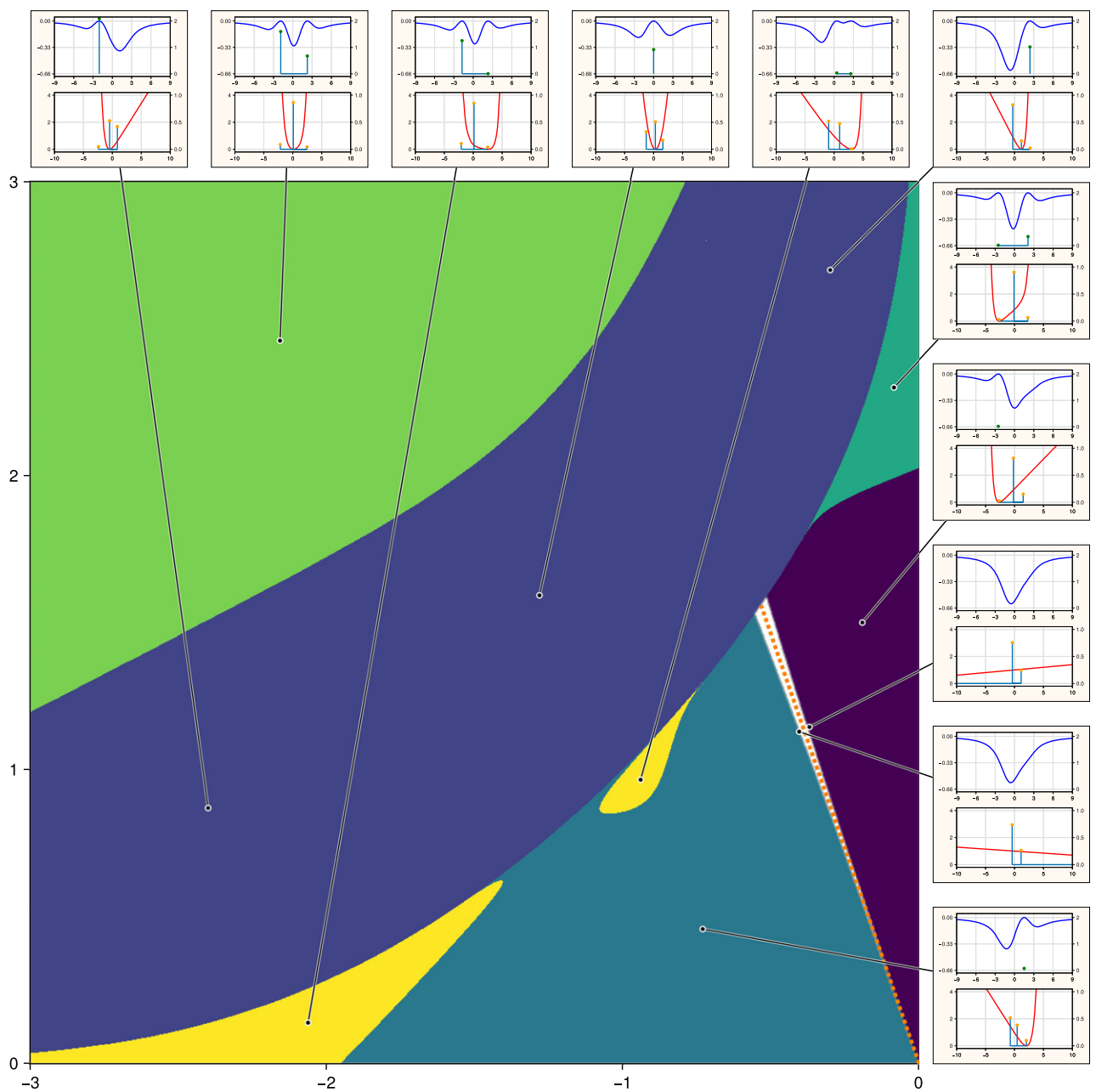
We recently characterised admissible sub-Gaussianity e-variables, as shown in the Figure. These variables are composed of a signed bet against the mean and positive bets against the sub-Gaussian constraints, reorganised to introduce a removable singularity (specifically, the limit as the constraint parameter approaches zero now results in the square).

Here we take this characterisation for a spin by numerically exploring the log-optimal e-variables for arbitrary two-point alternatives. Two-point distributions have three parameters (two points and one weight), and while we did all computations in 3d we can only present a 2d slice here. In the Figure we map the structure of these problems across the plane. Each coordinate represents a two-point alternative distribution, with the left and right support points on the axes and a fixed weight of 0.745 on the left point (a value chosen to reveal the most intricate structure). At each such alternative we computed the

KKT saddle point, obtaining the primal e-variable and its dual, the RPr (the reverse information projection, i.e., the sub-Gaussian distribution closest in KL-divergence to the alternative). Even though we are optimising over functions and measures, the major surprise here is that both primal and dual solutions have tractable short descriptions. The RPr is supported on exactly three points: the two points of the alternative and a third optimised point where the optimal e-variable vanishes. The optimal e-variable puts weight only on the one or two active sub-Gaussianity constraints for the RPr.

The Figure reveals the underlying geometry of the optimal e-variable, which bets against two constraints (yellow, green and

teal), or against one (blue, turquoise and purple). Within each, the colours flag whether the third support point is located before, between or beyond the two support points of the alternative. For a selection of 11 alternative distributions, the log-optimal e-variable is graphed in red, together with the RPr. In blue we graph the sub-Gaussianity constraints, together with the one- or two-point positive measure, which is supported precisely on the one or two points where the constraints are tight. In the white area our method runs into numerical trouble, as the third support point walks off to infinity and the optimal e-variable becomes the flat 1. The fourth selected alternative in the top row puts all mass on the sub-Gaussianity constraint at parameter value zero, where the optimal e-variable is indeed



A probability distribution P on \mathbb{R} is **sub-Gaussian** if $\forall \theta \in \mathbb{R} : \mathbb{E}_{X \sim P} e^{\theta X} \leq e^{\frac{1}{2}\theta^2}$
 $S : \mathbb{R} \rightarrow [0, \infty)$ is an **e-variable** if $\mathbb{E}_{X \sim P}[S(X)] \leq 1$ for every sub-Gaussian P
Characterisation Admissible e-variables are of the form $1 + bX + \mathbb{E}_{\theta \sim w} \left[\frac{e^{\theta X - \frac{1}{2}\theta^2} - 1 - \theta X}{\frac{1}{2}\theta^2} \right]$ for some $b \in \mathbb{R}$ and positive measure w on \mathbb{R} .

Figure 1: The landscape of optimal bets against sub-Gaussianity using two-point alternatives.:

a parabola (betting against the outcome and the outcome squared). Due to our parametrisation this smoothly fits in the general pattern.

The upshot?

Beyond providing a rigorous sanity check for our characterisation, this work yields a parametric family of optimal e-variables defined by at most five parameters. This makes it possible to employ adaptive sequential learning methods when data arrives from any unknown two-point alternative. Our results pave the way for a full understanding of the landscape beyond two-point alternatives, and for practical efficient anytime-valid refutation of sub-Gaussianity.

Each point in the plane represents a two-point alternative distribution, defined by its left (x-axis) and right (y-axis) support points, with a fixed weight of 0.745 on the former. The distributions on the dotted orange line segment are sub-Gaussian, the rest is not. The colours categorise the structural regimes of the log-optimal e-variable and its dual, the RIPr. In the yellow, green, and teal regions, the e-variable bets against two constraints; in the blue, turquoise, and purple regions, it bets against only one. The insets show (bottom graphs) the e-variable in red together with the RIPr, which is supported on the two points of the alternative and a third special point where the e-variable vanishes. The insets further show (top graphs) the sub-Gaussianity constraints of the RIPr in blue, together with the mixture of bets employed by the optimal e-variable against the tight constraints, of which there can be either one or two (in some cases these fall outside the range of the x-axis). In the white area, the third support point escapes to infinity, while its mass tends to zero, causing the e-variable to flatten to 1 at the sub-Gaussian submodel (dotted orange).

Martin Larsson acknowledges support from NSF grant DMS-2510965.

References:

- [1] P. Grünwald, R. de Heide, and W. Koolen, “Safe testing,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 86, no. 5, pp. 1091–1128, Nov. 2024.
- [2] M. Larsson, A. Ramdas, and J. Ruf, “Testing hypotheses generated by constraints,” *Mathematics of Operations Research*, vol. 0, no. 0, 2026.

Please contact:

Wouter Koolen
CWI and University of Twente, The Netherlands
wmkoolen@cwi.nl

E-Values for Contingency Tables

by Thorsten Dickhaus (University of Bremen), Francesca Giuffrida (Leiden University and IMT School for Advanced Studies Lucca) and Yongqi Wang (CWI)

We present powerful and easy-to-compute e-values for the classical statistical task of testing associations between two binary traits based on contingency table data. Genetic case-control association studies are our main intended use case.

Testing for association between two categorical variables on the basis of contingency table data is a classical task in inferential statistics. One prominent example is the “Lady tasting tea” experiment described by Sir R. A. Fisher in 1935. Analysing many contingency tables simultaneously and/or sequentially is important in the context of genetic association studies when analysing associations between categorical genetic markers and a categorical (often binary) disease status; see, e.g., [2].

The research question we address [1] is how to design an e-value for a contingency table, that is both easy to compute and powerful. In this, we mean by “easy to compute” that resource-intensive operations like a loop over all contingency tables with given marginal counts shall be avoided. This requirement refers to the situation that hundreds of thousands of such e-values have to be computed for one and the same dataset in the case of a genome-wide association study (GWAS). By “powerful”, we mean that the e-value should (with high probability) be larger than the “baseline” e-value obtained by calibrating the p-value based on Fisher’s exact test to the e-value scale using a standard p-to-e-calibrator. Since the null hypothesis of no association is composite, the Bayes factors proposed in [3] are prone to lack the e-value property.

Our primarily intended use case is a GWAS which is either multi-centric (independent patient groups are recruited at different locations) or group-sequential (independent patient groups are recruited at the same location at different time points). These two sampling schemes are realistic for GWAS, and e-values (if easy to compute and powerful) and their corresponding e-processes can allow data analysts to combine the evidence across centres or across time points, respectively, in a convenient and flexible manner (e.g., by multiplication or by averaging of e-values). In particular, e-processes allow for safe anytime-valid inference, implying the possibility of optional stopping.

For the special case of a 2×2 contingency table, we have investigated several possibilities to define such e-values, and we were able to characterize sampling schemes under which the usage of each of these e-values is particularly appropriate: while the e-process that was recently introduced by Turner et al. [4] has theoretical growth-optimality (under the alternative hypothesis of association) properties for paired data sequences, we have found that, when given a single large contingency table (the “batch setting”), it does not perform well - and another standard e-process, based on the principle of universal inference, performs even worse. In contrast, conditional types

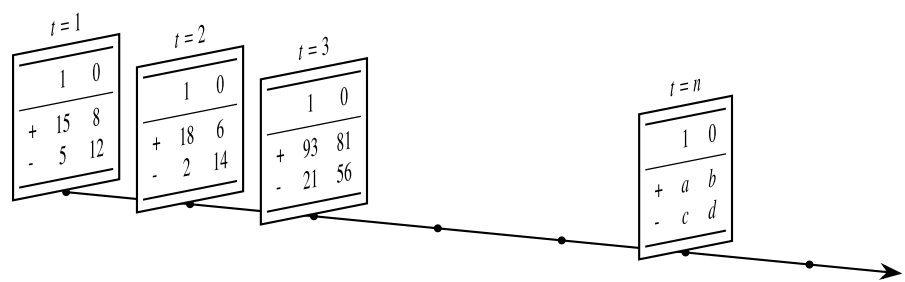


Figure 1: Illustration of the sequential 2×2 tables collected over time. At each time point t , binary observations from two groups are observed, coded as “+” or “-”. In the table at time $t=n$, entries are denoted by $\{a, b, c, d\}$, while the first three tables contain example numerical counts.

of e-variables tend to perform better in the batch setting, and among these, uniformly-most-powerful (conditional) e-variables perform best. When batches (i.e., tables) arrive sequentially, as illustrated in Figure 1, the picture becomes more complicated: Turner et al.’s e-variable is generally optimal asymptotically, and a conditional type of e-variable is often, but not always, preferable at smaller sample sizes.

This theoretical finding is confirmed empirically by means of computer simulations (as illustrated by Figure 2) and by re-analysing real genetic association datasets. Furthermore, we have revisited a meta-analysis replicating published psychological findings with peer-reviewed experimental protocols.

While our findings allow for analysing associations between a binary disease status and a binary genetic marker (e.g., a risk allele), future research will extend these investigations to categorical genetic markers with more than two categories. In particular, bi-allelic single nucleotide polymorphisms (SNPs) are often considered in GWAS. Such SNPs exhibit three categories.

References:

- [1] S. Arnold, et al., “E-Values for contingency tables, Revisited” [work in preparation; an initial presentation is given at the 2026 SAVI meeting at University of Twente], 2026.
- [2] T. Dickhaus, et al., “How to analyze many contingency tables simultaneously in genetic association studies,” *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 4, Art. no. 12, 2012. <https://doi.org/10.1515/1544-6115.1776>
- [3] T. Dickhaus, “Simultaneous Bayesian analysis of contingency tables in genetic association studies,” *Statistical Applications in Genetics and Molecular Biology*, vol. 14, no. 4, pp. 347–360, 2015. <https://doi.org/10.1515/sagmb-2014-0052>
- [4] R. J. Turner, et al., “Generic E-variables for exact sequential k -sample tests that allow for optional stopping,” *Journal of Statistical Planning and Inference*, vol. 230, Art. no. 106116, 2024. <https://doi.org/10.1016/j.jspi.2023.106116>

Please contact:

Thorsten Dickhaus
University of Bremen, Germany
dickhaus@uni-bremen.de

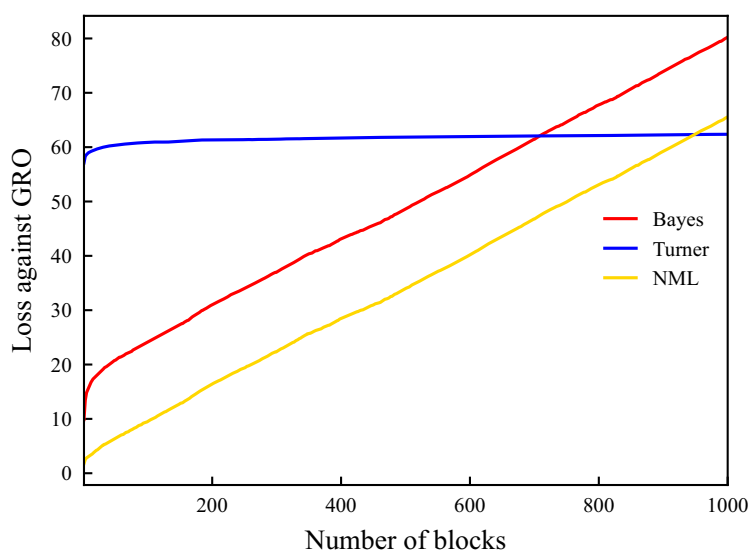


Figure 2: Growth-rate loss of several e-processes relative to the growth-rate-optimal method (GRO). This loss is a proxy for loss of statistical power compared to an “oracle” method that has knowledge of the underlying parameters. We simulated two independent Bernoulli streams up to a time horizon of 1000, with 1000 replications. At each time point, a 2×2 table was created under a fixed design with group sizes (200, 50). Within each group, the number of positive outcomes (+) was sampled according to the fixed corresponding Bernoulli means, (0.78, 0.01). Turner’s method is shown in blue, a conditional Bayesian method with an uninformative Gaussian prior in red, and a conditional normalized maximum likelihood approach (NML) in yellow. Reference [1] contains many more simulations, with small and large tables, balanced and unbalanced group sizes and small and large effect sizes, all confirming a similar general pattern.

