Safe anytime-valid inference: from theory to implementation in psychiatry research
Turner, R.J.

Acknowledgments

Without my supervisors Peter Grünwald, Floortje Scheepers and Aki Härmä it would have been impossible to complete the diverse body of work presented in this thesis. Peter introduced me to many new topics in mathematics and theoretical statistics. I really enjoyed our collaboration, combining further developing the safe statistics theory with actual software implementations and simulations concurrently. Floortje was incredibly fast at adapting to all complicated analysis techniques we proposed, and picking out exactly the right things to move psychiatry research along. Aki introduced me to the work of many interesting researchers at Philips, widening my scope and peeking my interest for the potential of the role of NLP in healthcare.

I want to give special mention to Karin Hagoort, teamlead innovation at the Psychiatry department of UMC Utrecht and head of the PsyData team. Apart from helping me with so many organizational challenges during the project and introducing me to many other interesting people in the UMC Utrecht and other mental healthcare institutes, we had many insightful discussions on the implementation of AI in healthcare and Karin's contributions to the three psychiatry-focused chapters in this thesis have been very valuable.

These three chapters would also have been impossible to write without the collaboration with and support from my PsyData colleagues at UMC Utrecht, especially Femke, Saskia, Kees, Vincent, Zimbo, and Willem. I also particularly want to mention the data scientists I collaborated with at Parnassia Groep: Roel, Rosa and Eline. Even though some of you have moved on to different jobs and positions by now, I would be happy to work together or exchange ideas on other data science projects in the future with each of you.

I also want to mention the other Psychiatrists of UMC Utrecht I collaborated with for the work in this thesis: Fleur Velders, Edwin van Dellen, Metten Somers and Yuri van der Does, and of course all of your colleagues who attended my presentations and with whom I worked on smaller data science questions over the years. It has been great to collaborate with clinicians with such an affinity for statistics and computer science, who could really think along critically with the work in this thesis.

My colleagues at CWI - over the years it have been too many to list them all here - greatly contributed toward my development as a statistician. I loved being able to get an insight in the work of theoretical mathematics and machine learning researchers: it was a real eye opener to follow all your work, this stimulated me

to think beyond the (small collection of) machine learning methods most-used in clinical research. I especially want to highlight the collaboration I had with Alexander Ly, collaborating on the safestats software package, who taught me lots of nifty coding tricks.

Lastly I want to mention my colleagues from the EPI consortium, Tim, Corinne, Saba, Jamila, Milen and in particular the PIs and project lead Paola Grosso, Cees de Laat and Sander Klous. I loved that my work was part of a bigger project and to learn about each of your respective fields of research.

Curriculum Vitae

From 2010 until 2013 Rosanne studied Medicine at the Leiden University Medical Center (LUMC). She was admitted to the Honours College of Leiden University, where she was awarded a grant to start a research project at the pathology department of LUMC under supervision of Dr. Hans Baelde, Prof. Kitty Bloemenkamp and Prof. J.A. Bruijn. In 2014 she got the opportunity to continue this research full time directly after obtaining her Bachelor degree, which resulted in the PhD thesis *Endothelial Pathology in Preeclampsia* at the Faculty of Medicine of Leiden University.

During her time as a researcher at LUMC, Rosanne discovered that scientific research and particularly methodology and mathematics interested her the most. Therefore she decided to continue her Master studies in this direction: from 2017 until 2019 she studied Statistical Science for the Life and Behavioral Sciences at the faculty of Science at Leiden University, for which she graduated *cum laude*. During her studies she also worked part-time as a software engineer at El Nino development. She wrote her Master thesis *Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test* under supervision of Prof. Peter Grünwald at CWI, which was awarded the Jan Hemelrijk Award by the Dutch society for statistics and operations research (VVSOR).

After graduating she continued the research on safe statistics started during her master thesis as part of a second PhD trajectory, this time in Mathematics. She worked in the *Enabling Personalized Interventions* consortium under supervision of Prof. Peter Grünwald, Prof. Floortje Scheepers (UMC Utrecht) and Dr. Aki Härmä (Philips research) on implementations of safe statistics and other methods suitable for real-time, federated learning. Rosanne spent half her time focusing on developing statistical methodology in the machine learning group at CWI, and the other half implementing new methods and working as a data scientist at the data science team PsyData at the Psychiatry department of UMC Utrecht. Since finishing her second PhD project, Rosanne has continued working at the Psychiatry department of UMC Utrecht as a clinical data scientist.

Bibliography

R. J. Adams. Safe hypothesis tests for the 2×2 contingency table. Master's thesis, Delft University of Technology, 2020.

C. G. Allaart, B. Keyser, H. Bal, and A. Van Halteren. Vertical split learning-an exploration of predictive performance in medical and other use cases. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington, DC: American Psychiatric Association, 1994.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (5th ed.)*, volume 21. American Psychiatric Publishing, 2013.

S. Amiri, A. Belloum, S. Klous, and L. Gommans. Compressive differentiallyprivate federated learning through universal vector quantization. In *AAAI Workshop on Privacy-Preserving Artificial Intelligence*, 2021.

S. Amiri, A. Belloum, E. Nalisnick, S. Klous, and L. Gommans. On the impact of non-iid data on the performance and fairness of differentially private federated learning. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pages 52–58. IEEE Computer Society, 2022.

V. Amrhein, S. Greenland, and B. McShane. Scientists rise up against statistical significance, 2019.

C. Andrade, S. S. Arumugham, and J. Thirthalli. Adverse effects of electroconvulsive therapy. *Psychiatric Clinics*, 39(3):513–530, 2016.

P. Arora, D. Boyne, J. J. Slater, A. Gupta, D. R. Brenner, and M. J. Druzdzel. Bayesian networks for risk prediction using real-world data: a tool for precision medicine. *Value in Health*, 22(4):439–445, 2019.

A. Bayes and G. Parker. How to choose an antidepressant medication. *Acta Psychiatrica Scandinavica*, 139(3):280–291, 2019.

C. Beard, A. J. Millner, M. J. Forgeard, E. I. Fried, K. J. Hsu, M. Treadway, C. V. Leonard, S. Kertz, and T. Björgvinsson. Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological medicine*, 46(16):3359–3369, 2016.

D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. Redefine statistical significance. *Nature human behaviour*, 2(1):6–10, 2018.

J. O. Berger, L. R. Pericchi, and J. A. Varshavsky. Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 307–321, 1998.

J. A. Berlin and R. M. Golub. Meta-analysis as evidence: building a better pyramid. *Jama*, 312(6):603–606, 2014.

D. Berner and V. Amrhein. Why and how we should join the shift from significance testing to estimation. *Journal of Evolutionary Biology*, 35(6):777–787, 2022.

D. de Beurs, C. Bockting, A. Kerkhof, F. Scheepers, R. O'Connor, B. Penninx, and I. van de Leemput. A network perspective on suicidal behavior: Understanding suicidality as a complex system. *Suicide Life Threat. Behav.*, 51(1):115–126, 2021. doi: 10.1111/sltb.12676.

O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:D267–D270, 2004.

D. Borsboom. A network theory of mental disorders. *World psychiatry*, 16(1): 5–13, 2017.

D. Borsboom and A. O. Cramer. Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, 9:91–121, 2013.

V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

J. E. Brazier, R. Harper, N. M. Jones, A. O'Cathain, K. J. Thomas, T. Usherwood, and L. Westlake. Validating the sf-36 health survey questionnaire: new outcome measure for primary care. *BMJ*, 305(6846):160–4, 1992. doi: 10.1136/bmj.305.6846.160.

G. Briganti, M. Scutari, and R. J. McNally. A tutorial on bayesian networks for psychopathology researchers. *Psychological methods*, 2022.

T. J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R. R. Coeytaux, G. Samsa, V. Hasselblad, J. W. Williams, M. D. Musty, et al. Effect of clinical decision-support systems: a systematic review. *Annals of internal medicine*, 157 (1):29–43, 2012.

J. Brunson and Q. Read. ggalluvial: Alluvial plots in 'ggplot2'. r package., 2020. URL `http://corybrunson.github.io/ggalluvial/`.

K. Bruynseels, F. Santoni de Sio, and J. Van den Hoven. Digital twins in health care: ethical implications of an emerging engineering paradigm. *Frontiers in genetics*, 9, 2018. doi: 10.3389/fgene.2018.00031.

P. B. Burns, R. J. Rohrich, and K. C. Chung. The levels of evidence and their role in evidence-based medicine. *Plastic and reconstructive surgery*, 128(1):305, 2011.

J. Busner and S. D. Targum. The clinical global impressions scale: applying a research tool in clinical practice. *Psychiatry (Edgmont)*, 4(7):28–37, 2007.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011. ISSN 1548-7660.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, UK, 2006.

A. Cipriani, T. A. Furukawa, G. Salanti, A. Chaimani, L. Z. Atkinson, Y. Ogawa, S. Leucht, H. G. Ruhe, E. H. Turner, J. P. T. Higgins, M. Egger, N. Takeshima, Y. Hayasaka, H. Imai, K. Shinohara, A. Tajika, J. P. A. Ioannidis, and J. R. Geddes. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet*, 391(10128):1357–1366, 2018. doi: 10.1016/s0140-6736(17)32802-7.

D. A. Ciraulo, J. Barnhill, and H. Boxenbaum. Pharmacokinetic interaction of disulfiram and antidepressants. *Am. J. Psychiatry*, 142(11):1373–4, 1985. doi: 10.1176/ajp.142.11.1373.

G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Annals of internal medicine*, 162(1):55–63, 2015.

P. Coorevits, M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, et al. Electronic health records: new opportunities for clinical research. *Journal of internal medicine*, 274(6): 547–560, 2013.

R. Cordier, T. Brown, L. Clemson, and J. Byles. Evaluating the longitudinal item and category stability of the sf-36 full and summary scales using rasch analysis. *Biomed Res Int*, 2018:1013453, 2018. doi: 10.1155/2018/1013453.

D. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66–68, 1967.

A. P. Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984.

C. E. Dean. Social inequality, scientific inequality, and the future of mental illness. *Philosophy, Ethics, and Humanities in Medicine*, 12(1):1–12, 2017.

T. M. Deist, F. Dankers, P. Ojha, M. Scott Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, Z. Zhang, E. Spezi, M. Button, J. Jan Nuyttens, R. Vernhout, J. van Soest, A. Jochems, R. Monshouwer, J. Bussink, G. Price, P. Lambin, and A. Dekker. Distributed learning on 20 000+ lung cancer patients - the personal health train. *Radiother Oncol*, 144: 189–200, 2020. doi: 10.1016/j.radonc.2019.11.019.