E-values Instead of P-values in Clinical Trials: What Happens?

by Yongxi Long and Erik van Zwet (Leiden University Medical Center)

Anytime valid testing with e-values offers great flexibility allowing both optional stopping (peeking) and optional continuation (collecting more data). The price to pay is a reduction of statistical power. Using data from more than 20,000 randomized trials, we evaluate how e-values compare with classical p-values in balancing flexibility and efficiency.

Clinical trials are conventionally conducted under a fixed sample design, in which the sample size is pre-determined and the analysis is performed only after data collection has ended. In practice, however, investigators often want greater flexibility. They may wish to monitor the accumulating evidence, and stop if the treatment benefit is evident or recruitment proves difficult. They may also want to collect more data if the results are promising but not yet conclusive (optional continuation). E-values [1] offer a unified framework of anytime-valid inference to achieve such flexibility; they remain valid no matter how frequently the data is inspected or how the sampling plan changes. However, this flexibility comes at the price of reduced power. We conducted a large-scale empirical study to assess the trade-off.

What are e-values?

Statistical tests are designed to limit the probability of a false positive result at a predefined significance level α . Classical p-value procedures exhaust their “ α budget” at the final analysis. Therefore, they are incompatible with optional continuation and require rigorous α -spending rules [2] to enable optional stopping.

E-values provide a more flexible alternative. They are designed so that, when there is truly no treatment effect (the null hypothesis), they are small (not greater than 1) on average.

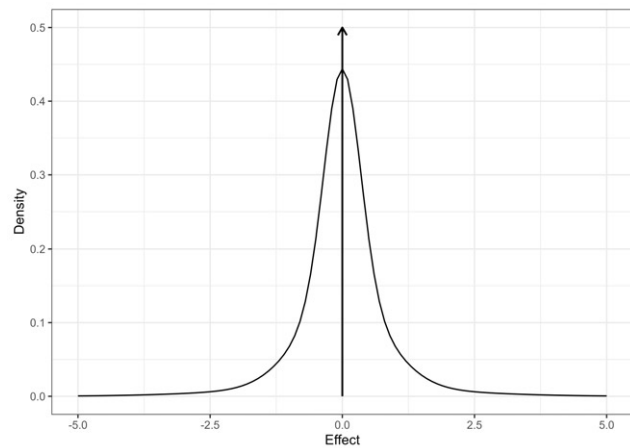


Figure 1: Distribution of effect sizes across 20,000+ clinical trials in CDSR. The spike at zero (44% proportion) reflects the reality that many trials investigated very small effects, and the long tail represents a few very successful interventions with strong effects (e.g., vaccine trials).

Large e-values therefore represent evidence against no effect. E-values can be used to build e-processes: these are sequences of e-values computed from accumulating trial data such that, under the null hypothesis, the probability that they will ever exceed a fixed threshold remains small. We can use an e-process to monitor the trial, and reject the null hypothesis whenever the e-process exceeds a certain threshold. As such, an e-process spends the α budget continuously. The user must decide if they want to spend α cautiously or more boldly.

Empirical evaluation

To answer the question of what happens if trialists had used e-values instead of p-values in clinical trials, we used a large collection of 20,000+ real-world trials from the Cochrane Database of Systematic Reviews (CDSR). We start by estimating the distribution of effect sizes. We find a spike around zero indicating many very small effects and a long tail of larger effects (Figure 1). We then generated synthetic trial trajectories up to and beyond their original sample sizes, which match the observed summary statistics (effect estimate and standard error) of the original trial at the original sample size. We can

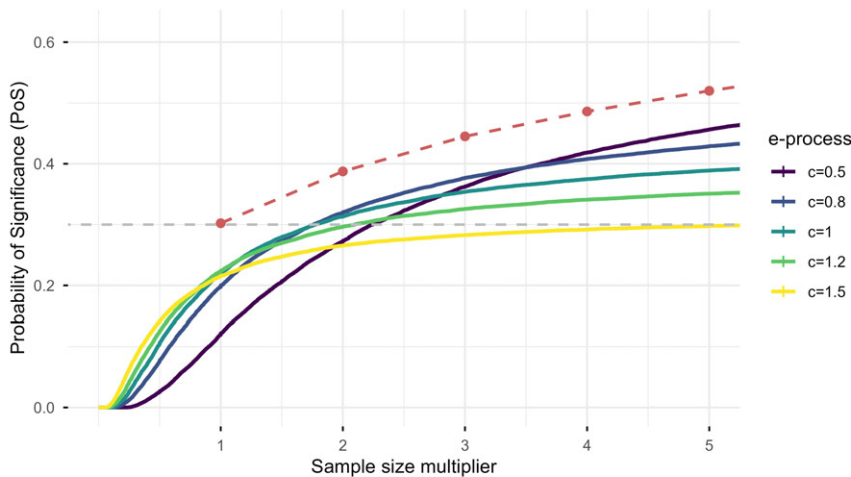


Figure 2: Probability of significant results (PoS) when trials are continued until 1,2,...,5 times their original sample sizes. The larger the sample size, the higher the chance a trial finds a statistically significant result. The coloured lines show the performance across a spectrum of e-processes, from aggressive ones with large values of c that seek early signals of large effects, to conservative ones with small values of c that hold conclusions for small effects until enough data have been collected. The red dashed line shows the PoS of the traditional p-value test as a reference. There are 30% of trials in the CDSR cohort that have reached statistical significance, which is plotted as a grey horizontal dotted line.

e-process	c=0.5	c=0.8	c=1	c=1.2	c=1.5
Sample size multiplier at 30% PoS	2.24	1.76	1.77	2.1	5.59
Mean sample size multiplier at 30% PoS	1.93	1.48	1.47	1.69	4.19

use this synthetic dataset to benchmark any given e-process against the fixed sample p-value. We focus on the probability of reaching statistical significance (probability of significance or PoS) and the time until rejection.

Various e-process constructions can be considered for the CDSR cohort which differ in the rate at which they spend their α budget. For every trial, we can infer the minimum clinically important difference (MCID) that was used for its sample size calculation. We construct e-processes as sequences of likelihood ratios comparing $c \times \text{MCID}$ to the null for several choices of c . Choosing a large value of c implies that we expect that the true effect is large and we should spend α quickly to detect it as soon as possible. Small values of c mean that we anticipate a small effect, and we should reserve enough α until we have collected a large sample.

Flexibility always costs power, but the cost can be reduced by designs

The PoS of fixed-sample p-value testing when every CDSR trial is continued until 1,2,...,5 times its original sample size is plotted in red dashed line as a reference in Figure 2. The PoS of likelihood ratio e-processes with different target effect sizes are plotted cumulatively against sample size multipliers. We find that targeting the e-process at the MCID (i.e., $c=1$) performs well across the CDSR cohort. It requires planning for approximately 1.8 times the original sample size to match the PoS of a fixed-sample p-value, but the expected sample size is reduced to approximately 1.5 times due to optional stopping: there are many cases when the e-process reaches significance long before the planned sample size.

Conclusion

The great flexibility of e-values offers attractive advantages, particularly in enabling anytime-valid testing. Their practical implementation in clinical trials requires a delicate balance between flexibility and power. If we spend our α -budget too fast, we make suboptimal use of the available sample size. On the other hand, if we are too cautious, we are effectively reserving α for finding small effects at sample sizes that are too large to be realistic given the practical and financial constraints of clinical trials. In a setting such as the CDSR which is characterized by many small effects and a few large ones, relatively aggressive e-process constructions yield good average efficiency.

References:

- [1] Grünwald, P., de Heide, R., & Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86 (5), 1091–1128. <https://doi.org/10.1093/jrssl/qkae011>
- [2] Cook, T. D. (2002). P-value adjustment in sequential clinical trials. *Biometrics*, 58 (4), 1005-1011.

Please contact:

Yongxi Long
Leiden University Medical Center, The Netherlands
y.long@lumc.nl

Erik van Zwet
Leiden University Medical Center, The Netherlands
E.W.van_Zwet@lumc.nl

Racing to the Truth: How E-values Can Speed Up Science

by Sebastian Arias , Alexander Ly, Michele Meziu (CWI) and Angel Reyero Lobo (CWI and Inria)

Modern science generates data continuously, but the statistical methods that still dominate many fields generally require data collection to end before reliable meta-analysis can begin. New research on e-values offers a way to analyse evidence in real time, without sacrificing statistical reliability. The approach could make science not just more robust to modern research practices, but significantly more efficient.

Science no longer proceeds one experiment at a time. Data now flow continuously from laboratories, countries and digital platforms, often through sprawling parallel collaborations. Yet statistical practice remains rooted in an era when evidence was assessed only after each study had been concluded.

Current ongoing research into e-values [1] suggests a more modern approach. Our preliminary findings indicate that e-values are not just theoretically appealing but also practically advantageous, allowing evidence to be monitored, combined and acted upon continuously during data collection – without sacrificing statistical reliability or sample efficiency (see Figure 1 for a schematic illustration).

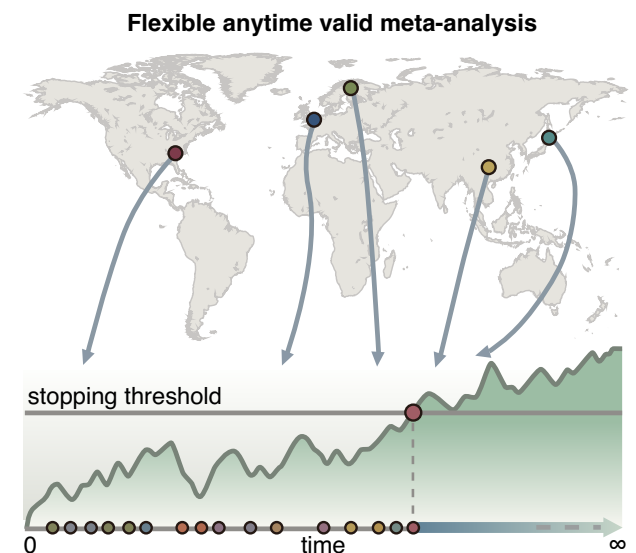


Figure 1: Sequential evidence monitoring across multiple data-collection sites. The map shows five geographically distinct data-collection sites, each represented by a different colour. The coloured dots along the x-axis mark observations as they arrive over time from the different sites. The curve shows the accumulation of meta-analytical evidence against either the hypothesis of no effect, or against the hypothesis that the effect is meaningfully large. The horizontal line marks the stopping threshold for the combined analysis, corresponding here to the evidential strength of four individually significant replication attempts.

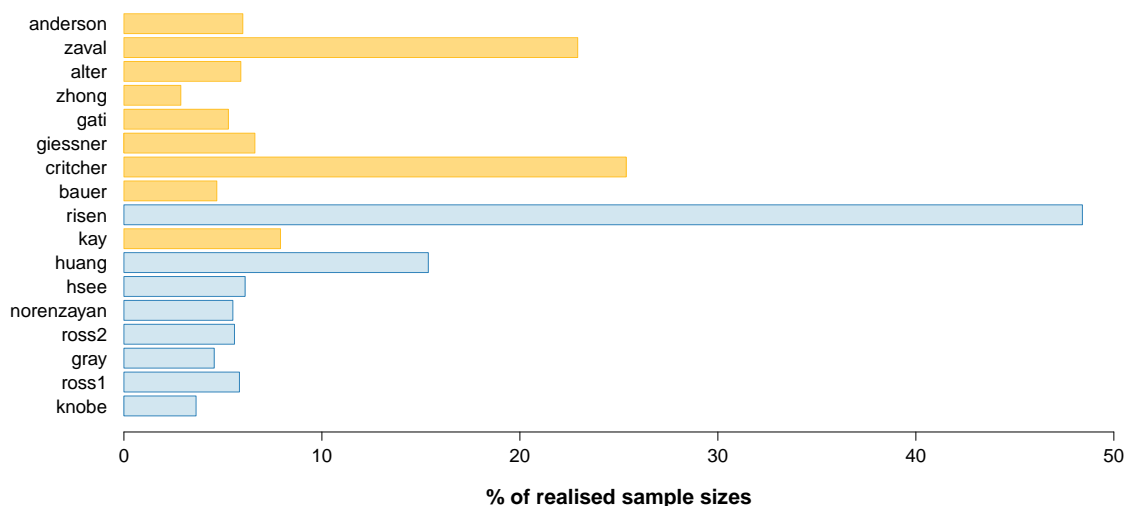


Figure 2: Percentage of the realised sample from the replication studies needed to reach a conclusion across seventeen different effects. The identifiers on the left refer to the (first) authors of the original published effects. Blue indicates evidence for replication success, whereas yellow indicates evidence for replication failure.

Projects such as the Many Labs 2 Project [2] illustrate the potential of e-value-based methods in multi-stream experimental settings. This vast replication effort involved 61 laboratories and more than 15,000 participants testing 28 published psychological effects. Under traditional p-value-based methods, researchers could not safely aggregate evidence until every participating laboratory had finished collecting data; interim analyses risked producing unreliable conclusions.

E-values could have changed this. Preliminary results suggest that, had they been used, robust conclusions about replicability could often have been reached using only a fraction of the realised sample sizes – and crucially, while the experiments were still running.

The method’s power [L1] lies in its structure. It races two e-values against each other: one accumulating evidence against the null hypothesis of no effect, and another accumulating evidence against the hypothesis that the effect is meaningfully large. Both remain statistically valid regardless of when – or even if – data collection has stopped. This allows interim evidence from multiple laboratories to be safely combined into a live meta-analysis, continuously weighing the case for and against replicability.

Figure 2 shows the percentage of the realised sample sizes needed to reach (strong) conclusions for 17 of the 28 effects. The meta-analytical evidential threshold used to determine success or failure of replicability across sites corresponds to the strength of four individually significant replication attempts combined. In several cases, the reductions are substantial. For one effect (“ross1” in Figure 2), reliable replication could already have been established after 421 participants – just 5.8% of the 7,205 eventually recruited. For another (“bauer”), evidence against replicability emerged after only 311 participants, or 4.7% of the final sample of 6,608. The gains were not universal: one effect (“risen”) still required nearly half of the realised sample size before a conclusion could be reached.

Interpreting these reductions precisely, however, requires some caution. Many Labs 2 was designed primarily to characterise variability in effect sizes across diverse samples and contexts, with replication success as a secondary consideration. Its design therefore prioritised breadth and robustness over statistical efficiency, suggesting that while our observed reductions should not be interpreted too literally, they nonetheless point to substantial untapped gains from continuous evidence aggregation.

The examples come from psychology, a field acutely aware of the costs of inefficient and irreproducible research. But the implications extend much further. Medicine, genomics and online experimentation increasingly depend on evidence accumulated simultaneously across many sites. Statistical tools capable of aggregating such evidence in real time may therefore become central to the organisation of science itself.

Our research is ongoing. But the early results already suggest that the e-value-based methods designed to make scientific inference robust to the adaptive realities of modern data collection can also make science substantially more efficient.

Link:

[L1] <https://github.com/AlexanderLyNL/safestats/tree/futility88>

References:

- [1] A. Ly, et al., “Dynamic evidence synthesis with e-values: Efficient sequential meta-analysis with early stopping for efficacy or futility with anytime-valid type I and II error control,” work in progress, 2026.
- [2] R. A. Klein, et al., “Many Labs 2: Investigating variation in replicability across samples and settings,” *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 4, pp. 443–490, 2018.

Please contact:

Alexander Ly, CWI, The Netherlands
a.ly@cwi.nl

How Small is Negligible? Rethinking Equivalence Testing with E-values

by Stan Koobs and Nick W. Koning (Erasmus University Rotterdam)

A classical statistical problem is to assess whether an unknown quantity is negligible: that is, practically equivalent to zero. It is standard to define ‘negligible’ as being smaller in magnitude than some threshold margin. Specifying this margin has plagued statisticians for decades: if it is set too large, then one can hardly speak of negligibility, but if the margin is set too small, then one may need an enormous amount of data to statistically establish negligibility. In recent work, we study this problem in depth and show how e-values can be used to bypass it, by enabling one to select the margin post-hoc: after seeing the data.

Many important decisions hinge on showing that some unknown quantity is negligible: a cheaper medical treatment is approved when its effect is close enough to the established treatment; the existence of a psychological effect may be called into question if a replication study shows the effect is negligible; a product feature is removed if an A/B test shows that it hardly improves customer engagement.

Establishing negligibility is traditionally framed as a hypothesis testing problem, commonly known as equivalence testing or non-inferiority testing. In this framework, the statistician first selects an equivalence margin: a threshold below which the quantity of interest is considered negligible. They then test the hypothesis that the quantity exceeds this margin. If this hypothesis is rejected, the quantity is declared to be negligible. This testing framework has become the standard tool for establishing negligibility, and is embedded in FDA, EMA, and ICH guidelines.

Unfortunately, the use of an equivalence test to establish negligibility faces a longstanding problem: such tests require the statistician to specify a margin, yet in many applications there simply is no natural margin below which a quantity is practically negligible and above which it is substantive. A common “solution” is to defer the choice of margin to an expert or to standard convention. But while this gives the statistician a usable margin, it does not solve the underlying problem: if there truly is no natural margin, then no expert or convention can make one appear.

If we do proceed with selecting a margin in an application where no natural margin exists, we are faced with a difficult trade-off: set it too large and the verdict of “negligible” becomes weak; set it too small and even a genuinely negligible quantity may demand an impractical amount of data to establish this. It may then be tempting to choose the margin post-hoc, based on the data, but this renders the statistical test invalid.

All of this leads to a natural question: should establishing negligibility really be treated as testing a single hypothesis?

This question is at the heart of our recent paper [1]. To answer it, we take a step back to consider the role of statistics within an application. We argue that its primary role is to provide statistical guarantees on the consequences of decisions, which can subsequently be used to inform decision makers. Reasoning backwards from these guarantees, we can determine the appropriate statistical methodology.

The underlying desire to show that the unknown quantity is negligible can be interpreted to mean that larger values of this quantity correspond to worse consequences. Under this natural assumption, we find that the appropriate methodology is not to report a single test outcome but a test outcome for every margin. The margins corresponding to hypotheses that are not rejected jointly form a confidence set for the unknown quantity, as shown in Figure 1. Only if the consequences hinge on a single margin, then a single equivalence test for this natural margin suffices.

We take this one step further by generalizing beyond hypothesis tests to e-values, reporting an e-value against every margin simultaneously. The result is what we call an equivalence curve, of which an example is shown as the solid line in Figure 1. It provides a continuous record of how much evidence the data carries against each margin. The larger the margin, the easier it is to establish that the unknown quantity is below it, so that the curve rises with the margin. For comparison, the dashed line shows the special equivalence curve that is built out of tests. This curve is very rigid, as tests are binary e-values that emit either no evidence (no rejection) or some large amount of evidence (rejection).

A remarkable feature of the equivalence curve is that it is simultaneously valid for every margin, allowing one to browse the equivalence curve and select a margin post-hoc while maintaining statistical validity. This means we can truly by-

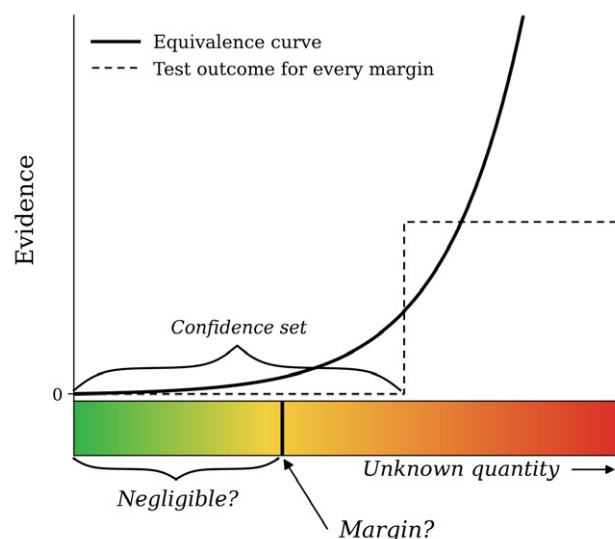


Figure 1: An equivalence curve (solid) reports continuous evidence (e-value) at every margin. A classical test for every margin produces a special equivalence curve (dashed) that flips at some data-dependent margin. The region where this equivalence curve is zero forms a confidence set. The color gradient describes the perceived negligibility of the unknown quantity, and the placement of the solid vertical line describes the problem of specifying a single margin.

pass the problem of pre-selecting a margin. This builds on recent work about hypothesis testing with a data-dependent significance level by Grünwald [2] and Koning [3], by implicitly coupling the margin to the significance level.

Taking a step back, a more general message that underlies our work is that not every statistical question must be forced into testing a single hypothesis. Establishing negligibility is a clean example: the real difficulty was never how to choose the right equivalence margin, but whether the problem should be reduced to testing a single margin in the first place.

References:

- [1] S. Koobs, N. W. Koning, “Equivalence testing with data-dependent and post-hoc equivalence margins”, arXiv:2603.16213, 2026.
- [2] P.D. Grünwald, “Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha”, PNAS 121(39):e2302098121, 2024, <https://doi.org/10.1073/pnas.2302098121>
- [3] N. W. Koning, “Post-hoc α hypothesis testing and the post-hoc p-value”, arXiv:2312.08040, 2025.

Please contact:

Stan Koobs
Erasmus University Rotterdam, The Netherlands
koobs@ese.eur.nl

Nick W. Koning
Erasmus University Rotterdam, The Netherlands
n.w.koning@ese.eur.nl

Safe Preference-Based Political Polls

by Adrienne Tuynman and Timothée Mathieu (Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 – CRISTAL)

Political polls before elections are useful to identify promising candidates, and to allow parties to make compromises or build alliances. We are interested in conducting polls sequentially, so that one can stop acquiring data as soon as possible while safely yielding statistically significant results.

While many countries use uninominal voting for their elections, where voters only choose the name of their preferred candidate, alternative comparison-based systems are seeing more and more usage, especially in smaller organisations or on the local scale. In those systems, voters give a ranking of candidates, so that their opinions are better taken into account. In this setting, how can one efficiently conduct an opinion poll? We propose to use sequential e-values for this purpose.

Sequential Hypotheses Testing for Political Surveys

We are looking into the Borda system: with N candidates, each voter awards θ points to their least liked candidate; the one slightly better gets one point; all the way up to the preferred candidate, who gets $N-1$ points. After tallying up the total

number of points that each candidate gets, the one with the most points wins. A candidate that is well liked by a majority of the population could therefore beat a candidate that is more polarising.

Formally, we assume the existence of a preference matrix, noting for each pair of candidates which one was preferred. The preference of i over j is represented by a Bernoulli random variable, of parameter equal to the probability that i is preferred to j . The theoretical Borda score can be computed through the matrix of these parameters, where the theoretical score of i is the sum over j of the probability that i beats j . At each iteration, we collect one sample from each of those Bernoulli random variables, thus yielding an observation of size $N(N-1)/2$.

We want to test which candidate wins an election determined by this unknown preference matrix. To be able to stop our survey as soon as possible, we use sequential e-values, allowing for optional stopping of the testing process.

E-values and reverse information projection

Formally, this means we test simultaneously for each candidate i the hypothesis $H_{0,i}$: “ i is a Borda winner” against $H_{1,i}$: “ i is not a Borda winner”.

A classical construction of e-values is based on the likelihood ratio. From [1], at the numerator, we use an estimate of the alternative distribution using the posterior distribution. On the denominator, we compute a projection p_0 of this posterior distribution onto H_0 via Reverse information projection, a projection that minimises the Kullback-Leibler divergence.

Formally, this yields

$$E_t := E_{t-1} \frac{\int P_\theta(X_t) dP(\theta | X_1, \dots, X_{t-1})}{p_0(X_t)}$$

For simplicity, the prior is a product of Jeffrey priors, which makes the posterior distribution easy to compute. Due to the multiplicative form of this e-value, we have the optional stopping property: we can stop collecting data whenever we want for whatever reason. The test which rejects H_0 if the e-value is larger than $1/\alpha$ has type I error below α . This describes how to test if a given candidate is a Borda Winner; to find the Borda Winner we use parallel tests as in [2], Section 3.2, which controls the probability of selecting a non-winner.

Application to sequential testing for Borda Winner in 2022 French Presidential Elections

We apply this method to a dataset of opinions about the 2022 French Presidential Elections from [L1]. This dataset provides pairwise votes coming from 2287 French voters, collected before and after the first round of the presidential elections. The responders also provided the ballot they actually cast in the official elections, and we use this actual ballot to reweight the sampling in order to match the official electoral results, in the same way that the original survey did.

At each timestep, for each pair of candidates, we sample a duel between those two. For each candidate, we try to reject the hypothesis that they are a Borda winner, with type I error 5%.

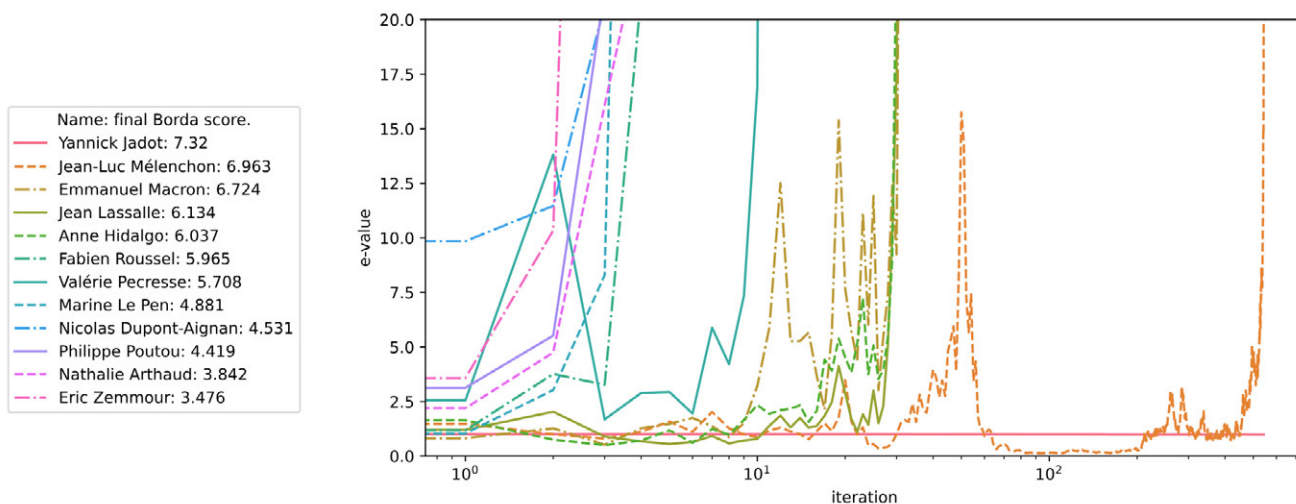


Figure 1: Evolution of the sequential e-values for the candidates in the 2022 French presidential election dataset as the number of sampled pairwise comparisons increases. Candidates are progressively eliminated until only one, Jadot, remains.

Based on that, we plot in Figure 1 the e-value corresponding to each candidate, depending on the number of iterations. When two candidates i and j are eliminated, it is no longer necessary to sample the preference i versus j , thus reducing the number of samples per iteration. In the end, only one candidate, Yannick Jadot, remains; the dataset confirms that he is the theoretical Borda winner. Interestingly, his e-value remains equal to 1 because the empirical distribution always lies within the hypothesis set corresponding to Jadot being a Borda winner. Consequently, no evidence accumulates against that hypothesis.