D. L. Demets and K. G. Lan. Interim analysis: the alpha spending function approach. *Statistics in medicine*, 13(13-14):1341–1352, 1994.

L. van Diermen, S. van den Ameele, A. M. Kamperman, B. C. Sabbe, T. Vermeulen, D. Schrijvers, and T. K. Birkenhäger. Prediction of electroconvulsive therapy response and remission in major depression: meta-analysis. *The British journal of psychiatry*, 212(2):71–80, 2018.

B. Duan, A. Ramdas, and L. Wasserman. Interactive rank testing by betting. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *PMLR*, pages 201–235, 2022.

Dutch National Healthcare Institute. Farmacotherapeutisch kompas, 2020. URL `https://www.farmacotherapeutischkompas.nl/`.

J. Eckhoff. *Helly, Radon, and Carathéodory type theorems, Handbook of Convex Geometry Part A*, pages 389–448. Elsevier, 1993.

S. Epskamp, A. O. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom. qgraph: Network visualizations of relationships in psychometric data. *Journal of statistical software*, 48:1–18, 2012.

N. J. Ermers, K. Hagoort, and F. E. Scheepers. The predictive validity of machine learning models in the classification and treatment of major depressive disorder: State of the art and future directions. *Frontiers in Psychiatry*, 11:472, 2020.

T. van Erven, P. Grünwald, and S. de Rooij. Catching up faster in bayesian model selection and model averaging. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

ESC Cardiovasc Risk Collaboration, SCORE2 working group, et al. Score2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in europe. *European Heart Journal*, 42(25):2439–2454, 2021.

B. van Es, L. C. Reteig, S. C. Tan, M. Schraagen, M. M. Hemker, S. R. Arends, M. A. Rios, and S. Haitjema. Negation detection in dutch clinical texts: an evaluation of rule-based and machine learning methods. *BMC bioinformatics*, 24(1):10, 2023.

B. de Finetti. *Theory of probability: A critical introductory treatment*, volume 6. John Wiley & Sons, 2017.

R. A. Fisher. *Statistical methods for research workers.* Oliver and Boyd, 1925.

W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57–57, 1992.

E. I. Fried, J. K. Flake, and D. J. Robinaugh. Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, 1(6):358–368, 2022.

P. Fusar-Poli, Z. Hijazi, D. Stahl, and E. W. Steyerberg. The science of prognosis in psychiatry: A review. *JAMA Psychiatry*, 75(12):1289–1297, 2018. doi: 10.1001/jamapsychiatry.2018.2530.

B. N. Gaynes, D. Warden, M. H. Trivedi, S. R. Wisniewski, M. Fava, and A. J. Rush. What did star*d teach us? results from a large-scale, practical, clinical trial for patients with depression. *Psychiatr Serv*, 60(11):1439–45, 2009. doi: 10.1176/ps.2009.60.11.1439.

S. A. Glied, B. D. Stein, T. G. McGuire, R. R. Beale, F. F. Duffy, S. Shugarman, and H. H. Goldman. Measuring performance in psychiatry: A call to action. *Psychiatr Serv*, 66(8):872–8, 2015. doi: 10.1176/appi.ps.201400393.

T. J. Gross, R. B. Araújo, F. A. C. Vale, M. Bessani, and C. D. Maciel. Dependence between cognitive impairment and metabolic syndrome applied to a brazilian elderly dataset. *Artificial intelligence in medicine*, 90:53–60, 2018.

P. Grünwald. *The Minimum Description Length Principle.* MIT Press, Cambridge, MA, 2007.

P. Grünwald. Beyond Neyman-Pearson. *arXiv preprint arXiv:2205.00901*, 2022.

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *accepted, pending minor revision, for publication in Journal of the Royal Statistical Society: Series B*, 2022a.

P. Grünwald, A. Henzi, and T. Lardy. Anytime valid tests of conditional independence under model-x. *arXiv preprint arXiv:2209.12637*, 2022b.

E. Gunel and J. Dickey. Bayes factors for independence in contingency tables. *Biometrika*, 61(3):545–557, 1974.

S. H. Hageman, A. J. McKay, P. Ueda, L. H. Gunn, T. Jernberg, E. Hagström, D. L. Bhatt, P. G. Steg, K. Läll, R. Mägi, et al. Estimation of recurrent atherosclerotic cardiovascular event risk in patients with established cardiovascular disease: the updated smart2 algorithm. *European Heart Journal*, 43(18): 1715–1727, 2022.

M. Hamilton. A rating scale for depression. *J Neurol Neurosurg Psychiatry*, 23: 56–62, 1960.

M. Hamilton. Development of a rating scale for primary depressive illness. *British journal of social and clinical psychology*, 6(4):278–296, 1967.

Y. Hao, P. Grünwald, T. Lardy, L. Long, and R. Adams. E-values for k-sample tests with exponential families. *arXiv preprint arXiv:2303.00471*, 2023.

A. U. Haq, A. F. Sitzmann, M. L. Goldman, D. F. Maixner, and B. J. Mickey. Response of depression to electroconvulsive therapy: a meta-analysis of clinical predictors. *The Journal of clinical psychiatry*, 76(10):18164, 2015.

B. J. Havaki-Kontaxaki, P. P. Ferentinos, V. P. Kontaxakis, K. G. Paplos, and C. R. Soldatos. Concurrent administration of clozapine and electroconvulsive therapy in clozapine-resistant schizophrenia. *Clinical Neuropharmacology*, 29 (1):52–56, 2006.

R. de Heide and P. D. Grünwald. Why optional stopping can be a problem for bayesians. *Psychonomic Bulletin & Review*, 28:795–812, 2021.

W. T. Heijnen, T. K. Birkenhäger, A. I. Wierdsma, and W. W. van den Broek. Antidepressant pharmacotherapy failure and response to subsequent electroconvulsive therapy: a meta-analysis. *Journal of clinical psychopharmacology*, 30(5): 616–619, 2010.

K. Hemming, M. Taljaard, J. E. McKenzie, R. Hooper, A. Copas, J. A. Thompson, M. Dixon-Woods, A. Aldcroft, A. Doussau, M. Grayling, et al. Reporting of stepped wedge cluster randomised trials: extension of the consort 2010 statement with explanation and elaboration. *BMJ*, 363, 2018.

A. Henzi and J. F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, 109(3):647–663, 2022.

M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine learning*, 32 (2):151–178, 1998.

M. A. Hernán, S. Hernández-Diaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spacy: Industrial-strength natural language processing in python, 2020. URL `https://doi.org/10.5281/zenodo.1212303`.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 2021.

T. Jamil, A. Ly, R. D. Morey, J. Love, M. Marsman, and E.-J. Wagenmakers. Default “Gunel and Dickey” Bayes factors for contingency tables. *Behavior Research Methods*, 49:638–652, 2017.

E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106 (4):620, 1957.

H. Jeffreys. *The theory of probability.* Oxford University Press, 1998.

Y. Jin, Y. Su, X.-H. Zhou, S. Huang, and A. D. N. Initiative. Heterogeneous multimodal biomarkers analysis for alzheimer's disease via bayesian network. *EURASIP Journal on Bioinformatics and Systems Biology*, 2016:1–8, 2016.

R. Johari, P. Koomen, L. Pekelis, and D. Walsh. Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 70(3):1806–1821, 2022.

L. K. John, G. Loewenstein, and D. Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532, 2012.

S. H. Jones, G. Thornicroft, M. Coffey, and G. Dunn. A brief mental health outcome scale-reliability and validity of the global assessment of functioning (gaf). *Br J Psychiatry*, 166(5):654–9, 1995. doi: 10.1192/bjp.166.5.654.

R. E. Kass and S. K. Vaidyanathan. Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):129–144, 1992.

J. A. Kassem, O. Valkering, A. Belloum, and P. Grosso. Epi framework: Approach for traffic redirection through containerised network functions. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 80–89. IEEE, 2021.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of a/b testing. In *Conference on Learning Theory*, pages 461–481. PMLR, 2014.

M. G. Kebede. Automating normative control for healthcare research. In *International Workshop on AI Approaches to the Complexity of Legal Systems, International Workshop on AI Approaches to the Complexity of Legal Systems, International Workshop on Explainable and Responsible AI and Law*, pages 62–72. Springer, 2021.

J. L. Kelly. A new interpretation of information rate. *The bell system technical journal*, 1956.

R. C. Kessler, W. T. Chiu, O. Demler, and E. E. Walters. Prevalence, severity, and comorbidity of 12-month dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):617–627, 2005.

K. H. Kho, M. F. van Vreeswijk, S. Simpson, and A. H. Zwinderman. A meta-analysis of electroconvulsive therapy efficacy in depression. *The journal of ECT*, 19(3):139–147, 2003.

O. J. Kirtley, K. van Mens, M. Hoogendoorn, N. Kapur, and D. de Beurs. Translating promise into practice: a review of machine learning in suicide research and prevention. *Lancet Psychiatry*, 9(3):243–252, 2022. doi: 10.1016/s2215-0366(21) 00254-6.

B. Klingenberg. A new and improved confidence interval for the mantel–haenszel risk difference. *Statistics in Medicine*, 33(17):2968–2983, 2014.

J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

W. M. Koolen and T. van Erven. Freezing and sleeping: Tracking experts that learn by evolving past posteriors. *arXiv preprint arXiv:1008.4654*, 2010.

W. M. Koolen and P. Grünwald. Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 141:69–82, 2022.

W. M. Koolen and S. de Rooij. Universal codes from switching strategies. *IEEE Transactions on Information Theory*, 59(11):7168–7185, 2013.

R. Koposov, S. Fossum, T. Frodl, Ø. Nytrø, B. Leventhal, A. Sourander, S. Quaglini, M. Molteni, M. de la Iglesia Vayá, H.-U. Prokosch, et al. Clinical decision support systems in child and adolescent psychiatry: a systematic review. *European Child & Adolescent Psychiatry*, 26:1309–1317, 2017.

Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, and R. J. B. Dobson. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artif Intell Med*, 117:102083, 2021.

R. Kroeze, D. C. van der Veen, M. N. Servaas, J. A. Bastiaansen, R. C. O. Voshaar, D. Borsboom, H. G. Ruhe, R. A. Schoevers, and H. Riese. Personalized feedback on symptom dynamics of psychopathology: A proof-of-principle study. *Journal for Person-Oriented Research*, 3(1):1, 2017.

H. M. Krumholz. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs*, 33(7):1163–1170, 2014.

S. Kundu. AI in medicine must be explainable. *Nature Medicine*, 27(8):1328–1328, 2021.

E. Kyrimi, K. Dube, N. Fenton, A. Fahmi, M. R. Neves, W. Marsh, and S. McLachlan. Bayesian networks in healthcare: What is preventing their adoption? *Artificial Intelligence in Medicine*, 116:102079, 2021.

T. L. Lai. On confidence sequences. *The Annals of Statistics*, 4(2):265–280, 1976.

T. A. Lang and D. G. Altman. Statistical analyses and methods in the published literature: The SAMPL guidelines. *Guidelines for reporting health research: A user's manual*, pages 264–274, 2014.

J. Lee, R. Henning, and M. Cherniack. Correction workers' burnout and outcomes: A bayesian network approach. *International journal of environmental research and public health*, 16(2):282, 2019.

W. Lee, J. Bindman, T. Ford, N. Glozier, P. Moran, R. Stewart, and M. Hotopf. Bias in psychiatric case–control studies: literature survey. *The British Journal of Psychiatry*, 190(3):204–209, 2007.

K. A. Leiknes, L. J.-v. Schweder, and B. Høie. Contemporary use and practice of electroconvulsive therapy worldwide. *Brain and behavior*, 2(3):283–344, 2012.

H. Leroux, A. Metke-Jimenez, and M. J. Lawley. Towards achieving semantic interoperability of clinical study data with FHIR. *Journal of biomedical semantics*, 8(1):1–14, 2017.

L. A. Levin. Uniform tests of randomness. *Soviet Mathematics Doklady*, 17(2): 337–340, 1976.

A. Levy, S. Taib, C. Arbus, P. Péran, A. Sauvaget, L. Schmitt, and A. Yrondi. Neuroimaging biomarkers at baseline predict electroconvulsive therapy overall clinical response in depression: a systematic review. *The journal of ECT*, 35(2): 77–83, 2019.

A. Lhéritier and F. Cazals. A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64(5):3361–3370, 2018.

J. Li. *Estimation of Mixture Models*. PhD thesis, Yale University, New Haven, CT, 1999.

J. Li and A. Barron. Mixture density estimation. In S. Solla, T. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 279–285, Cambridge, MA, 2000. MIT Press.

J. J. Liang, C.-H. Tsou, and M. V. Devarakonda. Ground truth creation for complex clinical nlp tasks–an iterative vetting approach and lessons learned. *AMIA Summits on Translational Science Proceedings*, 2017:203, 2017.

M. Lindon and A. Malek. Anytime-valid inference for multinomial count data. *Advances in Neural Information Processing Systems*, 35:2817–2831, 2022.

S. H. Lisanby. Electroconvulsive therapy for depression. *New England Journal of Medicine*, 357(19):1939–1945, 2007.

P. de Looff, M. L. Noordzij, M. Moerbeek, H. Nijman, R. Didden, and P. Embregts. Changes in heart rate and skin conductance in the 30 min preceding aggressive behavior. *Psychophysiology*, 56(10):e13420, 2019.

J. J. Luykx, D. Loef, B. Lin, L. van Diermen, J. O. Nuninga, E. van Exel, M. L. Oudega, D. Rhebergen, S. N. Schouws, P. van Eijndhoven, et al. Interrogating associations between polygenic liabilities and electroconvulsive therapy effectiveness. *Biological Psychiatry*, 91(6):531–539, 2022.

A. Ly, R. J. Turner, and J. ter Schure. R-package `safestats`, 2022. CRAN.

O.-A. Maillard. Mathematics of statistical sequential decision making, 2019. Thèse de Habilitation.

T. Manole and A. Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 2023.

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4): 719–748, 1959.

M. L. McHugh. The chi-square test of independence. *Biochemia medica*, 23(2): 143–149, 2013.

S. McLachlan, K. Dube, G. A. Hitman, N. E. Fenton, and E. Kyrimi. Bayesian networks in healthcare: Distribution by medical condition. *Artificial intelligence in medicine*, 107:101912, 2020.

R. McNally, P. Mair, B. Mugno, and B. Riemann. Co-morbid obsessive–compulsive disorder and depression: A bayesian network approach. *Psychological medicine*, 47(7):1204–1214, 2017.

J. Meiseberg and S. Moritz. Biases in diagnostic terminology: Clinicians choose different symptom labels depending on whether the same case is framed as depression or schizophrenia. *Schizophr Res*, 222:444–449, 2020. doi: 10.1016/j.schres.2020.03.050.

V. Menger. Psynlp, 2020. URL `https://github.com/vmenger/psynlp`.

V. Menger, F. Scheepers, and M. Spruit. Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences*, 8(6):981, 2018a.

V. Menger, F. Scheepers, L. M. van Wijk, and M. Spruit. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telematics and Informatics*, 35(4):727–736, 2018b.

V. J. Menger. *Knowledge Discovery in Clinical Psychiatry.* PhD thesis, Utrecht University, 2019.

S. A. Montgomery and M. Åsberg. A new depression scale designed to be sensitive to change. *The British journal of psychiatry*, 134(4):382–389, 1979.

L. B. Moreira and A. A. Namen. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. *Computer methods and programs in biomedicine*, 165:139–149, 2018.

J. Muglu, H. Rather, D. Arroyo-Manzano, S. Bhattacharya, I. Balchin, A. Khalil, B. Thilaganathan, K. S. Khan, J. Zamora, and S. Thangaratinam. Risks of stillbirth and neonatal death with advancing gestation at term: A systematic review and meta-analysis of cohort studies of 15 million pregnancies. *PLoS medicine*, 16(7):e1002838, 2019.

A. C. Naglich, A. Lin, S. Wakhlu, and B. H. Adinoff. Systematic review of combined pharmacotherapy for the treatment of alcohol use disorder in patients without comorbid conditions. *CNS Drugs*, 32(1):13–31, 2018. doi: 10.1007/s40263-017-0484-2.

Nederlandse Rijksoverheid. Wet hergebruik van overheidsinformatie, 2021. URL `https://wetten.overheid.nl/BWBR0036795/`.

Netherlands Federation of University Medical Centers. Guideline quality assurance of research involving human subjects, 2020.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

F. Orabona and K.-S. Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *arXiv preprint arXiv:2110.14099*, 2021.

M. Otto. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation–GDPR). In *International and European Labour Law*, pages 958–981, 2018.

S. Oxlad and M. Baldwin. Multiple case sampling of ect administration to 217 minors: Review and meta-analysis. *Journal of Mental Health*, 5(5):451–464, 1996.

L. Pace and A. Salvan. Likelihood, replicability and robbins' confidence sequences. *International Statistical Review*, 88(3):599–615, 2020.

M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

D. Pagnin, V. de Queiroz, S. Pini, and G. B. Cassano. Efficacy of ECT in depression: a meta-analytic review. *The journal of ECT*, 20(1):13–20, 2004.

T. Pandeva, T. Bakker, C. A. Naesseth, and P. Forré. E-valuating classifier two-sample tests. *arXiv preprint arxiv:2210.13027*, 2022.

G. Parmigiani, B. Barchielli, S. Casale, T. Mancini, and S. Ferracuti. The impact of machine learning in predicting risk of violence: A systematic review. *Frontiers in psychiatry*, 13, 2022.

S. van der Pas and P. Grünwald. Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection. *Statistica Sinica*, 28(1):229–255, 2018.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms.* The MIT Press, 2017.

D. Peterson. The replication crisis won't be solved with broad brushstrokes. *Nature*, 594(7862), 2021.

M. F. Pradier, T. H. McCoy Jr, M. Hughes, R. H. Perlis, and F. Doshi-Velez. Predicting treatment dropout after antidepressant initiation. *Translational psychiatry*, 10(1):1–8, 2020.

C. A. Prinsen, S. Vohra, M. R. Rose, S. King-Jones, S. Ishaque, Z. Bhaloo, D. Adams, and C. B. Terwee. Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. *Trials*, 15:247, 2014. doi: 10.1186/1745-6215-15-247.

J. Prudic, M. Olfson, S. C. Marcus, R. B. Fuller, and H. A. Sackeim. Effectiveness of electroconvulsive therapy in community settings. *Biological psychiatry*, 55(3): 301–312, 2004.

S. Qiu, W. Poon, M. Tang, and J. Tao. Construction of confidence intervals for the risk differences in stratified design with correlated bilateral data. *Journal of Biopharmaceutical Statistics*, 29(3):446–467, 2019.

R. Rabin and F. de Charro. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med*, 33(5):337–43, 2001. doi: 10.3109/07853890109002087.

A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *arXiv preprint arXiv:2210.01948*, 2022.

F. P. Ramsey. Truth and probability. In *The Foundations of Mathematics and other Logical Essays.* Routledge & Kegan Paul Ltd, 1931.

H. Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

R. Royall. *Statistical Evidence: A Likelihood Paradigm.* Chapman and Hall, 1997.

A. J. Rush, M. H. Trivedi, S. R. Wisniewski, A. A. Nierenberg, J. W. Stewart, D. Warden, G. Niederehe, M. E. Thase, P. W. Lavori, and B. D. Lebowitz. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. *American Journal of Psychiatry*, 163 (11):1905–1917, 2006.

Y. E. Rybak, K. S. P. Lai, R. Ramasubbu, F. Vila-Rodriguez, D. M. Blumberger, P. Chan, N. Delva, P. Giacobbe, C. Gosselin, S. H. Kennedy, H. Iskandar,

S. McInerney, P. Ravitz, V. Sharma, A. Zaretsky, and A. M. Burhan. Treatment-resistant major depressive disorder: Canadian expert consensus on definition and assessment. *Depress Anxiety*, 38(4):456–467, 2021. doi: 10.1002/da.23135.

L. Samalin, J.-B. Genty, L. Boyer, J. Lopez-Castroman, M. Abbar, and P.-M. Llorca. Shared decision-making: a systematic review focusing on mood disorders. *Current psychiatry reports*, 20(4):1–11, 2018. ISSN 1535-1645.

R. Sanfelici, D. B. Dwyer, L. A. Antonucci, and N. Koutsouleris. Individualized diagnostic and prognostic models for patients with psychosis risk syndromes: A meta-analytic view on the state of the art. *Biol Psychiatry*, 88(4):349–360, 2020. doi: 10.1016/j.biopsych.2020.02.009.

L. J. Savage. *The foundations of statistics*. John Wiley & Sons, Inc., 1954.

C. W. M. Schepper, R. J. Turner, L. Schaijk, and K. Hagoort. Bijwerkingen van ECT detecteren in klinische teksten, conference presentation at NVVP Voorjaarscongres 2022, 2022.