In Figure 2, we plot the number of samples required to retain only one potential winner, comparing our test (denoted TG) to two baselines: Universal Inference (see [3] Section 7), and the GLRT test (from [2] Section 3.2). We stop after collecting 2×10^5 samples, and - if the test hasn't yet yielded a single winner - we give the number of candidates still not rejected by the end of sample collection. Note that, while the number of samples seems to increase with the number of candidates, it is not always that straightforward: adding a very dominant candidate to a pool of equivalent candidates will make the test easier. Still, in those simulations, the e-value based test manages to conclude quicker than our two baselines, especially so as the instance gets more costly. For reference, official political surveys (IPSOS) conducted prior to the 2022 election used a total of 7,321 voters, in contrast TG uses 171 voters for a sample size of 4721 pairwise comparisons.

Our proposed e-values give a flexible and efficient way to do a political poll, allowing for optional stopping and using fewer samples than what is typically used in such surveys while having strong statistical guarantees.

Link:

[L1] <https://doi.org/10.5281/zenodo.10998451>

References:

- [1] R. J. Turner and P. D. Grünwald, "Exact anytime-valid confidence intervals for contingency tables and beyond," *Statistics & Probability Letters*, vol. 198, p. 109835, 2023.
- [2] E. Kaufmann, *Contributions to the Optimal Solution of Several Bandit Problems*. HDR dissertation, Université de Lille, 2020.
- [3] L. Wasserman, A. Ramdas, and S. Balakrishnan, "Universal inference," *Proceedings of the National Academy of Sciences*, vol. 117, no. 29, pp. 16880–16890, 2020.

Please contact:

Timothée Mathieu
 Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 – CRISAL, France

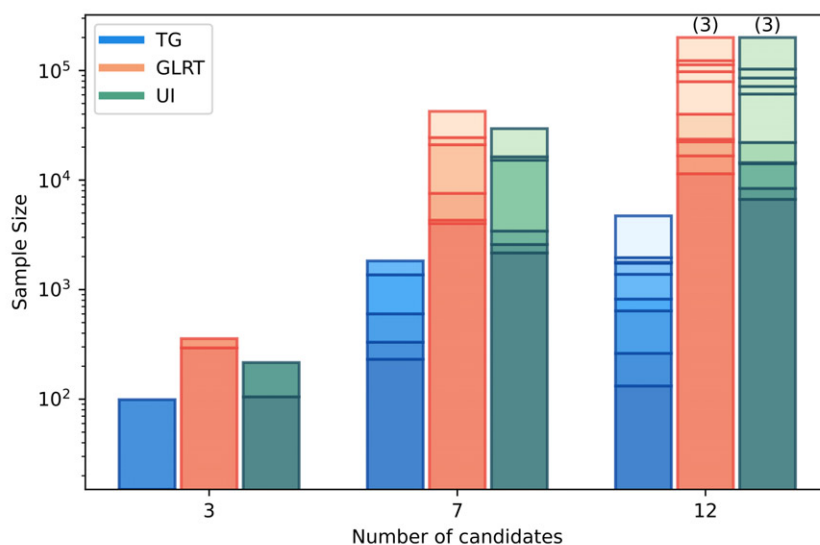


Figure 2: Comparison of the sample complexity required to identify a single winner using three sequential testing methods; the generalized likelihood ratio test (GLRT), and two e-value-based tests: our proposition (TG) and universal inference (UI), for different numbers of candidates.

Assessing Generative AI Systems Using E-scores

by Guneet Singh Dhillon (University of Oxford), Teodora Pandeva (Microsoft Research), and Alicia Curth (Microsoft Research)

Generative AI systems are becoming ubiquitous, but their outputs can still be inaccurate or misleading. Using e-values, the e-scores framework provides a statistically rigorous assessment of AI-generated responses while accommodating the adaptive and post-hoc nature of human-AI interactions.

Generative AI is actively shaping our everyday lives. Large language models now draft emails, summarise documents, generate software code, and answer millions of conversational queries each day. Their outputs are often coherent and persuasive, yet they may still contain subtle inaccuracies, fabricated facts, or logical inconsistencies. Such spurious outputs are especially problematic in high-stakes settings like healthcare, education, finance, and scientific research. This creates a central challenge for trustworthy AI: how can we reliably assess whether an AI-generated response is actually correct?

Statistical methods based on p-values have long offered a principled way to control errors in scientific inference. P-value-

based methods can also provide guarantees in our setting by filtering generated responses [1]. For example, suppose a user is willing to tolerate a 5% error rate. After filtering responses by their corresponding p-values, the probability of mistakenly retaining an incorrect response is at most 5%. In this way, the retained responses are reliable for the chosen tolerance level. A user asking for restaurant recommendations may accept a relatively high tolerance for error, while another seeking medical or legal advice may require a tolerance closer to zero.

These guarantees, however, only hold if the user fixes the tolerance level before examining the data. Modern-day usage of generative AI systems rarely follows such a rigid protocol because adaptivity is intrinsic to human-AI interaction. Users naturally revise their tolerance after inspecting the data. Figure 1 illustrates a setting in which the filtered responses remain unchanged for any tolerance level between 0.01 and 1. Since a lower tolerance corresponds to a stronger reliability guarantee, the user would want to choose 0.01 after observing the data. From a statistical perspective, this creates a serious problem that closely resembles the well-known issue of p-hacking in scientific research, and invalidates the original error-control guarantees.

Recent years have seen a surge of interest in hypothesis testing with e-values, which extend classical p-value methodology to more complex testing scenarios. Importantly, this includes settings with data-dependent or post-hoc tolerance levels. Unlike p-values, e-values retain post-hoc validity, i.e., they continue

PROMPT

Seth is twice as old as Brooke. In 2 years, the sum of their ages will be 28. How old is Seth?

RESPONSE

Let's denote Seth's current age as S and Brooke's current age as B . From the problem, we know that Seth is twice as old as Brooke, so we can write this relationship as an equation: $S = 2B$.

✓ e-score: 0.01

In two years, Seth will be $S + 2$ years old and Brooke will be $B + 2$ years old. The sum of their ages in two years will be 28, so we can write another equation: $(S + 2) + (B + 2) = 28$.

✓ e-score: 0.01

Substituting the first equation into the second equation, we get:
 $2B + 4 + B + 2 = 28 \iff 3B + 6 = 28 \iff 3B = 22 \iff B = 22/3$.

✗ e-score: 4.95

Since B represents Brooke's age, we need to find Seth's age. From the first equation, we substitute B with $22/3$:
 $S = 2 \times 22/3 \iff S = 44/3$.

✗ e-score: 6.01

So, Seth is $44/3$ or approximately 14.67 years old. The answer is: 14.67.

✗ e-score: 6.28

Figure 1: E-scores example for mathematical factuality. The large language model's response consists of five sub-responses, each a step in the mathematical reasoning (starting from the inner and ending on the outer block). The checks/crosses on the bottom left and the green/red colour of each block represent the response's (in)correctness up to that point. We highlighted part of the third sub-response that, on manual inspection, caused the incorrectness, which cascades to subsequent sub-responses. The e-scores on the bottom right of each block are measures of incorrectness, i.e., low for correct and high for incorrect responses. (Credit: [3].)

to provide guarantees even when the user chooses the tolerance level after examining the data [2]. This property makes them particularly well-suited to the interactive and adaptive use of modern generative AI systems.

In our recent work, together with Javier González, we developed a framework based on e-values for assessing the correctness of AI-generated responses [3]. At the heart of the approach is the idea of assigning each generated response a non-negative score that measures evidence of incorrectness, i.e., low scores indicate reliable responses, while high scores indicate incorrect ones. Embracing the adaptive nature of human-AI interaction, the framework allows users to choose their tolerance level (and therefore the filtering threshold) after observing the responses and their scores. Crucially, because these scores are derived from e-values, they provide guarantees for post-hoc validity. We therefore call these scores e-scores.

Another appealing feature of the e-scores approach is its ability to assess smaller portions of a generated response. For instance, a long AI-generated text may contain both highly reliable and spurious segments. As illustrated in Figure 1, the framework can assign e-scores not only to the complete response but also to partial responses, allowing users to identify which portions deserve greater trust and which may require additional verification. This provides much finer-grained information than a single global assessment while maintaining statistical validity. The framework is also applicable beyond large language models and textual outputs. In principle, it can broadly apply to any generative AI system and any output domain. This flexibility opens the door to diverse applications and use cases. In our work, we also demonstrate the framework's efficacy in two real-world scenarios [3]. The first focuses on mathematical factuality to identify the first incorrect step in a chain of reasoning steps; Figure 1 presents one such example. The second evaluates the desirability of a response according to criteria such as helpfulness and truthfulness. In both cases, the empirical results corroborate that the theoretical guarantees translate effectively into practical utility.

Altogether, the e-scores framework offers a new mechanism for assessing generative AI systems. It incorporates e-values to bridge rigorous guarantees and adaptive human-AI interactions. In doing so, it empowers users to make reliable decisions while providing statistical post-hoc validity. We believe this presents an important step toward trustworthy AI and the adoption of AI systems in our everyday lives.

References:

- [1] C. Mohri and T. Hashimoto, "Language models with conformal factuality guarantees," in Proceedings of the International Conference on Machine Learning, 2024.
- [2] P. D. Grünwald, "Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha," Proceedings of the National Academy of Sciences, 2024.
- [3] G. S. Dhillon, J. González, T. Pandeva, and A. Curth, "E-scores for (in)correctness assessment of generative model outputs," in Proceedings of the International Conference on Artificial Intelligence and Statistics, 2026.

Please contact:

Guneet Singh Dhillon, University of Oxford, United Kingdom
guneet.dhillon@stats.ox.ac.uk

Winning by Watching: Evaluating Policies Safely for Reinforcement Learning

by Stephan Bongers (CWI)

Standard statistical guarantees fail when analysts repeatedly check incoming data. By integrating anytime-valid inference with reinforcement learning, this work enables safe policy evaluation under continuous monitoring.

Making safe decisions with predictable outcomes in an unknown environment is a longstanding goal in science. For example, in healthcare we want to predict treatment outcomes before giving the treatment, or in economics we seek interventions with reliable effects despite uncertainty. Although many off-the-shelf methods exist that test whether a policy has an outcome with a certain statistical guarantee provided that the sample size is fixed in advance, they generally break down when the decision maker is allowed to start and stop data collection at any point in time. By combining reinforcement learning and anytime-valid methods, we aim to develop such statistical methods that allow for safe decision making at any point in time.

Reinforcement learning (RL) studies how a decision maker can learn a policy that maximizes long-term cumulative reward through interaction with an environment (see Figure 1(a)). When the environment is known, as in games like chess or Go, policies can be evaluated reliably by simulating many games. But in domains such as healthcare or economics, where the environment is not known beforehand, reliably evaluating new policies before deployment becomes much harder.

Off-policy evaluation (OPE) aims to estimate the value of a policy of interest, known as the target policy, using data collected under a different policy, referred to as the behavior policy. A major challenge in OPE is the curse of horizon: as the decision horizon increases, the overlap between the target and behavior policies can decay exponentially fast. This diminishing overlap severely undermines the reliability of value estimation. Consequently, many practical applications require not only accurate point estimates but also principled measures of uncertainty, such as confidence intervals (CIs). In response to this need, a growing body of work has developed OPE methods for constructing statistically valid CIs [2].

Consider the following example. Suppose you have access to the gameplay records of a friend who plays the Gridworld game every day (see Figure 1(b)). Can you determine, with certainty, whether you could play the game better than your friend? One possible approach is to compute confidence intervals (CIs) with some existing OPE method for both your friend's strategy and your own hypothetical strategy using the n games observed so far. Although your strategy may appear more promising, the resulting CIs may still be inconclusive because they overlap (see Figure 1(c)). At this point, it may seem natural to collect additional gameplay data and check again whether the comparison becomes conclusive. However, doing so invalidates the CIs, since the sample size would then depend on the ob-

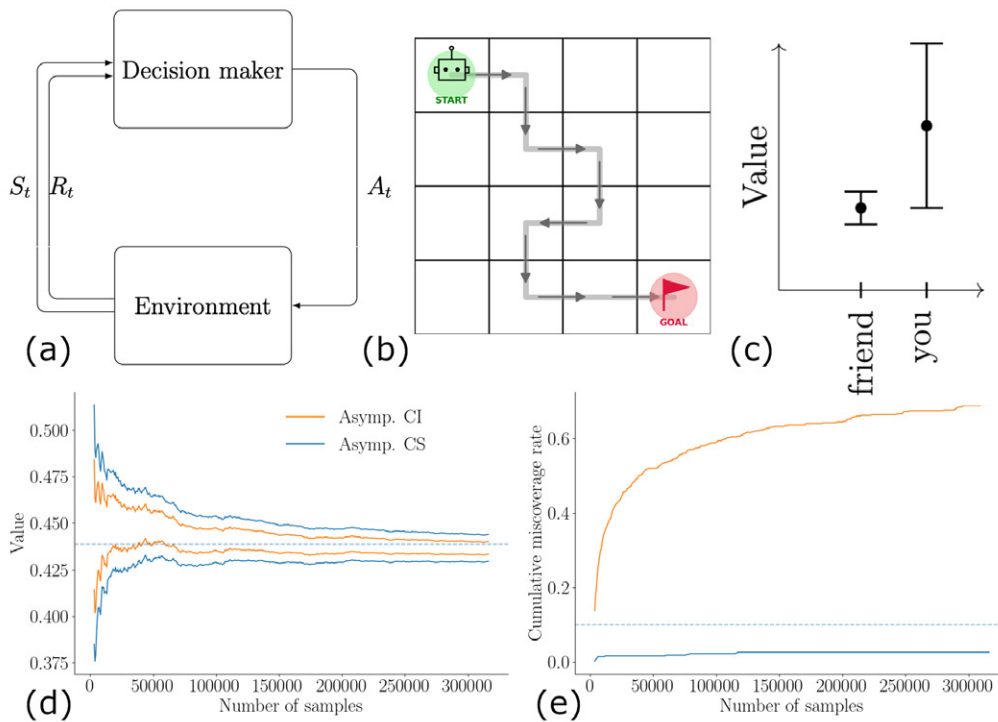


Figure 1: (a) Reinforcement learning interaction diagram, (b) Gridworld game illustration, (c) confidence interval overlap for a point estimate, (d) asymptotic confidence interval and confidence sequence for the value of the target policy based on the number of samples, (e) cumulative miscoverage rate for the asymptotic confidence interval and confidence sequence.

served data. To avoid this issue, the sample size must be fixed independently of the results. Consequently, we cannot reuse the first n samples to draw a new conclusion. Instead, any further inference must rely on a newly collected set of gameplay records that is independent of the initial n games.