J. Schneider, M. Patterson, and X. F. Jimenez. Beyond depression: Other uses for tricyclic antidepressants. *Cleveland Clinic Journal of Medicine*, 86(12):807–814, 2019.

J. ter Schure. *ALL-IN Meta-Analysis*. PhD thesis, Leiden University, 2022.

J. ter Schure, M. F. Pérez-Ortiz, A. Ly, and P. Grunwald. The safe logrank test: Error control under continuous monitoring with unlimited horizon. *arXiv preprint arXiv:2011.06931*, 2020.

J. ter Schure, A. Ly, L. Belin, C. S. Benn, M. J. Bonten, J. D. Cirillo, J. A. Damen, I. Fronteira, K. D. Hendriks, A. P. Junqueira-Kipnis, et al. Bacillus calmette-guerin vaccine to reduce covid-19 infections and hospitalisations in healthcare workers: a living systematic review and prospective all-in meta-analysis of individual participant data from randomised controlled trials. *medRxiv*, pages 2022–12, 2022.

J. A. ter Schure and P. Grünwald. Accumulation bias in meta-analysis: the need to consider time in error control. *F1000Research*, 8, 2019.

M. Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35:1–22, 2010.

M. Scutari and K. Strimmer. Introduction to graphical modelling. *Handbook of Statistical Systems Biology*, 2011.

T. Seidenfeld. Why I am not an objective Bayesian; some reflections prompted by Rosenkrantz. *Theory and Decision*, 11(4):413–440, 1979.

J. Sevilla. Finding, scoring and explaining arguments in bayesian networks. *arXiv preprint arXiv:2112.00799*, 2021.

S. Shaer, G. Maman, and Y. Romano. Model-free sequential testing for conditional independence via testing by betting. *arXiv preprint arXiv:2210.00354*, 2022.

G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, pages 84–101, 2011.

G. Shafer et al. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021.

S. Shekhar and A. Ramdas. Nonparametric two-sample testing by betting. *arXiv preprint arXiv:2112.09162*, 2021.

D. Siegmund. *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 2013.

L. Simon, M. Blay, F. Galvao, and J. Brunelin. Using EEG to predict clinical response to electroconvulsive therapy in patients with major depression: a comprehensive review. *Frontiers in Psychiatry*, 12:643710, 2021.

K. A. Spackman, K. E. Campbell, and R. A. Côté. SNOMED RT: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association, 1997.

D. Stacey, F. Légaré, K. Lewis, M. J. Barry, C. L. Bennett, K. B. Eden, M. Holmes-Rovner, H. Llewellyn-Thomas, A. Lyddiatt, R. Thomson, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane database of systematic reviews*, 2017.

M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.

E. Steyerberg. *Clinical prediction models*. Springer New York, 2009.

A. Susaiyah, A. Härmä, E. Reiter, and M. Petković. Neural scoring of logical inferences from data using feedback. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(5):90–99, 2021.

The EPI Consortium. Epi: Enabling personalized interventions., 2019. URL `https://enablingpersonalizedinterventions.nl`. Last accessed 18 January 2023.

J. W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25, 1980.

R. J. Turner. Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test. Master's thesis, Leiden University, 2019.

R. J. Turner. PsyNLP, 2021. URL `https://github.com/rosanneturner/psynlp_outcome_measures`.

R. J. Turner. Netwerkanalyse van antidepressiva behandeltrajecten. In *NVVP Voorjaarscongres*, 2022.

R. J. Turner. safeSequentialTestingAISTATS2023, 2023. Code corresponding to AISTATS Paper, accessible at `https://github.com/rosanneturner/safeSequentialTestingAISTATS2023`.

R. J. Turner and P. D. Grünwald. Exact anytime-valid confidence intervals for contingency tables and beyond. *Statistics & Probability Letters*, page 109835, 2023.

R. J. Turner, A. Ly, and P. D. Grünwald. Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *arXiv preprint arxiv:2106.02693*, 2021.

R. J. Turner, F. Coenen, F. Roelofs, K. Hagoort, A. Härmä, P. D. Grünwald, F. P. Velders, and F. E. Scheepers. Information extraction from free text for aiding transdiagnostic psychiatry: constructing nlp pipelines tailored to clinicians' needs. *BMC Psychiatry*, 22(1):407, 2022. doi: 10.1186/s12888-022-04058-z.

J. Ville. Étude critique de la notion de collectif, 1939.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

E.-J. Wagenmakers and A. Ly. Bayesian Scepsis About SWEPIS: Quantifying the Evidence That Early Induction of Labour Prevents Perinatal Deaths. *PsyArXiv*, 2020. doi: 10.31234/osf.io/5ydpb.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

A. Wald. *Sequential Analysis*. Wiley, 1947.

R. Wang and A. Ramdas. False discovery rate control with e-values. *arXiv preprint arXiv:2009.02824*, 2020.

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

R. L. Wasserstein and N. A. Lazar. The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2020.

U. Wennerholm, S. Saltvedt, A. Wessberg, M. Alkmark, C. Bergh, S. B. Wendel, H. Fadl, M. Jonsson, L. Ladfors, V. Sengpiel, et al. Induction of labour at 41 weeks versus expectant management and induction of labour at 42 weeks (SWEdish Post-term Induction Study, swepis): multicentre, open label, randomised, superiority trial. *British Medical Journal*, 367, 2019.

H. A. Whiteford, A. J. Ferrari, L. Degenhardt, V. Feigin, and T. Vos. The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. *PLoS One*, 10(2):e0116820, 2015. doi: 10.1371/journal.pone.0116820.

P. Whiting, J. Savović, J. P. Higgins, D. M. Caldwell, B. C. Reeves, B. Shea, P. Davies, J. Kleijnen, R. Churchill, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. *Journal of clinical epidemiology*, 69: 225–234, 2016.

J. T. Wigman, J. van Os, E. Thiery, C. Derom, D. Collip, N. Jacobs, and M. Wichers. Psychiatric diagnosis revisited: towards a system of staging and profiling combining nomothetic and idiographic parameters of momentary mental states. *PLoS One*, 8(3):e59559, 2013. doi: 10.1371/journal.pone.0059559.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18.

J. B. Williams. Standardizing the hamilton depression rating scale: past, present, and future. *Eur Arch Psychiatry Clin Neurosci*, 251 Suppl 2:II6–12, 2001. doi: 10.1007/BF03035120.

World Health Organization. The World Health Organization Quality of Life assessment (WHOQOL): position paper from the World Health Organization. *Soc Sci Med*, 41(10):1403–9, 1995. doi: 10.1016/0277-9536(95)00112-k.

World Health Organization. Fact sheet depression 2022, 2022. URL `https://www.who.int/news-room/fact-sheets/detail/depression`.

A. G. Yip, K. J. Ressler, F. Rodriguez-Villa, S. H. Siddiqi, and S. J. Seiner. Treatment outcomes of electroconvulsive therapy for depressed patients with and without borderline personality disorder: a retrospective cohort study. *The Journal of Clinical Psychiatry*, 82(2):28439, 2021.

M. Zimmerman, M. Posternak, M. Friedman, N. Attiullah, S. Baymiller, R. Boland, S. Berlowitz, S. Rahman, K. Uy, and S. Singer. Which factors influence psychiatrists' selection of antidepressants? *Am. J. Psychiatry*, 161(7): 1285–9, 2004. doi: 10.1176/appi.ajp.161.7.1285.

Appendix with Supplementary Material

Supplementary material for chapter 2

Appendix S2.A contains detailed proofs. Appendix S2.B contains a detailed description of the numerical approach to calculating e-variables for restricted $\mathcal{H}_1$. Appendix S2.C contains a detailed description of Gunel-Dickey Bayes factors. Appendix S2.D contains optional stopping experiments. Appendix S2.E explains how to adapt the block group sizes n_a and n_b based on past data.

S2.A Proofs

The proofs below repeatedly use Theorem 1 of Grünwald et al. [2022a] and a direct corollary (called Corollary 2 by Grünwald et al. [2022a]), which we restate here for convenience, combined as a single statement. We use the notation adopted later in the paper: for $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ and, for W a distribution on Θ_0, we write $P_W = \int P_\theta dW(\theta)$.

Theorem (Theorem 1 of Grünwald et al. [2022a]) Let Y be a random variable taking values in a set $\mathcal{Y}$. Suppose Q is a probability distribution for Y with density q that is strictly positive on all of $\mathcal{Y}$ and let $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ be a set of distributions for Y where each P_θ has density p_θ. Let $\mathcal{W}_0$ be the set all distributions on Θ_0. Assume $\inf_{W_0 \in \mathcal{W}_0(\Theta_0)} D(Q \| P_{W_0}) < \infty$. Then (a) there exists a (potentially sub-) distribution P_0^* with density p_0^* such that

$$S^* := \frac{q(Y)}{p_0^*(Y)}$$

is an e-variable (p_0^* is called the *Reverse Information Projection (RIPr) of* q *onto* $\{p_W : W \in \mathcal{W}_0\}$ [Li, 1999, Li and Barron, 2000, Grünwald et al., 2022a]). Moreover, (b), S^* satisfies

$$\sup_{S \in \mathcal{E}(\Theta_0)} \mathbf{E}_{Y \sim Q}[\log S] = \mathbf{E}_{Y \sim Q}[\log S^*] = \inf_{W_0 \in \mathcal{W}_0(\Theta_0)} D(Q \| P_{W_0}) = D(Q \| P_0^*). \tag{A.1}$$

and is thus the Q-GRO e-variable for Y. If the minimum is achieved by some W_0^*, i.e. $D(Q\|P_0^*) = D(Q\|P_{W_0^*})$, then $P_0^* = P_{W_0^*}$. Moreover, (c), if there exists an e-variable S of the form $q(Y)/p_{W_0}(Y)$ for some $W_0 \in \mathcal{W}_0$ then W_0 must achieve the infimum in (A.1) and S must be essentially equal to S^* in the sense that for all $P \in \mathcal{H}_0 \cup \{Q\}$, $P(S^* = q(Y)/p_{W_0}(Y)) = 1$. Similarly (d), if there exists a $W_0^* \in \mathcal{W}_0$ that achieves the infimum in (A.1) then $S = q(Y)/p_{W_0^*}(Y)$ is an e-variable and S is again essentially equal to S^*.

S2.A.1 Proof of Propositions

Proof of Proposition 1 Below we state and prove a slight generalization of Proposition 1.

Proposition 4 (generalization)**.** Let $\mathcal{H}_1 = \{Q\}$ be a singleton and let $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ be such that for some distribution W on Θ_0, $D(Q\|P_W) < \infty$. For general $\theta \in \Theta_0$ and distributions W on Θ_0, define $S_{\theta,(j)} := q(Y_{(j)})/p_\theta(Y_{(j)})$ and $S_{W,(j)} = q(Y_{(j)})/p_W(Y_{(j)})$. We have:

1. Suppose there exists a distribution W on Θ_0 such that $S_{W,(1)}$ is an e-variable. Then $S_{W,(1)}$ is the Q-GRO e-variable for $Y_{(1)}$. In particular, if W puts mass 1 on a particular $\theta^\circ \in \Theta_0$, then $S_{W,(1)} = S_{\theta^\circ,(1)}$ is the Q-GRO e-variable.

2. If $\Theta_0 = \{\theta_0\}$ is simple then, with the prior W_0 putting mass 1 on θ_0, $S_{W_0,(1)} = S_{\theta_0,(1)}$ is an e-variable and hence, by the above, also the Q-GRO e-variable.

3. If, for some $\theta^\circ \in \Theta_0$, $S_{\theta^\circ,(1)}$ is an e-variable and we further assume that $Y_{(1)}, Y_{(2)}, \ldots$ are i.i.d. according to all distributions in $\mathcal{H}_0 \cup \mathcal{H}_1$, then $S^{(m)}_{\text{GRO}(Q)} = \prod_{j=1}^m S_{\theta^\circ,(j)}$; that is, the Q-GRO optimal (unconditional) e-variable for $Y^{(m)}$ is the product of the individual Q-GRO optimal e-variables.

Proof. Part 1 The theorem above, part (b), implies, with $Y = Y_{(1)}$, that some Q-GRO e-variable S^* for $Y_{(1)}$ exists. Part (c) then implies that we can take S^* to be equal to $S_{W,(1)}$. This implies the statement.

Part 2 is immediate.

Part 3 We assume that $S_{\theta^\circ,(1)}$ is an e–variable. Then the i.i.d. assumption implies that $S^{(m)}_{\theta^\circ} := \prod_{j=1}^m S_{\theta^\circ,(j)} = \prod q(Y_{(j)})/p_{\theta^\circ}(Y_{(j)})$ is also an e-variable. But [Grünwald et al., 2022a, Theorem 1], part (c) as stated above implies (by taking a distribution W putting mass 1 on θ) that for $\mathcal{H}_0$ for which data are i.i.d., for each $m \geq 1$, that if a $\theta \in \Theta_0$ exists such that $S^{(m)}_\theta$ is an e-variable, then $S^{(m)}_\theta$ must be the Q-GRO e-variable for $Y^{(m)}$. This proves the statement. □

Proof of Proposition 2 The formulae for $\breve{\theta}_a|Y^{(j-1)}$ and $\breve{\theta}_b|Y^{(j-1)}$ are standard expressions for the Bayes predictive distribution based on the given beta priors; we omit further details. As to the expression for $\breve{\theta}_0|Y^{(j-1)}$ in terms of $\kappa = n_b/n_a$: Straightforward rewriting gives, for general $\alpha_a, \alpha_b, \beta_a, \beta_b$:

$$\breve{\theta}_0|Y^{(j-1)} = \frac{1}{1+\kappa}\breve{\theta}_a|Y^{(j-1)} + \frac{\kappa}{1+\kappa}\breve{\theta}_b|Y^{(j-1)}. \tag{A.2}$$

If we plug in the expressions for $\breve{\theta}_a|Y^{(j-1)}, \breve{\theta}_b|Y^{(j-1)}$ and we instantiate to $\alpha_b = \kappa\alpha_a$, and $\beta_b = \kappa\beta_a$, this becomes

$$\begin{aligned}\breve{\theta}_0|Y^{(j-1)} &= \frac{1}{1+\kappa}\frac{U_a+\alpha_a}{n_a(j-1)+\alpha_a+\beta_a} + \frac{\kappa}{1+\kappa}\frac{U_b+\alpha_b}{\kappa(n_a(j-1)+\alpha_a+\beta_a)} \\ &= \frac{1}{1+\kappa}\frac{U_a+U_b+(1+\kappa)\alpha_a}{n_a(j-1)+\alpha_a+\beta_a} = \frac{U+(1+\kappa)\alpha_a}{n(j-1)+(1+\kappa)\alpha_a+(1+\kappa)\beta_a},\end{aligned}$$

which is what we had to prove.

S2.A.2 Proof of Theorem 1

We first restate Theorem 1 in its extended version that holds for $k \geq 2$ data streams. Let $\vec{n} = (n_1, \ldots, n_k), n = \sum_{g=1}^{k} n_g, \vec{\theta} = (\theta_a, \ldots, \theta_k) \in \Theta^k$ and $\vec{y}^n$ be as defined in the main text (3.3). We use '$\vec{Y}^n \sim P_{\theta^*}$' as an abbreviation for '$Y_1^{n_1} \sim P_{\theta_1^*}; \ldots; Y_k^{n_k} \sim P_{\theta_k^*}$'.

Theorem .1 (extended)**.** Let

$$s(\vec{y}^n; \vec{n}, \vec{\theta}^*) := \prod_{g=1}^{k} \frac{p_{\theta_g^*}(y_g^{n_g})}{\prod_{i=1}^{n_g}\left(\sum_{g'=1}^{k}\frac{n_{g'}}{n}p_{\theta_{g'}^*}(y_{i,g})\right)}.$$

The random variable $S_{[\vec{n},\vec{\theta}^*]} := s(\vec{Y}^n; \vec{n}, \vec{\theta}^*)$ is an e-variable, i.e. we have:

$$\sup_{\theta\in\Theta} \mathbf{E}_{V^n\sim P_\theta}\left[s(V^n; \vec{n}, \vec{\theta}^*)\right] \leq 1.$$

Moreover, if $\{P_\theta : \theta \in \Theta\}$ is a convex set of distributions, then $S_{[\vec{n},\vec{\theta}^*]}$ is the $(\vec{\theta}^*)$-GRO e-variable: for any non-negative function s' on $\mathcal{Y}^n$ satisfying $\sup_{\theta\in\Theta} \mathbf{E}_{V^n\sim P_\theta}[s'(V^n)] \leq 1$, we have:

$$\mathbf{E}_{\vec{Y}^n\sim P_{\theta^*}}[\log s(\vec{Y}^n; \vec{n}, \vec{\theta}^*)] \geq \mathbf{E}_{\vec{Y}^n\sim P_{\theta^*}}[\log s'(\vec{Y}^n)].$$

Proof of Theorem .1 The following fact plays a central role in the proof:

Fact For $g \in (1, ..., k)$, let $n_g \in \mathbf{N}, n := \sum_{g=1}^{k} n_g$ and let $u_g \in \mathbf{R}^+$. Suppose that $\sum_{g=1}^{k} n_g u_g \leq n$. Then $\prod_{g=1}^{k} u_g^{n_g} \leq 1$.

This result follows from the following standard generalization of Young's inequality to k numbers: for any k numbers $u_1, \ldots, u_k \in \mathbf{R}_0^+$ and any k nonnegative numbers $p_1, \ldots, p_k$ with $\sum_{g=1}^{k} p_g = 1$, we have $\prod_{g=1}^{k} u_g^{p_g} \leq \sum_{g=1}^{k} p_g u_g$. Applying this with $p_g = n_g/n$ to u_g and n_g as above, we get $\prod_{g=1}^{k} u_g^{n_g/n} \leq \sum_{g=1}^{k}(n_g u_g)/n \leq 1$, and the result follows by exponentiating to the power n.

Part 1 For $y \in \mathcal{Y}$, set set $p^\circ(y) := \sum_{g=1}^{k}(n_g/n)p_{\theta_g^*}(y)$ and $p^\circ(y^m) = \prod_{i=1}^{m} p^\circ(y_i)$.

For all $\theta \in \Theta$ we have:

$$\mathbf{E}_{V^n \sim P_\theta}\left[s(V^n; \vec{n}, \vec{\theta}^*)\right] = \prod_{g=1}^{k} \mathbf{E}_{Y_g^{n_g} \sim P_\theta}\left[\frac{p_{\theta_g^*}(Y_g^{n_g})}{p^\circ(Y_g^{n_g})}\right] = \prod_{g=1}^{k}\left(\mathbf{E}_{Y \sim P_\theta}\left[\frac{p_{\theta_g^*}(Y)}{p^\circ(Y)}\right]\right)^{n_g}. \tag{A.3}$$

We also have

$$\sum_{g=1}^{k} \frac{n_g}{n} \mathbf{E}_{Y \sim P_\theta}\left[\frac{p_{\theta_g^*}(Y)}{p^\circ(Y)}\right] = \mathbf{E}_{Y \sim P_\theta}\left[\sum_{g=1}^{k} \frac{n_g}{n} \cdot \frac{p_{\theta_g^*}(Y)}{\sum_{g'=1}^{k} \frac{n_{g'}}{n} p_{\theta_{g'}^*}(Y)}\right] = 1. \tag{A.4}$$

The result now follows by combining (A.3) with (A.4) using the Fact further above.

Part 2 By convexity of $\{P_\theta : \theta \in \Theta\}$, there exists $\theta^\circ \in \Theta$ such that $p_{\theta^\circ} = \sum_{g=1}^{k}(n_g/n)p_{\theta_g^*}$ and then the numerator in (A.4) can we rewritten as $p_{\theta^\circ}(\vec{y})$. The GRO-property is now an immediate consequence of Proposition 4, Part 1.

S2.B Numerical approach to calculating e-variables for restricted $\mathcal{H}_1$

In this subsection we describe how we propose to approximate the beta prior and posterior on the restricted $\mathcal{H}_1$ with parameter space $\Theta(\delta)$, as defined in (5.1). Note that we limit ourselves to $\delta > 0$ in this detailed description; for $\delta < 0$ one can apply an entirely equivalent approach, with an extra term in the reparameterization. We define

$$\zeta = \begin{cases} \delta \text{ if } d((\theta_a, \theta_b)) = \theta_b - \theta_a, \\ 0 \text{ if } d((\theta_a, \theta_b)) = \text{log-odds-ratio}(\theta_a, \theta_b), \end{cases}$$

such that we have $\theta_a \in (0, 1-\zeta)$ and in both cases, θ_b is completely determined by θ_a: $\theta_b = d^{-1}(\delta; \theta_a)$. Hence, our density estimation problem now becomes one-dimensional, which enables us to put a discretized prior on the restricted parameter space.