This example illustrates an important limitation: the CIs are only valid at some non-data dependent sample size. Any inference made at a data-dependent stopping time (e.g., under continuous monitoring) invalidates the statistical guarantee. This limitation becomes even more apparent for asymptotic CIs that require large prespecified sample sizes, especially when the collection of data is expensive. Asymptotic CIs are central limit theorem-based CIs which are only approximately valid for large sample sizes and have the advantage that the intervals are approximately valid under weaker modeling assumptions.

Rather than relying on asymptotic confidence intervals (CIs), we consider the recently developed notion of asymptotic confidence sequences (CSs), which retain their validity under arbitrary data-dependent stopping times [3]. In other words, asymptotic CSs provide anytime-valid uncertainty quantification. A key advantage of asymptotic CSs is that they enable continuous monitoring of the estimates while allowing additional data to be collected adaptively. Although non-asymptotic CSs have long been studied in the statistics literature, their applicability to general reinforcement learning settings is often limited by restrictive assumptions required for validity.

We derived asymptotic CSs for the general RL setting [1]. In Figure 1(d) we compare the CS and CI at several sample sizes for the Gridworld example. We see that the CI did not cover the true value at several sample sizes. Figure 1(e) shows the cumulative miscoverage rate, i.e., the probability of failing to

capture the true value at any sample size up to n . The cumulative miscoverage rate is estimated by repeating the experiment in Figure 1(d) many times. In contrast to the asymptotic CI, we see that the miscoverage rate of the asymptotic CS is bounded uniformly over the sample sizes by the Type-I error rate.

Our research demonstrates that it is possible to make safe decisions with predictable outcomes in unknown environments while continuously monitoring incoming data without inflating the Type-I error rate. This paves the way toward truly safe reinforcement learning, where data collection under a current policy can be terminated as soon as there is sufficient statistical evidence that an alternative policy performs better. By enabling anytime-valid uncertainty quantification throughout the decision process, carefully designed policies can support more reliable and trustworthy decision-making.

References:

- [1] S. Bongers, F. A. Oliehoek, and M. T. J. Spaan, “Asymptotic anytime-valid off-policy evaluation for reinforcement learning,” work in progress, 2026.
- [2] N. Kallus and M. Uehara, “Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning,” *Operations Research*, vol. 70, no. 6, pp. 3282–3302, 2022.
- [3] I. Waudby-Smith, D. Arbour, R. Sinha, E. H. Kennedy, and A. Ramdas, “Time-uniform central limit theory and asymptotic confidence sequences,” *Annals of Statistics*, vol. 52, no. 6, pp. 2613–2640, Dec. 2024.

Please contact:

Stephan Bongers
 CWI, The Netherlands
 stephan.bongers@cwi.nl

Scrutinizing Banks' Risk Forecasts with E-backtesting

by Ruodu Wang (University of Waterloo)

A model-free method lets regulators and financial institutions continuously monitor tail-risk forecasts using e-values that remain valid whenever checked.

After every market shock, the same question arises for financial institutions and regulators: were market-risk forecasts adequately computed and reported? Under the Basel regulatory framework [1], banks report daily risk measures that determine how much capital they must hold. Regulators, as well as banks themselves, then check whether these forecasts are consistent with data; this task is referred to as backtesting.

Backtesting risk forecasts is almost as old as measuring financial risk itself. For a long time, the main target in finance was Value-at-Risk (VaR). In simple terms, VaR is a high-level, usually 99%, quantile of the future loss distribution over a fixed horizon. It has a simple model-free backtesting culture: count the days on which losses exceed the forecast. This requires a fixed testing period; in statistical jargon, it is not an online method and does not naturally handle data arriving sequentially, typically every trading day. The first challenge is how to backtest risk forecasts online.

The task became much harder when Expected Shortfall (ES), defined as the expected loss given that the loss has reached a high level, replaced VaR as the standard risk measure for market risk in Basel's 2019 regulatory framework. ES looks into the tail and captures the severity of large losses, which is why it is favoured by regulators and risk management scholars. It

has also become popular in many areas involving risk assessment, including portfolio selection, robust optimisation, and machine learning. But this extra tail sensitivity comes at a statistical cost: ES does not admit a simple model-free backtesting procedure like VaR, even with a fixed data collection horizon. Thus, the second challenge is how to backtest ES in a model-free manner.

Evidence that can be checked at any time

With Qiuqi Wang and Johanna Ziegel, we developed the methodology of e-backtesting [2], a procedure that uses e-values and e-processes to monitor risk forecasts sequentially. An e-value can be read as the payoff of a fair bet against the hypothesis that the forecasts are adequate. If that hypothesis is true, the expected payoff is at most one. Multiplying daily e-values gives an evidence process whose guarantee is anytime-valid: a regulator or risk manager may look today, tomorrow or after a crisis week without invalidating the test. If the e-process crosses a threshold, the probability of such a false alarm under adequate forecasting is controlled by that threshold, no matter when the crossing occurs. This differs from a conventional p-value workflow, where tests are calibrated for one planned analysis and repeated peeking can distort error guarantees.

Addressing both challenges at once

The paper [2] introduces backtest e-statistics for risk-measure forecasts and characterises essentially unique optimal choices for VaR and for the ES-VaR pair. Each day's loss and forecasts are converted into an e-value. Conservative forecasts, which are acceptable in prudential regulation, keep the process low; forecasts that repeatedly fall short of realised tail risk push it upward.

This methodology leaves flexibility in how the e-process is constructed. To make it powerful in practice, we propose data-driven rules that approximate a growth-rate-optimal principle for e-processes. The recommended default, GREM, combines

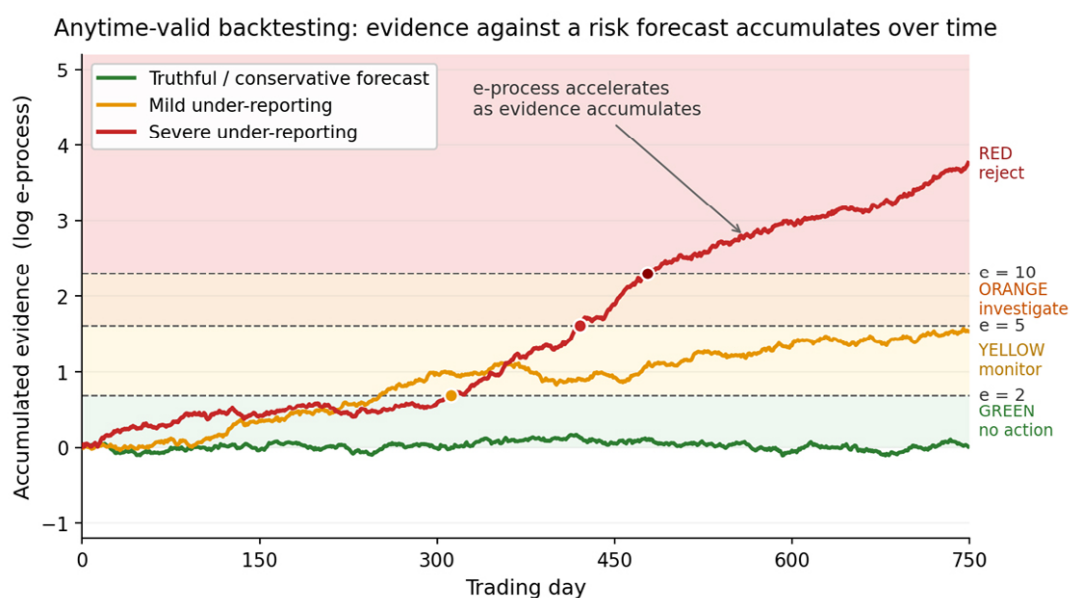


Figure 1: Stylised illustration of e-backtesting. Three banks report risk forecasts every trading day. An accurate or conservative forecast keeps the e-process near or below 1; mild or severe under-forecasting pushes it upward. Banks do not know their true risk values either; they can only try to forecast responsibly.

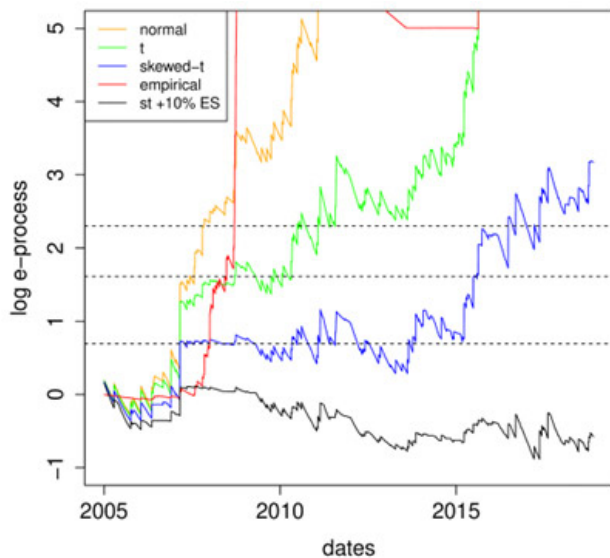


Figure 2: E-processes (GREM method) for backtesting daily ES forecasts on the NASDAQ Composite, January 2005 to December 2021. Dashed lines mark thresholds $e = 2, 5$ and 10 . Empirical and Gaussian forecasts are rejected quickly; Student-t and skewed-t forecasts are rejected more slowly; the deliberately conservative forecast is never rejected. Reproduced from [2].

two adaptive strategies and performs competitively while avoiding assumptions on the forecasting mechanism and the loss distributions.

Because evidence accumulates gradually, the method naturally supports a multi-zone alarm system. A regulator could treat moderate thresholds as early warnings, higher thresholds as substantial evidence, and still higher thresholds as decisive evidence. The interface could be one curve per institution: green for no action, yellow for monitoring, orange for investigation and red for rejection (Figure 1).

What happens on market data

As a simple example, we applied the method to daily NASDAQ Composite returns from January 2005 to December 2021. Several standard forecast methods produced 97.5% ES forecasts. Empirical and AR-GARCH forecasts with Gaussian innovations were rejected decisively within months of the 2007-2008 financial crisis. Forecasts based on Student-t or skewed-t innovations survived longer but eventually accumulated evidence against them. A deliberately conservative forecast, set 10% above the skewed-t estimate, was not rejected (Figure 2).

This is the behaviour a regulatory backtest should encourage. The method is not trying to punish cautious capital estimates, nor does it presume bad intent. It seeks persistent evidence that forecasts are inadequate, whether because of model misspecification, structural change, statistical difficulty or under-reporting. Even a skewed-t AR-GARCH model, often regarded as a strong model for financial log-returns, still underestimated the market risk in this period.

Beyond banking

Although motivated by financial regulation, the same idea applies more broadly, as shown by the theory developed in the

paper. Whenever a forecast target, such as the mean, the variance, or another risk measure, admits a suitable backtest e-statistic, an e-process can turn repeated forecast evaluation into an anytime-valid monitoring tool. E-backtesting marks a shift from fixed-deadline tests to cumulative evidence that respects how decisions are actually made.

References:

- [1] Basel Committee on Banking Supervision: Minimum Capital Requirements for Market Risk. BIS, 2019. <https://www.bis.org/bcbs/publ/d457.pdf>
- [2] Q. Wang, R. Wang and J. Ziegel, “E-backtesting”, *Management Science*, 72(6):4952-4973, 2026. <https://doi.org/10.1287/mnsc.2023.01659>

Please contact:

Ruodu Wang
 Department of Statistics and Actuarial Science
 University of Waterloo, Canada
wang@uwaterloo.ca

Anytime-Valid Testing in the Age of AI-Assisted Software Development

by Michael Scott Lindon (Netflix)

The statistical guarantees designed to protect against human failures in sequential experimentation turn out to be exactly what is needed to govern autonomous AI agents conducting experiments.

Since the early 2010s, technology companies have been running hundreds to thousands of A/B tests per day. The canonical example of an A/B test is one testing button colours or reworded headlines – visible product elements designed to improve outcomes such as engagement. In reality, the majority of modern A/B tests are not of this kind, but are employed as quality control gates for safely rolling out new features. This is especially true of A/B tests used to roll out new code into production.

In software engineering, observability refers to the ability to infer the internal state of a system from its external outputs – logs, metrics, and traces. App-load time, response latency, and number of stream rebufferers are just a few examples of measurements found in observability data; a modern emerging metric is the LLM-as-a-judge evaluation, in which a large language model scores the quality of AI-generated outputs. While the earliest uses for observability data were dashboards and alerting, such data can be used to test the performance of an incumbent software version against a newer replacement across an extremely heterogeneous population of hardware devices – what the industry calls a canary test. The idea is simple. Before a software update reaches the entire user population, it is first exposed to a small, randomized subset of devices. Observability data is ingested from control and treatment devices and monitored in real time. If a performance degradation

is detected, the release is aborted, preventing bugs from reaching end-users.

Detection speed is essential. The longer degraded software remains live, the worse the cumulative harm to users. This necessitates sequential statistical methodologies that preserve Type-I error guarantees under continuous monitoring [1]. Canary tests, by the real-time nature and sheer volume of their data, are one of the clearest examples where anytime-valid inference is superior to group sequential testing methods, which require a small, finite number of pre-specified interim analyses.

The richness of observability data presents a further challenge: a single canary experiment may monitor hundreds of metrics simultaneously, making multiple testing corrections unavoidable. These metrics further exhibit a complicated dependency structure. A further advantage of e-processes, beyond preserving Type-I error under optional stopping, is their amenability to multiple testing procedures. The e-BH procedure [2], for example, controls false discovery rate under arbitrary dependence, in contrast to the positive regression dependence assumption required by the original BH procedure.

Catching a failure, however, is only half the problem. When a canary test fails, the engineer must causally attribute the shift in performance metrics to lines of code changed between software versions. This has traditionally been a very manual and painstaking process. Large language models change this picture. AI coding assistants are enabling developers to ship faster than ever, with decreasing direct supervision over what is pushed to production. An LLM can read a code diff and reason about which changes plausibly caused which metric movements. An agent can go further: generate a hypothesis, make a targeted code change, run a new canary, and evaluate the result. This closed loop may iterate until the root cause is isolated and resolved. The canary experiment in the software delivery pipeline is exactly the quality control gate needed to address this concern.

The analogy to clinical diagnosis is apt. A physician orders an initial battery of tests, forms hypotheses from the results, orders further tests to accumulate evidence, and ultimately confirms a diagnosis by observing whether the patient responds to a drug designed for a specific disease. Agentic software debugging follows the same logic, with canary experiments in place of diagnostic tests and targeted code changes in place of drug therapy.

The entire machinery of safe anytime-valid inference fits this agentic workflow with striking coherence. Within a single experiment, anytime-valid tests allow the agent or human operator to stop as soon as there is enough evidence of improvement, degradation, or futility, making experimentation both more efficient and less risky for users. Across the many metrics monitored in each experiment, e-value based multiple testing procedures control false discoveries even under complex dependence. E-values from distinct experiments can also be merged to combine evidence.

Across the unbounded sequence of hypotheses generated by the agent's learn-edit-test loop, online FDR procedures control the rate at which false discoveries enter the agent's evolving

understanding of users. When the agent surfaces conclusions to a human reviewer, e-values also support valid post-selection inference over the hypotheses that survived.

This becomes essential when the workflow is not merely autonomous debugging, but autonomous statistical research. An agent can run an experiment, inspect the results, infer something about users, generate follow-up hypotheses, and launch the next experiments without waiting for a human analyst. Run this flywheel without online FDR and false discoveries are no longer isolated reporting errors. They corrupt the trajectory of the research – steering the agent toward hypotheses built on noise rather than genuine insight into users. Online FDR is the governance layer that keeps this loop from optimizing around noise.

The fit between agentic workflows and the toolkit of anytime-valid inference, e-values, and online FDR is not a coincidence – it runs deeper than analogy. The problems these methods solve are exactly the problems autonomous agents recreate: peeking at accumulating data, launching follow-up studies based on borderline results, and reporting only the post-selected outcomes. Agents do not intend any of this, but they recreate the same statistical failure modes by optimizing for finding an answer. Formal statistical guarantees are the only protection available. Safe anytime-valid inference provides the tools to build a statistical harness for agentic experimentation.

Links:

[L1] <https://kwz.me/hIK>

[L2] <https://kwz.me/hIS>

References:

- [1] M. Lindon et al., “Rapid regression detection in software deployments through sequential testing,” *KDD, ACM*, 2022.
- [2] R. Wang and A. Ramdas, “False discovery rate control with e-values,” *J. R. Stat. Soc. B*, vol. 84, no. 3, pp. 822–852, 2022.
- [3] Z. Xu and A. Ramdas, “Online multiple testing with e-values,” *AISTATS, PMLR*, 2024.

Please contact:

Michael Scott Lindon, Netflix
michael.s.lindon@gmail.com

European Research and Innovation

Large Language Models as Design Partners: Automating Graphical Mockups to Refine Requirements

by Giovanna Broccia, Maurice H. ter Beek (CNR-ISTI), and Alessio Ferrari (University College Dublin and CNR-ISTI)

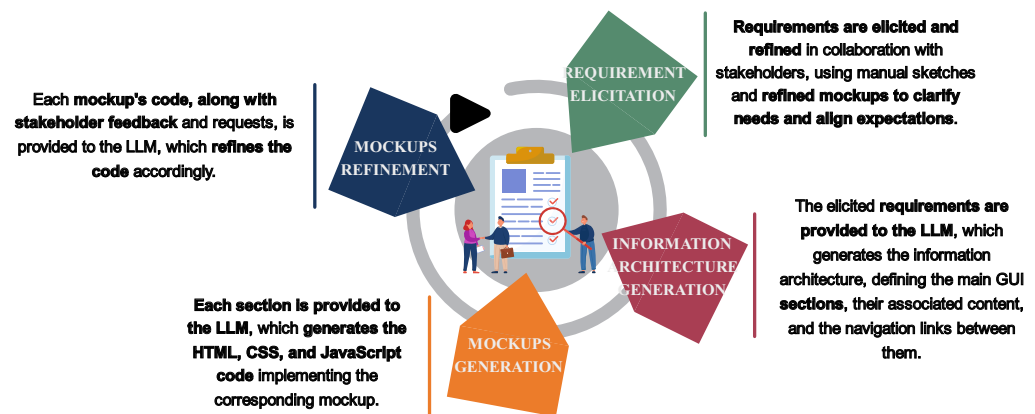
Can large language models help designers move faster without sacrificing human centrality? Researchers from CNR-ISTI and University College Dublin are exploring how large language models can support the rapid creation and refinement of industrial graphical user interfaces through a case study involving the Italian railway operator Trenord, helping development teams move more quickly from textual requirements to interactive mockups, while keeping humans at the centre of the design process.

Creating graphical user interface (GUI) mockups is a crucial step in software design. Mockups help designers, engineers, and end-users discuss requirements, validate ideas, and refine functionalities before implementation begins. However, the process is often time-consuming, requiring multiple design iterations and repeated stakeholder reviews. It is also technically demanding, involving manual coding and debugging activities, especially in industrial contexts where interfaces evolve continuously and many stakeholders are involved [1].

Researchers from CNR-ISTI and University College Dublin collaborated with Trenord [L1], a railway operator in Northern Italy, to investigate how large language models (LLMs) can support GUI design activities by acting as “design partners”. The work was done within Spoke 4 “Railway Transportation” of MOST – the National Centre for Sustainable Mobility [L2], which received EU funding through NextGenerationEU – and focused on the design of a predictive maintenance dashboard for railway operations. The dashboard is intended to support maintenance and engineering personnel in monitoring train fleets, analysing diagnostic information, and inspecting failures predicted by maintenance algorithms. The goal was not to replace designers, but to understand whether LLMs could automate repetitive coding activities and accelerate the transition from requirements to interactive mockups [2].

The proposed approach adopted a human-LLM co-design process consisting of four main activities (see Figure 1). In the first phase, requirements were elicited collaboratively through focus groups, analysis of existing documentation, and discussions with Trenord stakeholders. During this stage, designers also created manual sketches to clarify stakeholders’ needs and align expectations. In the second phase, the elicited requirements were provided to the LLM, which transformed textual requirements into a structured information architecture by identifying dashboard sections and navigation paths. In the third phase, the LLM generated interactive mockups in HTML, CSS, and JavaScript for each identified dashboard section. Finally, the generated mockups were refined iteratively according to

Figure 1: Human-LLM co-design process illustrating how requirements elicitation, information architecture generation, mockup generation, and iterative refinement are collaboratively performed by human designers, stakeholders, and LLMs to accelerate the creation of interactive GUI mockups.



stakeholder feedback. The code of each mockup, together with requests and comments from stakeholders, was provided to the LLM, which updated and refined the interfaces accordingly.

The resulting dashboard included six mockups representing different sections of the dashboard, including fleet overview pages, train configuration views, and detailed prediction panels showing the status of onboard components. Rather than static images, the generated outputs were interactive interfaces that stakeholders could immediately visualise, interact with, and discuss. This rapid feedback cycle enabled participants to identify missing requirements and propose changes early in the design process.

One of the main outcomes of the project was the observation that LLMs can significantly reduce the effort spent on repetitive technical activities, such as coding layouts and refining HTML/CSS structures. This allowed designers to focus more on stakeholder interaction, requirement clarification, and design reasoning. Importantly, the proposed methodology accelerates clerical activities such as programming and debugging while keeping humans central throughout the entire process – from requirements elicitation and refinement, via information architecture definition, to the design and evaluation of the mockups themselves. The most important contribution of the approach is therefore not full automation, but the possibility of supporting collaboration and accelerating iteration cycles.

The generated mockups were later presented and discussed in a focus group involving three members of Trenord's engineering personnel. During the session, the interfaces were shown as an operational dashboard supporting two classes of users: maintenance personnel and engineering personnel. Stakeholders were asked to provide feedback both on the generated sections and on the overall process.

The participants described the generated interfaces as more "concrete" and closer to the final product than manually created sketches. Because the interfaces could be modified rapidly, discussions became more dynamic and productive. Stakeholders were able to request immediate changes, validate functionalities earlier, and better understand how the final system could evolve. Rather than constraining creativity through predefined solutions, the approach appeared to support continuous exploration and refinement of ideas.

The experience also highlighted several practical lessons. First, requirements quality strongly influences the quality of the generated interfaces: ambiguous or incomplete requirements frequently lead to unsatisfactory outputs. Second, human involvement remains essential throughout the process: the LLM can generate and refine interfaces quickly, but humans are still responsible for defining goals, validating outputs, and guiding the design direction.

Another important lesson concerns the choice of the LLM itself. Different models showed different strengths in terms of visual quality, consistency, and adherence to requirements. We also observed that prompting style plays a key role: structured and sequential prompts significantly improved the coherence of generated interfaces and reduced the number of refinement iterations needed.

Beyond the railway domain, we believe that this approach could support many industrial design activities where rapid prototyping and continuous stakeholder feedback are essential. Future work will involve analysing how LLMs could generate multiple interface alternatives automatically and how mockups themselves could help refine requirements in future design cycles.

Links:

[L1] <https://www.trenord.it/en/>

[L2] <https://www.centronazionalemost.it/en/>

References:

- [1] T. R. Silva et al., "A Comparative Study of Milestones for Featuring GUI Prototyping Tools", *J. Softw. Eng. Appl.*, vol. 10, no. 6, 2017, 564–589. DOI: 10.4236/jsea.2017.106031
- [2] G. Broccia et al., "An experience report on leveraging LLMs for GUI generation: Automating coding to prioritise creativity," in *Joint Proc. of REFSQ 2025 Co-Located Events, CEUR Workshop Proc.*, vol. 3959, 2025. <https://ceur-ws.org/Vol-3959/CreaRE-paper2.pdf>

Please contact:

Giovanna Broccia
CNR-ISTI, Italy
giovanna.broccia@isti.cnr.it

Challenges in Small-Scale Medical Data Exchange Platform Development

by Hubert Schölnast, Peter Kieseberg, Patrick Kochberger and Henri Ruotsalainen (University of Applied Sciences St. Pölten)

While the concepts of data sharing and data reuse are simple in theory, they face a plethora of challenges and obstacles when transferred into real-life applications. In this article we discuss the major challenges encountered in the successful construction of a sharing infrastructure for oncological data, as well as best practices and learnings in order to overcome similar issues.

Utilization of health data has a long tradition, especially when referring to concepts like P4 medicine [1]. While the task seems trivial in theory, there are many obstacles in practice, especially when considering small teams targeting very specialized data and research questions below the multi-million euro development budgets. In this paper we discuss learnings from previous projects that acted as basis for the practical implementation of an oncological platform in Lower Austria together with new challenges derived from the changes in the geopolitical landscape. Figure 1 provides an overview of these selected challenges.

Obstacle 1 – Anonymisation versus Pseudonymization: One major issue for misunderstanding is the term “anonymization”. While it is clearly reserved in IT Security and regulations like the General Data Protection Regulation (GDPR) as methods that do not allow re-identification of persons in data sets, neither directly through identifiers nor indirectly through a combination of so-called “quasi identifiers”, it is often used as a word for “masking” data in the sense of using pseudonyms for the identifiers in medical environments. While this seems to be a trivial issue on first glance, it results in completely different protection levels: Regulations like the GDPR enforce strict protection of non-anonymous data, which also explicitly includes pseudonymized data with respect to e.g., consent, right to deletion and others. Furthermore, typically applicable anonymization mechanisms result in data distortion which could be detrimental to the analysis conducted by AI algorithms.

This topic is a typical example for different interpretations of technological terms in different fields. Thus, defining each of these terms is of major importance. Furthermore, we found

that referencing specific standards or regulations can greatly enhance the quality of understanding between experts from different fields.

Obstacle 2 – Digital Sovereignty: Due to recent reshaping of the geopolitical landscape, the topic of European digital sovereignty emerged as a major issue, especially considering the high risk involved when dealing with personal medical information for the patients. In addition, as technological support becomes a bargaining chip in geopolitical games, requiring availability of foreign technological platforms becomes a liability. While the decision to stay with purely European platforms needs to be decided on case by case, the principle issue of digital sovereignty needs to be taken into consideration.

Furthermore, upcoming regulations are currently foreseen to strongly emphasize on utilizing European platforms wherever possible, in order to decouple from US dominance, adding further complexity to the design phase. Thus, the requirement for using sovereign software and technologies needs to be discussed and clarified.

Obstacle 3 – Bills of Materials: In tandem with the topic of digital sovereignty, so-called “Bills of Material”

(BOMs) [2] become increasingly important. While this is already required for Software under certain circumstances and is a major mechanism for ensuring digital sovereign as outlined in Obstacle 2, the nature of modern machine learning algorithms also requires a BOM for the data that is used for training the models, since these models define the actual behaviour of the algorithms. Furthermore, this test data might not only unwillingly introduce biases but can also be used to include specific miscalculations on purpose. Thus, the Data BOM needs to make sure that it covers the actual data that was used for training, which is typically hard to prove in case of “training as a service” or “model as a service” environments. While this can be solved by training the model on premise under the control of a trusted supervisor, this certainly increases the costs, as model training requires a lot of hardware resources. In case of utilizing cryptography for data protection, “Cryptographic Bills of Materials” (CBOMs) are required as well, not only catering for the algorithm that was used, but for the specific implementation, as this is a major weak spot in the application of encryption.

Obstacle 4 – Legal and regulatory realities: While sharing seems to be trivial, and the FAIR criteria lay out respected guidelines on what needs to be fulfilled for future proof information exchange accompanied by a lot of research and best practices on how to achieve it, the practical implementation in actual projects requires a lot of legal involvement, especially when dealing with different partners for data provisioning and analysis: Different legislations have different rules on the

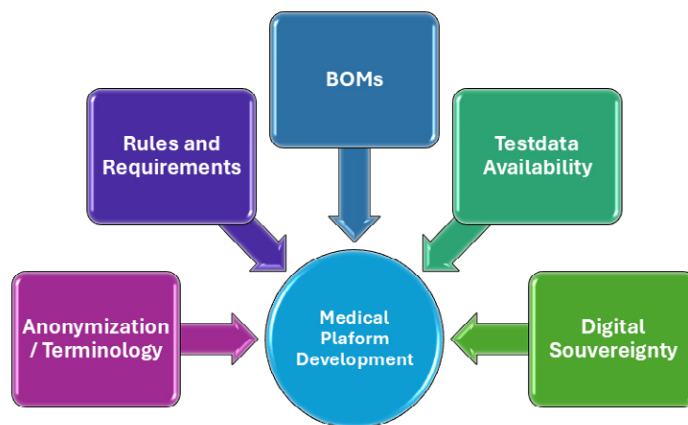


Figure 1: Five selected challenges for real-world medical data exchange environments.

treatment of patient data, as this topic is typically heavily nationalised. Even within the same legal framework, partners have different internal rules and responsibilities, as well as different strategies regarding risks. This alignment is often overlooked in projects driven from technical / research side and can result in long delays and critical uncertainties.