First, we discretize the parameter space Θ_a to a grid (a vector) with precision $K, K \in (0, 1-\zeta)$ and $1/K \in \mathbb{N}^+$: $\bar{\boldsymbol{\theta}}_{\boldsymbol{a}} = (K, 2K, 3K, \ldots, 1-\zeta)$. Then, we reparameterize $\theta_a = (1-\zeta)\rho$, with $\rho \in (0,1)$. Then, we have $\bar{\boldsymbol{\rho}} = (K/(1-\zeta), 2K/(1-\zeta), \ldots, 1)$. For the discretized grid $\bar{\boldsymbol{\rho}}$, we compute the prior $W = \text{Beta}(\alpha, \beta)$ densities and normalize them, which also gives us the discretized densities for each $\theta_a^i \in \bar{\boldsymbol{\theta}}_{\boldsymbol{a}}$ (with $i \in (1, 2, \ldots, 1/K)$):

$$\pi_{\alpha,\beta,\zeta}(\theta_a^i) = \frac{\text{Beta}(\frac{\theta_a^i}{1-\zeta}; \alpha, \beta)}{\sum_{k=1}^{\frac{1}{K}} \text{Beta}(\frac{\theta_a^k}{1-\zeta}; \alpha, \beta)}.$$

For all elements of $\bar{\boldsymbol{\theta}}_{\boldsymbol{a}}$, the corresponding θ_b is retrieved and the likelihood of incoming data points $p_{\theta_a, \theta_b}(Y^{(j-1)})$ is calculated. We can then estimate the posterior density of $\theta_a^i \in \bar{\boldsymbol{\theta}}_{\boldsymbol{a}}$:

$$p(\theta_a^i | Y^{(j-1)}) = \frac{\pi_{\alpha,\beta,\zeta}(\theta_a^i) p_{\theta_a^i, \theta_b^i}(Y^{(j-1)})}{\sum_{k=1}^{\frac{1}{K}} \pi_{\alpha,\beta,\zeta}(\theta_a^k) p_{\theta_a^k, \theta_b^k}(Y^{(j-1)})}.$$

We can then estimate $\breve{\theta}_a | Y^{(j-1)} = \mathbf{E}_{\theta_a \sim W | Y^{(j-1)}}[\theta_a]$ as $\sum_{i=1}^{\frac{1}{K}} p(\theta_a^i | Y^{(j-1)}) \theta_a^i$, and $\breve{\theta}_b | Y^{(j-1)} = d^{-1}(\delta; \theta_a | Y^{(j-1)})$.

S2.C The Gunel-Dickey Bayes Factors do not give rise to e–variables

Sampling	Fixed	Bayes factor (10) for 2x2 table
Poisson	none	$\frac{8(n+1)(n_1+1)}{(n+4)(n+2)} \left[\frac{n_{a1}! n_{b1}! n_{a0}! n_{b0}! n!}{(n_1+1)! n_0! n_a! n_b!} \right]$
Joint multinomial	n	$\frac{6(n+1)(n_1+1)}{(n+3)(n+2)} \left[\frac{n_{a1}! n_{b1}! n_{a0}! n_{b0}! n!}{(n_1+1)! n_0! n_a! n_b!} \right]$
Independent multinomial	n_a, n_b	$\binom{n}{n_1} / (\binom{n_a}{n_{a1}} \binom{n_b}{n_{b1}}) \frac{(n+1)}{(n_a+1)(n_b+1)}$
Hypergeometric	n_a, n_b, n_1	$\frac{n_{a1}! n_{b1}! n_{a0}! n_{b0}! n!}{\prod_{i \in \{a,b,0,1\}} (n_i + \mathbb{I} n_i = min(n_a, n_b, n_0, n_1))!}$

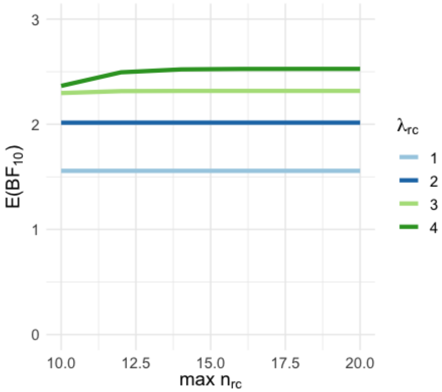
Table S2.1: Overview of (objective) Bayes factors for contingency table testing provided by Gunel and Dickey [1974] and Jamil et al. [2017].

We will not consider the hypergeometric and joint multinomial scenarios for this paper, where the number of successes n_1 is fixed, as they do not match the block-wise data design in this paper. The Bayes factor for the Poisson sampling scheme is not an e-variable, as the expectation under the null hypothesis with Poisson distributions on individual cell counts exceeds 1 for rates $\lambda \geq 1$:

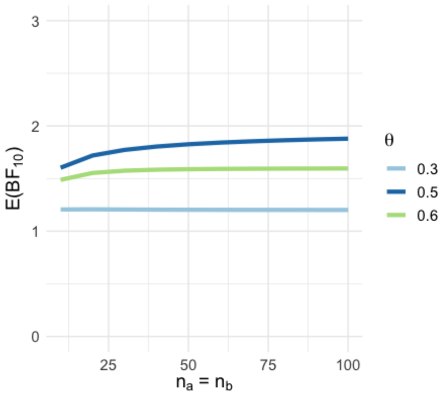
$$\mathbb{E}_{n_{rc} \sim \text{Poisson}(\lambda_{rc})} \left[BF_{10}(N_{a1}, N_{b1}, N_{a0}, N_{b0}) \right] =$$
$$\sum_{n_{a1}=0}^{\infty} \cdots \sum_{n_{b0}=0}^{\infty} \pi_{\lambda_{a1}}(n_{a1}) \ldots \pi_{\lambda_{b0}}(n_{b0}) BF_{10}(n_{a1}, n_{b1}, n_{a0}, n_{b0}) =$$
$$\frac{8}{\exp(\lambda_{a1} + \ldots + \lambda_{b0})} \sum_{n_{a1}=0}^{\infty} \cdots \sum_{n_{b0}=0}^{\infty} \lambda_{a1}^{n_{a1}} \ldots \lambda_{b0}^{n_{b0}} \frac{(n+1)(n_1+1)}{(n+4)(n+2)} \frac{n!}{(n_1+1)! n_0! n_a! n_b!},$$

as illustrated numerically in Figure S2.1 for increasing limits for the sums $\sum_{n_{rc}=1}^{\max n_{rc}}$.

For the independent multinomial sampling scheme, let, without loss of gener-



(a) The Gunel-Dickey Bayes factor for the Poisson sampling scheme is not an e-variable: $\sum_{n_{a1}=0}^{\max n_{rc}} \ldots \sum_{n_{b0}=0}^{\max n_{rc}} \pi_{\lambda_{a1}}(n_{a1}) \ldots \pi_{\lambda_{b0}}(n_{b0})$ $BF_{10}(n_{a1}, n_{b1}, n_{a0}, n_{b0})$ for various $\max n_{rc}$ and λ_{rc}.



(b) The Gunel-Dickey Bayes factor for the independent multinominal sampling scheme is not an e-variable: $\mathbb{E}_{N_{a1}, N_{b1} \sim \text{Binomial}(\theta)} [BF_{10}(N_{a1}, N_{b1} | n_a, n_b)]$ for various choices of θ and n_g.

Figure S2.1: GD

ality, $n_a < n_b$. We get, with $n_0 = n - n_1$,

$$\mathbb{E}_{N_{a1}, N_{b1} \sim \text{Binomial}(\theta)} \left[BF_{10}(N_{a1}, N_{b1} | n_a, n_b) \right] =$$
$$\sum_{n_{a1}=0}^{n_a} \sum_{n_{b1}=0}^{n_b} \binom{n_a}{n_{a1}} \binom{n_b}{n_{b1}} \theta^{n_1} (1-\theta)^{n_0} \frac{\binom{n}{n_1}}{\binom{n_a}{n_{a1}}\binom{n_b}{n_{b1}}} \frac{(n+1)}{(n_a+1)(n_b+1)} =$$
$$\frac{(n+1)}{(n_a+1)(n_b+1)} \sum_{n_{a1}=0}^{n_a} \sum_{n_{b1}=0}^{n_b} \binom{n}{n_1} \theta^{n_1} (1-\theta)^{n_0}$$

Numerical simulations show that, for a range of choices for n, n_a and θ this exceeds 1; see Figure S2.1.

S2.D Type-I error guarantee under optional stopping

Type-I Error In Figure S2.2 type-I error rates of several e-variables and Fisher's exact test estimated through a simulation experiment are depicted. 2000 samples of length 1000 were drawn according to a Bernoulli(0.1) distribution to represent 1000 data streams in two groups. After each complete block $m \in \{1, \ldots, 1000\}$ an e-value or p-value was calculated and the proportion of rejected experiments up until m with each test type was recorded. As the stream lengths increase, the type-I error rate under (incorrectly applied) optional stopping with Fisher's exact test increases quickly. The type-I error rate of the e-variables remains bounded.

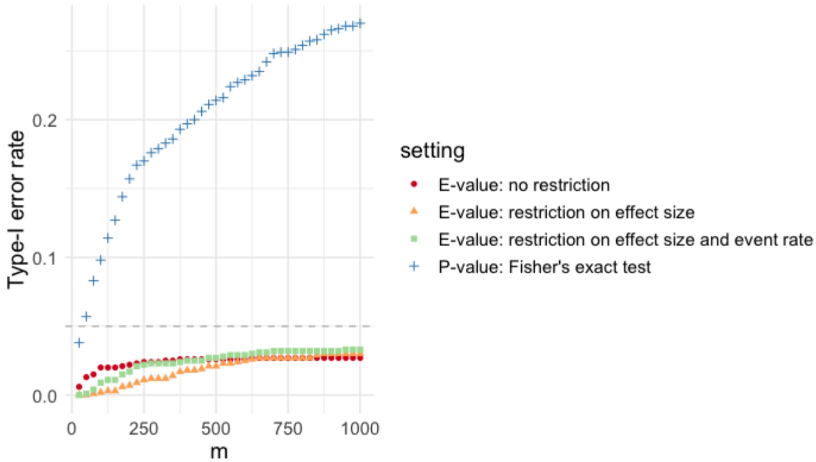


Figure S2.2: Type-I error rates for various e-variables and Fisher's exact test under optional stopping estimated with 1000 simulations of two Bernoulli(0.1) data streams of length 1000, with $n_a = n_b = 1$. Significance level $\alpha = 0.05$ was used (grey dashed line). For the safe tests, beta prior parameter values used were $\gamma = \alpha_a = \beta_a = \alpha_b = \beta_b = 1/2$ ($\gamma = 0.18$ gave comparable results). For the e-variables with restrictions on $\mathcal{H}_1$, we used $\delta = 0.05$ and $\theta_a = 0.1$.