Obstacle 5 – Testdata for Development: A major obstacle in the development of AI-enhanced medical research platforms lies in the provisioning of data for development and testing purposes: While this is often clarified for the actual operational workflow, where the platform is e.g., under the control of the original data owner, things become far more muddled in the development phase, when external partners need access to data and formats, especially when externalizing AI model training. While synthetic data can of course solve this problem, this means that data generators need to be developed that cover the full width of possible test cases. This test coverage is even a problem when using real data in case of rare cases and thus requires far more data engineering than typically assigned to this task.

In conclusion, the practical implementation of data driven processes encounters practical difficulties, especially when dealing with multiple partners in complex legal and ethical environments. In order to navigate around these obstacles, the APOG project [L1] has defined guidelines for qualified interviews that clarify these issues in a structured and comprehensive manner that will be published in the course of the project.

Link:

[L1] <https://kwz.me/h14>

References:

- [1] P. Sobradillo, F. Pozo, and Á. Agustí, “P4 medicine: The future around the corner,” *Archivos de Bronconeumología (English Edition)*, vol. 47, no. 1, pp. 35–40, 2011.
- [2] É. Ó. Muirí, “Framing software component transparency: Establishing a common software bill of material (SBOM),” *NTIA*, Nov. 12, 2019.
- [3] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

Please contact:

Hubert Schölnast
University of Applied Sciences St. Pölten, Austria
Hubert.Schoelnast@ustp.at

Building Sovereign AI for Europe’s Energy Networks

by Antonello Monti (Fraunhofer Institute for Applied Information Technology FIT, Germany)

European electricity networks are becoming increasingly complex as renewable energy sources, electrification and cross-border interconnections continue to grow. The AI.Grids initiative brings together 48 European organizations to develop open, trustworthy and sovereign AI models and data foundations tailored to the needs of Europe’s critical energy infrastructure.

Artificial intelligence is rapidly becoming a key enabling technology for the operation of critical infrastructures. In the energy sector, AI promises to improve forecasting, operational planning, system monitoring and decision support. At the same time, growing dependence on proprietary technologies developed outside Europe raises concerns about transparency, trustworthiness, long-term availability and technological sovereignty.

To address these challenges, 48 European organizations have joined forces in the AI.Grids initiative [L1], a large-scale collaborative effort dedicated to developing open, trustworthy and European-built AI models for electricity network applications. Supported by the European Commission and aligned with the broader European strategy for the digitalization of the energy system [1], the initiative represents one of the most ambitious public-private collaborations currently underway at the intersection of artificial intelligence and energy systems.

The increasing complexity of modern electricity networks creates a strong need for advanced analytical capabilities. The large-scale integration of renewable energy sources, the electrification of transport and industry, and the growing interconnection of national energy systems require operators to manage unprecedented amounts of data and increasingly dynamic operating conditions. AI technologies offer considerable potential to support these tasks while improving efficiency, resilience and sustainability, reflecting priorities identified in the European Commission’s Digitalization of the Energy System strategy [1].

AI.Grids aims to establish a distinctly European approach to AI for critical infrastructures. Beyond technological sovereignty, the initiative also addresses a significant gap in the current landscape: there are currently no foundation models specifically designed for electricity grids. AI.Grids therefore seeks to explore and evaluate different architectural approaches for developing AI models tailored to power system applications. While international initiatives such as GridFM have started exploring this area, AI.Grids seeks to evaluate alternative approaches and to develop models trained on European data and infrastructure characteristics. Rather than relying exclusively on externally developed technologies, the initiative seeks to create AI models that are developed, governed and maintained within Europe. The objective is not only technological excellence but also transparency, accountability and alignment with European values and regulatory frameworks.



Figure 1: Launch of the AI.Grids initiative in Brussels on 3 June 2026. Representatives of 48 European organizations signed the agreement establishing a collaborative effort to develop sovereign and open AI solutions for Europe's electricity networks. Photo: EU/Claudio Centonze.

The initiative originated from a proposal developed by Antonello Monti of the Fraunhofer Institute for Applied Information Technology FIT together with Ricardo Bessa from INESC TEC and Elissaios Sarmas from the National Technical University of Athens. The project has evolved through close collaboration between the three institutions and a growing European community of energy and AI researchers. The concept received support from the European Commission's Directorate-General for Energy and subsequently evolved into the current AI.Grids program.

A central principle of AI.Grids is openness. Participating organizations have agreed to adopt an open-source development model that enables broad participation across the European energy ecosystem. This approach encourages collaboration among researchers, technology providers, grid operators and public institutions while fostering transparency and trust in the resulting AI solutions.

The governance framework is coordinated by CRESYM, a European research association responsible for the organizational and legal foundations of the initiative. Participating organizations contribute expertise, data and technological resources. Within the consortium, the Fraunhofer Center Digital Energy plays a key operational role by leading activities related to the definition and establishment of the data foundations required for future AI model development.

The quality and representativeness of data are critical factors for the successful deployment of AI in electricity systems. One of the first objectives of AI.Grids is therefore to establish a common data foundation for model development and training. The initiative combines operational data contributed by participating organizations with synthetic datasets derived from information about European energy infrastructures. These data assets will support the evaluation of different AI architectures and provide the foundation for AI models trained on European operational and infrastructure data.

Initial work within AI.Grids focuses on two priority application domains. The first concerns the static representation of electricity networks and related tasks such as power flow analysis. The second addresses time-series forecasting for electricity generation and demand. These areas were selected following a series of workshops in which candidate use cases were

identified, prioritized and grouped according to their relevance and potential impact.

The agreement signed in Brussels on 3 June 2026 marks the first operational milestone of AI.Grids. The initiative remains open to additional participants, allowing further organizations to contribute during the coming years and helping to establish a broad European community around AI for energy systems.

The first operational software release is planned for early 2027. Developed according to open-source principles, the resulting software will be made available to participating organizations, with particular focus on the needs of European transmission and distribution system operators. By facilitating interoperability and supporting the deployment of innovative technologies, the initiative is expected to contribute to a more resilient and efficient European energy system. Representatives of participating organizations gathered in Brussels for the signing ceremony that officially launched the initiative (Figure 1).

AI.Grids demonstrates how Europe can pursue an alternative path for the development of artificial intelligence in critical infrastructures. By combining scientific excellence, industrial expertise and public-sector engagement, the initiative seeks to ensure that the benefits of AI-driven innovation remain under transparent governance and serve the broader interests of European society.

Link:

[L1] <https://cresym.eu/ai-grids/>

Reference:

[1] European Commission, Directorate-General for Energy, "Digitalisation of the energy system." [Online]. Available: <https://kwz.me/hlc>

Please contact:

Antonello Monti, Fraunhofer Institute for Applied Information Technology FIT, Germany
antonello.monti@fit.fraunhofer.de

SCIANCE: AI for Scientific Discovery in Europe

by András Benczúr, Edina Nemeth (SZTAKI), Jonas L'Haridon (European Science Foundation) and Magdalena Brus (EGI Foundation)

Artificial Intelligence (AI) is changing how scientific research is conceived, executed and interpreted, from analysing massive astrophysical data streams to accelerating drug discovery and improving climate and environmental modelling. Yet, the European landscape of AI enabled research remains fragmented: scientific communities, AI experts and research infrastructures often work in parallel rather than together, and strategic guidance on where to invest and how to coordinate efforts is still emerging. The SCIANCE project was launched to address this fragmentation and to help Europe turn AI into a coherent, shared engine for scientific discovery.

SCIANCE (Grant Agreement No. 101293570) is a 30 month coordination and support action that brings together 13 scientific organisations, research infrastructures and digital infrastructure providers from across Europe. Its overarching ambition is to help Europe develop a coordinated AI for Science ecosystem by developing a Strategic Research and Innovation Agenda (SRIA) for AI in Science, an implementation roadmap for infrastructure upgrades, and a long term cooperation framework within the Resource for AI Science in Europe (RAISE) initiative of the EU [L1].

At the heart of SCIANCE is the recognition that AI in Science is not a single domain, but a cross cutting transformation that affects many disciplines and infrastructures in different ways. To capture this diversity, the project adopts a cross domain perspective, focusing on five pilot areas where Europe has strong scientific leadership: astronomy and fundamental physics, materials science, Earth and environmental sciences, life sciences, and social sciences and humanities. Across these pilots, SCIANCE also examines transversal AI methods and approaches – such as machine learning, generative and foundation models, symbolic and knowledge based AI, hybrid and physics informed models, autonomous and agentic systems, and frugal, resource efficient AI – and how they map onto different stages of the scientific lifecycle, from hypothesis generation and experimental design to data analysis, simulation, automation and open science.

AI in Science Working Groups (AISWGs) play a central role throughout this process. They connect domain scientists, AI experts, infrastructure operators, industry innovators and open science advocates around thematic priorities, ensuring that landscape analyses and priority

setting exercises remain grounded in real research practices and infrastructure constraints.

The project structure reflects this dual focus on disciplinary depth and cross cutting foundations. In its first phase, SCIANCE consolidates knowledge through a landscape analysis. This includes systematic state of the art reviews of AI applications in the pilot domains, a differential mapping of AI research and technological developments relevant for science, and a landscape analysis of European AI infrastructures and initiatives, supported by the OpenAIRE Graph as an evidence base. In parallel, SCIANCE documents good practices in AI enabled science using a context mechanism outcome framework, making visible not only successful applications but also the conditions under which they work, their limitations and their implications for transparency, ethics, legal compliance, openness and frugality.

In the second phase, SCIANCE moves from evidence consolidation to co creation of priorities. A series of domain specific workshops brings together leading scientists, AI researchers, infrastructure providers and stakeholders from relevant European initiatives to identify long term research challenges where AI can make a decisive difference. Complementing these are transversal AI in Science workshops that focus on AI enhanced literature analysis and open science, data collection and processing, experimental design and policy support, automation of laboratory workflows and collaboration, and the application of frugal AI models. A dedicated interdisciplinary workshop then addresses cooperation priorities for AI model development and sharing, looking at skill profiles, governance and regulation, and practices and infrastructures for FAIR and responsible data sharing.

The third phase of SCIANCE translates these priorities into the Strategic Research and Innovation Agenda and a roadmap for actionable infrastructure scenarios. A distinctive feature of SCIANCE is its emphasis on sustainable coordination and community building. Beyond the SRIA and roadmap, the project pilots RAISE – the Resource for AI Science in Europe – as a long term coordination structure that will continue to support AI enabled science after the project ends. This encom-

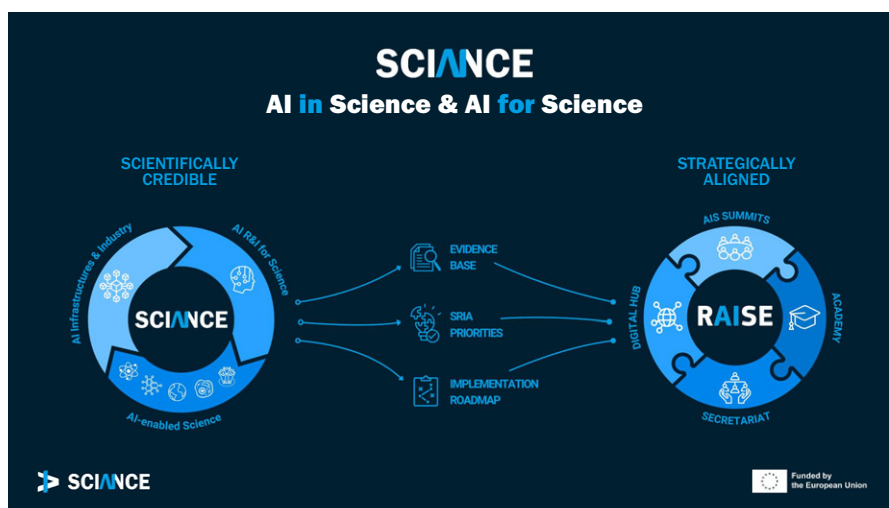


Figure 1: The SCIANCE framework links AI-enabled scientific communities with AI research and infrastructures to develop an evidence base, a Strategic Research and Innovation Agenda (SRIA), and an implementation roadmap that contribute to the long-term RAISE initiative.

passes multiple elements: a Secretariat that provides governance and programme management; a Digital Hub that serves as a central entry point for information, engagement and knowledge exchange; contributions to AI in Science Summits that convene communities annually; and the RAISE Academy, which equips policymakers, funders and executives with tools to translate SRIA priorities into concrete policy and funding measures. In this way, SCIANCE goes beyond roadmapping to address strategic capacity and governance, recognising that AI in Science requires not just algorithms and hardware, but informed, coordinated decisions.

For the ERCIM community, SCIANCE is particularly relevant because it treats AI in Science as both an informatics and a mathematical challenge, and because it builds on and complements existing European networks rather than duplicating them. By providing a structured, community validated SRIA and roadmap, SCIANCE offers a shared frame of reference for future projects, infrastructure investments and policy measures in AI for Science, including those emerging from ERCIM members.

Ultimately, SCIANCE aims to ensure that AI methods and tools are adopted in science in ways that are scientifically robust, socially responsible and strategically aligned with European values and goals. Its integrated methodology – combining meta analysis, landscape mapping, good practice registries, co creation workshops, feasibility assessments and capacity building – is designed to turn disparate initiatives into an evolving, interconnected AI in Science ecosystem. As AI continues to redefine what is possible in scientific research, SCIANCE seeks to help Europe harness these possibilities deliberately, transparently and collaboratively.

The consortium brings together 13 leading scientific organisations, research infrastructures, and digital infrastructure providers from 9 European countries, combining expertise in science, artificial intelligence, policy, and infrastructure operation.

Consortium partners include: European Science Foundation (ESF), EGI Foundation (EGI), OpenAIRE, Consiglio Nazionale delle Ricerche – ISTI (CNR-ISTI), EuroBioImaging ERIC, Constructor University, University of Manchester, University of Twente (ITC), National Institute for Subatomic Physics (NIKHEF), German Research Center for Artificial Intelligence (DFKI), HUN-REN Institute for Computer Science and Control, and the Big Data Value Association (BDVA).

Link:

[L1] <https://science.eu>

Please contact:

Jonas L'Haridon, Project Coordinator
European Science Foundation (ESF), France,
jlharidon@esf.org

Magdalena Brus, Communications Manager
EGI Foundation (EGI), The Netherlands
magdalena.brus@egi.eu

Landscape analysis:

András Benczúr and Edina Németh, SZTAKI, Hungary,
benczur@sztaki.hu, edina.nemeth@sztaki.hu



HORIZON Europe: ERCIM Project Support Services

European projects provide an excellent opportunity to advance research and innovation while collaborating with leading partners across Europe. Through ERCIM, our member institutes have participated in more than 100 European Commission-funded ICT projects, combining scientific excellence with professional project management. The ERCIM Office supports these collaborations by providing expertise in project coordination, administration, financial management, as well as communication and dissemination.

HORIZON Europe: How can you get involved?

The ERCIM Office provides expert support across the full project lifecycle, including:

- Identification of funding opportunities
- Recruitment of project partners (within ERCIM and through our networks)
- Proposal writing and grant negotiation
- Contract and consortium management
- Communication and IT systems support
- Organisation of meetings, workshops and conferences
- Dissemination of project results.

How does it work in practice?

Contact the ERCIM Office to present your project idea. A panel of experts from the ERCIM Science Task Group will review it and provide recommendations. Based on this assessment, the ERCIM Office will decide whether to support the preparation of your proposal. Please note that at least one ERCIM member institute must participate in the proposal for the ERCIM Office to become involved.

If the ERCIM Office agrees to participate, it may join the consortium either as project coordinator or as a project partner, providing the services described above.

For more information, please contact:

Jessica Michel Assoumou
ERCIM Office
jessica.michel@ercim.eu



SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Call for Proposals

Dagstuhl Seminars and Perspectives

Workshops

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is accepting proposals for scientific seminars/workshops in all areas of computer science, in particular also in connection with other fields.

If accepted, the event will be hosted in the seclusion of Dagstuhl's well known, own, dedicated facilities in Wadern on the western fringe of Germany. Moreover, the Dagstuhl office will assume most of the organisational/ administrative work, and the Dagstuhl scientific staff will support the organizers in preparing, running, and documenting the event. Thanks to subsidies the costs are very low for participants.

Dagstuhl events are typically proposed by a group of three to four outstanding researchers of different affiliations. This organizer team should represent a range of research communities and reflect Dagstuhl's international orientation. More information, in particular details about event form and setup, as well as the proposal form and the proposing process, can be found on

<https://www.dagstuhl.de/dsproposal>

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is funded by the German federal and state government. It pursues a mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

Important Dates

- *Next submission period:*
15 October to 1 November 2026
- *Seminar dates:*
Between October 2027
and September 2028 (tentative).

Call for Participation

ERCIM Forum Beyond Compliance on Digital Ethics in Research

The 5th edition of the ERCIM Forum Beyond Compliance on Research Ethics in the Digital Age will take place on 29–30 October 2026 in Porto, Portugal, hosted by INESC TEC.

The forum provides a space for researchers and practitioners to discuss ethical challenges arising from digital technologies in real-world research settings. This year's programme will address research involving young publics, AI companions, limitations of explainable AI (XAI), AI in healthcare, and digital ethics assessment procedures.

Participation is free of charge. The event will be co-located with the ERCIM Fall Meeting, an AIOLIA training workshop, and the ECSS Conference of Informatics Europe.

Training in Digital Ethics

Registrations are now open for a one-day Training in Digital Ethics, organised within the Horizon Europe AIOLIA project (Artificial Intelligence in Human Cognition and Behaviour), in which ERCIM is a project partner. The training is delivered by INESC TEC, an ERCIM member institute, and explores the ethical challenges arising from AI systems that shape human decision-making and interaction.

The course will take place **28 October 2026** at INESC TEC in Porto, Portugal, and is intended for early-career researchers, PhD students, and professors working in computer science, artificial intelligence, and digital technologies. Through lectures, discussions, and practical exercises, participants will gain an understanding of the main ethical principles of trustworthy AI and learn how to identify and address ethical issues throughout the AI system lifecycle.

The afternoon session focuses on applying AI ethics in practice. Using real-world case studies developed within AIOLIA, participants will explore ethical trade-offs—for example, between safety, privacy, explainability, non-discrimination, and human autonomy—and examine practical measures for responsible AI design and governance.

The training will be led by Alexei Grinbaum and Alexandra Prigent (CEA-Saclay), together with Laurynas Adomaitis (RISE Research Institutes of Sweden), internationally recognised experts in AI ethics and governance. Participation is free of charge, but places are limited and registration is required. Researchers interested in integrating ethical reflection into AI research and development are encouraged to apply early.

More information: <https://www.inesctec.pt/en/events/training-in-digital-ethics>



The forum can only be attended in person; no online participation or live streaming will be available. However, recordings of most talks and event summaries will be published online after the event.

The forum is organised by the ERCIM Working Group on Digital Ethics, which welcomes new members and contributors.

We look forward to welcoming you to Porto.

Emma Beauxis-Aussalet and Laurynas Adomaitis, on behalf of the Organising Committee.

More information:
<https://kwz.me/hIE>

ERCIM Working Group “Inclusive Digital Futures” Held Kick-off Meeting

The kick-off meeting of the ERCIM Working Group “Inclusive Digital Futures: Developing Technology and Culture” took place online on 18 May 2026. The meeting introduced the Working Group’s vision and scope, which focuses on moving beyond isolated institutional initiatives towards a more systematic approach within the Computer Science community to foster inclusion and diversity in both the digital sphere and the broader research and innovation ecosystem.

Participants introduced themselves, shared their professional backgrounds, outlined their expectations for the Working Group, and proposed ideas for future activities and contributions. This was followed by a round-table discussion on the Group’s objectives, priorities, and future work plan.

A detailed report summarizing the discussions and contributions will be prepared and shared with all Working Group members.

Several key themes emerged during the meeting. Participants emphasized the importance of community building and research networking to support the development of interdisciplinary research activities. The Working Group was also seen as a platform for translating research outcomes into practical applications and best practices that reflect the diversity of cultures, societies, and user communities. Additional priorities include the development of interdisciplinary pilot studies, knowledge exchange and dissemination activities, and the preparation of joint proposals for relevant European funding opportunities.

Although established within ERCIM, the Working Groups are open to the wider research community and encourage participation from researchers beyond ERCIM member institutions.

More information and contact details:
<https://ercim-ide-wg.ics.forth.gr/>

Call for Participation

MODELS 2026 in Málaga: Advancing Model- Driven Engineering in the Age of AI

Málaga, Spain, 4-9 October 2026,

The ACM/IEEE 29th International Conference on Model Driven Engineering Languages and Systems (MODELS 2026) will take place in Málaga, Spain, bringing together researchers, practitioners and industry leaders working on model driven engineering, modelling languages, and model-based software and systems engineering. Since 1998, the MODELS conference series has been the leading international forum for advances in modelling methods, languages, tools and applications.

MODELS 2026 provides an opportunity for the modelling community to further advance the foundations of modelling, and come up with innovative applications of modelling in emerging areas of digital twins, cyber-physical systems, embedded systems, socio-technical systems, cloud computing, big data, machine learning, security, open source, and sustainability.

The programme features Foundations and Practice tracks, 14 workshops, tutorials, tools & demonstrations, an Industry Day, doctoral and educators symposia, posters, and a student research competition. Co-located with MODELS 2026, there will be the 3rd International Conference on Engineering Digital Twins (EDTconf 2026), the 18th System Analysis and Modelling conference (SAM) and LangDev 2026.

The early-bird registration deadline is 17 August 2026; standard registration opens on 18 August, and on-site registration will be available from 4 October 2026.

More information:
<https://kwz.me/hiH>

Call for Participation

FMICS 2026: 31st International Conference on Formal Methods for Industrial Critical Systems

Liverpool, UK, 2-4 September 2026

FMICS is the annual conference of the ERCIM Working Group on Formal Methods for Industrial Critical Systems and the leading forum at the intersection of formal methods research and industrial applications. The conference aims to bring together researchers, practitioners and tool developers interested in the development and deployment of formal methods for safety-critical systems. ERCIM is pleased to sponsor this year's edition.

As in previous years, FMICS is part of the CONFEST umbrella event consisting of three main conferences CONCUR, FMICS and QEST+FORMATS, and affiliated workshops. The 2026 edition will take place on the main campus of the University of Liverpool, on 1-5 September 2026.

The technical programme features presentations on topics including industrial adoption of formal methods, deductive software verification, concurrent systems, autonomous systems, railway interlocking, model merging and formal verification tools. Invited presentations will be given by Julia Badger (NASA Johnson Space Center, USA), who will discuss the application of formal methods to the design and verification of autonomous space systems, and Colin O'Halloran (University of York and D-RiSQ Ltd, UK), who will address the role of formal methods in reducing time to market for industrial software.

The programme committee is chaired by Peter Gorm Larsen (Aarhus University, Denmark) and Kristin Yvonne Rozier (Iowa State University, USA).

More information and registration:
<https://confest-2026.github.io/fmics/>

ERCIM “Alain Bensoussan” Fellowship Programme

The ERCIM Postdoctoral Fellowship Programme is one of the flagship initiatives of ERCIM. Open to young researchers from around the world, the programme covers a broad range of fields in computer science and applied mathematics.

The fellowship scheme aims to help young scientists deepen their knowledge of European research structures and networks, while gaining valuable experience within leading European research institutions. Fellowships have a duration of 12 months, with the possibility of extension, and are hosted by one of the ERCIM member institutes.

Hosting institutions

Only ERCIM members can host fellows. When an ERCIM member is a consortium, the hosting institute may be any of its member organisations. When an ERCIM member is a funding body, the hosting institute may be one of its affiliated institutions.

Fellowships are offered according to the needs and available funding of the member institutes. Fellows are appointed either through a stipend (a research training agreement) or a work contract, depending on the hosting institute. The type of contract and the

“ The ERCIM Fellowship is a unique opportunity to grow as a researcher in a dynamic, international environment. It offers excellent exposure to cutting-edge research, fosters collaboration across Europe, and supports valuable networking with both academia and industry. I highly recommend it to researchers seeking to broaden their horizons and make a meaningful impact in their field.



Gautam SETHIA
Former ERCIM Fellow



monthly allowance or salary vary by host institution.

ERCIM encourages applications from researchers both in academia and in industry.

Why apply for an ERCIM Fellowship?

The Fellowship Programme enables talented early-career scientists from all over the world to work on challenging research problems alongside leading European experts. In addition to research excellence, the programme fosters collaboration and knowledge exchange within the European research community.

The programme offers ERCIM Fellows the opportunity to:

- Work with internationally recognised experts;
- Gain a deeper understanding of European research structures and networks;
- Become familiar with the working conditions in leading European research centres;

- Promote cross-fertilisation and cooperation between research groups working in similar areas across Europe.

Equal opportunities

ERCIM is committed to ensuring equal opportunities and promoting diversity. Candidates are not discriminated against on the basis of race, colour, religion, gender, national origin, age, marital status, or disability.

Conditions

Candidates must:

- Have obtained a PhD degree within the last eight years (prior to the application deadline), or be in the final year of their doctoral studies with an outstanding academic record. Proof of the PhD degree must be provided before the start of the fellowship;
- Be fluent in English.

Application deadlines

Applications are accepted twice a year, with deadlines on 30 April and 30 September.

Since its inception in 1991, more than 800 fellows have participated in the programme.

The Fellowship Programme is named in honour of Alain Bensoussan, former president of Inria, one of the three founding institutes of ERCIM.

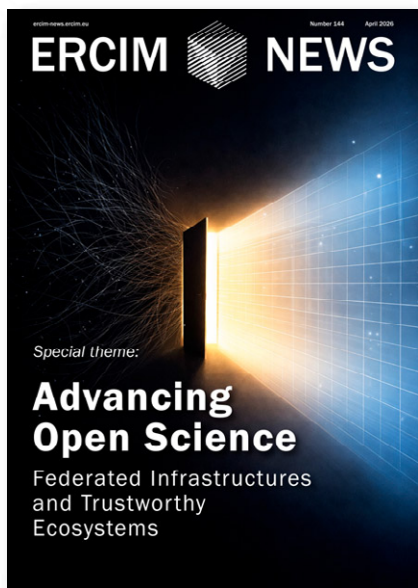
<http://fellowship.ercim.eu>

“ This fellowship will have a strong and lasting impact on my career. It has enabled me to gain valuable international research experience, work in a dynamic research environment, and collaborate with leading researchers in my field. The knowledge, skills and professional connections I have gained during this fellowship will help me to further develop my research, initiate new collaborations, and enhance my academic career in the future.



Jobin FRANCIS
Former ERCIM Fellow





Showcase your EU Project in **ERCIM News**

Reach up to **10,000 readers**
in Europe's ICT research
and innovation community.

ERCIM News is the quarterly magazine of ERCIM,
connecting researchers, innovators, technology leaders
and policy makers across Europe and beyond.

WHY PUBLISH A SPONSORED CONTRIBUTION?



Visibility in a trusted European
research publication



Reach a specialised audience of
researchers, innovators and decision
makers, and stakeholders in the
European Commission, European
Parliament and national
administrations



Distribution through print and digital
channels



Receive a ready-to-use
dissemination text for Horizon
Europe reporting



Long-term online visibility through
the ERCIM News website

SPONSORED CONTRIBUTION

1000 €

Includes:

- Publication in ERCIM News (print and online)
- Professional editorial presentation
- Permanent online availability
- Social media promotion
- Dissemination-ready reporting

REPORTING-READY TEXT

“ The project was featured in ERCIM News,
reaching researchers, innovators and
technology stakeholders across Europe
through print and online dissemination
channels. ”

CONTACT

Peter Kunz
Editor, ERCIM News
peter.kunz@ercim.eu
ercim-news.ercim.eu



ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.



ERCIM is the European Partner of the World Wide Web Consortium.



Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1, 56124 Pisa, Italy
www.iit.cnr.it



I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
www.isi.gr



Centrum Wiskunde & Informatica

Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
www.cwi.nl



Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
<http://www.ntnu.no/>



Fonds National de la
Recherche Luxembourg

Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
www.fnrlu



RISE SICS
Box 1263,
SE-164 29 Kista, Sweden
<http://www.sics.se/>



Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
www.ics.forth.gr



SBA Research gGmbH
Floragasse 7, 1040 Wien, Austria
www.sba-research.org/



Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
www.iuk.fraunhofer.de



Eötvös Loránd Research Network
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
www.sztaki.hu/



INESC
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, n° 378,
4200-465 Porto, Portugal
www.inesc.pt



University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
www.cs.ucy.ac.cy/



Institut National de Recherche en Informatique
et en Automatique
B.P. 105, F-78153 Le Chesnay, France
www.inria.fr



UNIVERSIDAD DE MÁLAGA

Institute for Software Engineering and Software Technology
“Jose María Troya Linero”, University of Malaga
Calle Arquitecto Francisco Peñalosa, 18, 29010 Málaga
<https://gp.uma.es/itis>