S2.E Adjusting n_a and n_b based on past data

To see how to choose n_a and n_b for subsequent blocks based on past data, we first need to formalize the fact that data in different streams may arrive asynchronously. Thus, let $t = 1, 2, \ldots$ represent global ('calendar') time, and introduce corresponding random variables V_t and G_t: at each t, we obtain an outcome V_t in $\mathcal{Y}$ in group $G_t \in \{a, b\}$. We make no assumptions about the relative ordering of outcomes from the two groups. At time t, we have that t_a, the number of a's that are observed so far, and t_b, the number of b's observed so far, satisfy $t_a + t_b = t$, but subject to this constraint we allow them coming in any order. We now introduce a function $f : \bigcup_{t \geq 0} \mathcal{Y}^t \times \{0,1\}^t \rightarrow \{\text{STOP-BLOCK}, \text{CONTINUE}\}$ that, at each point in time t, decides whether the current block should end ($f(V^t, G^t) = $ STOP-BLOCK) or not ($f(V^t, G^t) = $ CONTINUE). As long as the value of this function does not depend on the actual outcomes V_t observed after the last block that was completed, all requirements for having a test martingale and thus for safe optional stopping are met. For example, suppose that on data $V_1, G_1, V_2, G_2, \ldots, V_t, G_t$ observed so-far, f has output STOP-BLOCK at m occasions, the last time at $t' = t - k$ for some $k > 0$. Then $f(t)$ is allowed to depend on $Y^{(m)}$ and G^t, but for any fixed $Y^{(m)} = y^{(m)}, G^t = g^t$, for all $y^k, y'^k \in \mathcal{Y}^k$, we must have $f((y^{(m)}, y^k), g^t) = f((y^{(m)}, y'^k), g^t)$.

Supplementary material for chapter 3

Appendix section S3.A contains proofs and section S3.B contains extended simulation results.

S3.A Proofs

Both proofs below use Theorem 1 of Grünwald et al. [2022a] and a direct corollary (called Corollary 2 by Grünwald et al. [2022a]), which we re-state here, for convenience, combined as a single statement. Recall that we use notation $P_W := \int P_{\vec{\theta}} dW(\vec{\theta})$.

Theorem (Theorem 1 of Grünwald et al. [2022a]) Let Y be a random variable taking values in a set $\mathcal{Y}$. Suppose Q is a probability distribution for Y with density q that is strictly positive on all of $\mathcal{Y}$ and let $\mathcal{H}_0 = \{P_{\vec{\theta}} : \vec{\theta} \in \vec{\Theta}_0\}$ be a set of distributions for Y where each $P_{\vec{\theta}}$ has density $p_{\vec{\theta}}$. Let $\mathcal{W}_0$ be the set of all distributions on $\vec{\Theta}_0$. Assume $\inf_{W_0 \in \mathcal{W}_0(\vec{\Theta}_0)} D(Q \| P_{W_0}) < \infty$. Then (a) there exists a (potentially sub-) distribution P_0^* with density p_0^* such that

$$S^* := \frac{q(Y)}{p_0^*(Y)}$$

is an e-variable (p_0^* is called the *Reverse Information Projection (RIPr) of q onto $\{p_W : W \in \mathcal{W}_0\}$*). Moreover, (b), S^* satisfies

$$\sup_{S \in \mathcal{E}(\vec{\Theta}_0)} \mathbf{E}_{Y \sim Q}[\log S] = \mathbf{E}_{Y \sim Q}[\log S^*] = \inf_{W_0 \in \mathcal{W}_0(\vec{\Theta}_0)} D(Q \| P_{W_0}) = D(Q \| P_0^*). \tag{A.5}$$

(where $\mathcal{E}(\vec{\Theta}_0)$ is the set of all e-variables relative to null hypothesis $\mathcal{H}_0$) and S^* is thus the Q-GRO e-variable for Y. If the minimum is achieved by some W_0^*, i.e. $D(Q\|P_0^*) = D(Q\|P_{W_0^*})$, then $P_0^* = P_{W_0^*}$. Moreover, (c), if there exists an e-variable S of the form $q(Y)/p_{W_0}(Y)$ for some $W_0 \in \mathcal{W}_0$ then W_0 must achieve the infimum in (A.5) and S must be essentially equal to S^* in the sense that for all $P \in \mathcal{H}_0 \cup \{Q\}$, $P(S^* = q(Y)/p_{W_0}(Y)) = 1$. Similarly (d), if there exists a $W_0^* \in \mathcal{W}_0$ that achieves the infimum in (A.5) then $S = q(Y)/p_{W_0^*}(Y)$ is an e-variable and S is again essentially equal to S^*.

Proof of Theorem 3.1 *Part 1* The real idea behind the proof is the formulation of the modified testing problem in which only a single outcome per block is observed. This we already did in the main text. Linking the two is simply the last, very simple step, with analogies to the proof of Part 1 of Theorem 1 in Turner et al. [2021].

Let $n_a, n_b \in \mathbf{N}, n := n_a + n_b$ and let $u, v \in \mathbf{R}^+$. Suppose that $n_a u + n_b v \leq n$. Then $u^{n_a} v^{n_b} \leq 1$, which follows immediately from applying Young's inequality to

$u^{n_a/n}, v^{n_b/n}$ but can also be derived directly by writing v as function of u and differentiating $\log(u^{n_a} v^{n_b})$ to u.

Further, by independence, for $(\theta_a, \theta_b) \in \vec{\Theta}_0$,

$$
\begin{aligned}
&\mathbf{E}_{Y_a^{n_a} \sim P_{\theta_a}, Y_b^{n_b} \sim P_{\theta_b}} [s'(Y_a^{n_a}, Y_b^{n_b})] = \\
&\mathbf{E}_{Y_a^{n_a} \sim P_{\theta_a}} \left[\frac{p_{\theta_a^*}(Y_a^{n_a})}{p^\circ(Y_a^{n_a}|a)} \right] \cdot \mathbf{E}_{Y_b^{n_b} \sim P_{\theta_b}} \left[\frac{p_{\theta_b^*}(Y_b^{n_b})}{p^\circ(Y_b^{n_b}|b)} \right] = \\
&\left(\mathbf{E}_{Y \sim P_{\theta_a}} \left[\frac{p_{\theta_a^*}(Y)}{p^\circ(Y|a)} \right] \right)^{n_a} \cdot \left(\mathbf{E}_{Y \sim P_{\theta_b}} \left[\frac{p_{\theta_b^*}(Y)}{p^\circ(Y|b)} \right] \right)^{n_b} = \\
&\left(\mathbf{E}_{Y \sim P'_{\vec{\theta}}|a} \left[\frac{p'_{\vec{\theta}^*}(Y|a)}{p^\circ(Y|a)} \right] \right)^{n_a} \cdot \left(\mathbf{E}_{Y \sim P'_{\vec{\theta}|b}} \left[\frac{p'_{\vec{\theta}^*}(Y|b)}{p^\circ(Y|b)} \right] \right)^{n_b} .
\end{aligned} \tag{A.6}
$$

Combining the two facts stated above, (3.6) implies that the latter quantity is bounded by 1.

Part 2 By lower-semicontinuity of the KL divergence in its second argument (Posner's theorem, used as in Grünwald et al. [2022a]) the infimum in (3.4) is achieved by some prior distribution W° so that by Theorem 1 of Grünwald et al. [2022a] (part (b) in the formulation above), $p^\circ(\cdot \mid \cdot) = p'_{W^\circ}(\cdot \mid \cdot)$ and hence also $P^\circ(G, Y) = P'_{W^\circ}(G, Y)$. By convexity of $\mathcal{H}'_0$ and finiteness of the support of $P'_{\vec{\theta}}(G, Y)$, there must be some $\vec{\theta}$ such that $P'_{W^\circ}(G, Y) = P_{\vec{\theta}}(G, Y)$ and hence also $p'_{W^\circ}(\cdot \mid \cdot) = p'_{\vec{\theta}}(\cdot \mid \cdot)$, which shows (a). This means that we have now created an e-variable for the original problem which can be written as $p_{\theta_a^*, \theta_b^*}/p_{W_0}$ with p_{W_0} a prior distribution on $\vec{\theta}_0$ (namely, the one that puts mass 1 on $\vec{\theta}$). (b) is then an immediate consequence of Theorem 1 of Grünwald et al. [2022a] (part (c) in the formulation above). (note that we *cannot* draw this conclusion if $\mathcal{H}'_0$ is not convex; for then the distribution p'_{W° may not correspond to the distribution p_{W° in the original problem — this correspondence is only guaranteed if p'_{W° coincides with some $p'_{\vec{\theta}}$.

Proof of Theorem 3.2 Recall that we assume that $\vec{\Theta}_0$ is convex and compact. We set $\text{KL}'(\theta_a, \theta_b) := D(P'_{\theta_a^*, \theta_b^*} \| P'_{\theta_a, \theta_b})$ where D is the KL divergence as in (3.5), i.e. for the modified setting in which P'_{θ_a, θ_b} is a distribution on a single outcome, as discussed before Theorem 3.1. For the 2×2 model this KL divergence can be written explicitly as

$$
\begin{aligned}
D(P'_{\theta_a^*, \theta_b^*} \| P'_{\theta_a, \theta_b}) &= \mathbf{E}_{G \sim Q'} \mathbf{E}_{Y \sim P'_{\vec{\theta}^*}|G} \left[\log \frac{p'_{\vec{\theta}^*}(Y|G)}{p'_{\vec{\theta}}(Y|G)} \right] \qquad \text{(A.7)} \\
&= \frac{n_a}{n} \mathbf{E}_{Y \sim P'_{\theta_a^*}} \left[\log \frac{p_{\theta_a^*}(Y)}{p_{\theta_a}(Y)} \right] + \frac{n_b}{n} \mathbf{E}_{Y \sim P'_{\theta_b^*}} \left[\log \frac{p_{\theta_b^*}(Y)}{p_{\theta_b}(Y)} \right] \\
&= \frac{n_a}{n} \sum_{y_a \in \{0,1\}} p_{\theta_a^*}(y_a) \log \frac{p_{\theta_a^*}(y_a)}{p_{\theta_a}(y_a)} + \frac{n_b}{n} \sum_{y_b \in \{0,1\}} p_{\theta_b^*}(y_b) \log \frac{p_{\theta_b^*}(y_b)}{p_{\theta_b}(y_b)}.
\end{aligned}
$$

From (3.8) we now see that $n\text{KL}'(\theta_a, \theta_b) = \text{KL}(\theta_a, \theta_b)$. We will prove the theorem with KL replaced by KL$'$ and $\mathcal{H}_0$ by $\mathcal{H}'_0$; since the two KL's agree up to a constant factor of n, all results transfer to the KL mentioned in the theorem statement.

Since $\vec{\Theta}_0$ is compact in the Euclidean topology and all distributions in $\mathcal{H}'_0$ can be represented as 2-dimensional vectors, i.e. they have common and finite support, we must have that $\mathcal{H}_0$ is compact in the weak topology so we can use the lower-semicontinuity of KL divergence in its second argument (Posner's theorem) as in [Grünwald et al., 2022a] to give us that the minimum KL divergence $\min \text{KL}'(\theta_a, \theta_b)$ is achieved by some $(\theta_a^\circ, \theta_b^\circ)$. Since KL divergence is strictly convex in its second argument and $\mathcal{H}'_0$ is convex (this is the place where we need to use KL$'$ rather than KL: $\mathcal{H}_0$ may *not* be convex!), the minimum must be achieved uniquely. Since KL divergence $\text{KL}'(\theta_a, \theta_b)$ is nonnegative and 0 only if $(\theta_a, \theta_b) = (\theta_a^*, \theta_b^*)$, it follows that $(\theta_a^\circ, \theta_b^\circ) = (\theta_a^*, \theta_b^*)$ if $\min \text{KL}(\theta_a, \theta_b) = 0$. Otherwise, since we assume (θ_a^*, θ_b^*) to be in the interior of $[0, 1]^2$, $\text{KL}(\theta_a, \theta_b) = \infty$ iff (θ_a, θ_b) lies on the boundary of $[0, 1]^2$. Thus, $(\theta_a^\circ, \theta_b^\circ)$ must lie in the interior of $[0, 1]^2$ as well. $(\theta_a^\circ, \theta_b^\circ)$ cannot lie in the interior of $\vec{\Theta}_0$ though: for any point (θ_a, θ_b) in the interior of $\vec{\Theta}_0$ we can draw a line segment between this point and (θ_a^*, θ_b^*). Differentiation along that line gives that $\text{KL}'(\theta_a, \theta_b)$ monotonically decreases as we move towards (θ_a^*, θ_b^*), so the minimum within the closed set $\vec{\Theta}_0$ must lie on its boundary.

It remains to show that (3.9) is the (θ_a^*, θ_b^*)-GRO e-variable relative to $\mathcal{H}_0$. To see this, note that, by convexity of $\mathcal{H}'_0$, from Theorem 3.1, we must have that the GRO e-variable for this original problem is of the form

$$\frac{p_{\theta_a^*}(y_a^{n_a}) p_{\theta_b^*}(y_b^{n_b})}{p_{\theta_a^+}(y_a^{n_a}) p_{\theta_b^+}(y_b^{n_b})}$$

for some (θ_a^+, θ_b^+). The result then follows again by Theorem 1 of Grünwald et al. [2022a] (part (c) in the formulation above): this shows that the distribution W_0 that puts mass 1 on (θ_a^+, θ_b^+) minimizes, among all distributions W on $\vec{\Theta}_0$, $D(P_{\theta_a^*, \theta_b^*} \| P_W)$. Since the set of such distributions includes all distributions that put mass 1 on *some* $(\theta_a, \theta_b) \in \vec{\Theta}_0$, we must have that $(\theta_a^+, \theta_b^+) = (\theta_a^\circ, \theta_b^\circ)$.

S3.B Extended simulation results

Numerical example We here give a small numerical example to illustrate the construction of our confidence sequences. For this example, we will look in detail at the data used to generate the second row of Figure 3.2a, the second panel, where we have observed 500 data blocks, with 27 "successes" ($y = 1$) in group a, and 136 "successes" in group b. To estimate δ_{L} and δ_{R}, $S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0]}$ as in (7.14) was calculated for that specific data stream, for a grid of possible δ, each defining one $\vec{\Theta}_0$; here, a grid with size 100 and a precision of 0.02 on $[-1, 1]$ was applied. The prior W_1 for the posterior mean was chosen as a Beta prior with $\alpha = \beta = 0.18$ according to Turner et al. [2021]. The area corresponding to values of δ for which $S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0]} < \frac{1}{0.05}$ after block $m = 500$ represents the confidence interval. For example, for the lower bound, δ_{L}, the smallest value of δ that did not lead to rejection was 0.15, with a corresponding e-value of 2.23. The e-value corresponding to $\delta = 0.13$ was 24.17, hence this risk difference was excluded from the confidence interval.

Running intersection In Figure S3.1, confidence sequence width is compared with and without applying the running intersection.

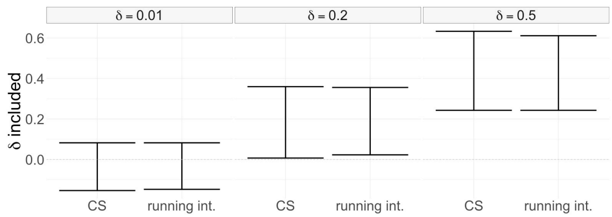


Figure S3.1: Confidence sequence with and without running intersection, for data generated under $P_{\theta_a,\theta_a+\delta}$ with $\theta_a = 0.05$, for a data stream of length 100. The significance threshold was set to 0.05. The design was balanced, with data block sizes $n_a = 1$ and $n_b = 1$.

Supplementary material for chapter 4

The following contains Section 1, *examples of theme and change phrases used for filtering sentences in the NLP pipeline*, of the supplementary material for Chapter 4 in this thesis. The other sections of the supplementary material can be found online in the publication corresponding to this chapter in *BMC Psychiatry* [Turner et al., 2022].

Table S4.1: Examples from the lists used for rule-based filtering of the four themes and change phrases

Category	**Dutch**	**Translation to English**	**Sentiment score**
Symptom reduction	Angstiger	More anxious	-1
	Angstigheid	Anxiety	-1
	Agressie	Aggression	-1
	Agresie	Aggression (misspelled)	-1
	Somber	Sad	-1
	Somer	Sad (misspelled)	-1
	Rotgevoel	Bad feeling	-1
	Doelloosheid	Aimlessness	-1
Social functioning	Zelfstandig	Independent	1
	Zelfstandige	Independent (conjugation)	1
	Zelfstandigheid	Independence	1
	Resocialiseren	Resocialize	1
	Participeert	Participates	1
	Vriendinnen	Girlfriends	1
	Vriendschappen	Friendships	1
	Verantwoordelijkheid	Responsibility	1
General well-being	Welbevinden	Well-being	1
	Welzijn	Well-being (synonym)	1
	Hoop	Hope	1
	Zingeving	Meaning	1
	Zinvol	Meaningful	1
	Zelfwaardering	Self-esteem	1
	Eigenwaarde	Self-esteem (synonym)	1
	Zelfvertrouwen	Self-confidence	1
	Zelfvetouwen	Self-confidence (misspelled)	1
Patient experience	Voelde	Felt	1
	Nez	In their own words (abbreviated)	1
	Voelt	Feels	1

Table with examples, continued

Category	**Dutch**	**Translation to English**	**Sentiment score**
	Uitte	Expressed	1
	Verwoorde	Articulated	1
	Constateert	Noted	1
	Merkt	Notes	1
	Mekrt	Notes (misspelled)	1
Change indicator	Afnam	Decreased	-1
	Afname	Decrease	-1
	Afgenomen	Decreased (conjugation)	-1
	Toenemende	Increasing	1
	Toenemde	Increasing (misspelled)	1
	Verbeter	Improve	1
	Verminder	Reduce	-1
	Vermindern	Reduce (misspelled)	-1

Supplementary material for chapter 5

Table S5.1: Overview of antidepressant prescription groups and specific antidepressants present in the data

Group	Antidepressant
MAOI	Tranylcypromine
	Moclobemide
	Phenelzine
nSSRI	Trazodone
	Duloxetine
	Venlafaxine
Other	Bupropion
	Vortioxetine
	Agomelatine
	Hyperici herba
SSRI	Sertraline
	Citalopram
	Escitalopram
	Fluoxetine
	Paroxetine
	Fluvoxamine
TetraCA	Mirtazapine
	Mianserine
TriCA	Nortriptyline
	Amitriptyline
	Clomipramine
	Imipramine
	Doxepine
	Maprotiline
	Dosulepine

Table S5.2: Overview of therapeutic dose range for selection of antidepressant treatment trajectories

antidepressant	Minimal dose	Maximal dose
tranylcypromine	10	60
phenelzine	8	120
moclobemide	100	600
clomipramine	10	250
nortriptyline	20	250
amitriptyline	10	150
imipramine	10	300

Table with dose ranges, continued

antidepressant	Minimal dose	Maximal dose
dosulepin	50	225
doxepin	25	300
trimipramine	NA	NA
venlafaxine	75	375
mirtazapine	15	45
trazodone	100	400
bupropion	150	300
duloxetine	60	120
agomelatine	25	50
vortioxetine	5	20
hyperici herba	NA	NA
sertraline	50	200
citalopram	10	40
fluoxetine	20	60
escitalopram	5	20
paroxetine	20	50
fluvoxamine	50	300

Table S5.3: Detailed summary of outcome measures per antidepressant prescription group.

AD Type	Facility	N	Continuation	Med. dur. until switch	Prescription duration	Core complaints	Social	Well-being	Experience
SSRI	PG	2244	0.680	77	162	-0.166	0.337	0.301	-0.084
	UMCU	316	0.924	16	92	-0.344	0.386	0.094	-0.217
nSSRI	PG	774	0.625	97	188	-0.174	0.324	0.302	-0.117
	UMCU	147	0.878	21	143	-0.119	0.567	0.229	-0.1128
TriCA	PG	853	0.742	86	175	-0.117	0.322	0.257	-0.077
	UMCU	192	0.901	42	122	-0.098	0.493	0.201	-0.079
TetraCA	PG	827	0.573	45	126	-0.115	0.280	0.308	-0.115
	UMCU	44	0.886	51	49	-0.182	0.689	0.140	-0.222
MAOI	PG	62	0.613	122	212	-0.102	0.167	0.250	0.101
	UMCU	45	0.733	14.5	137	-0.057	0.390	0.187	-0.121
Other	PG	224	0.558	85	170	-0.180	0.399	0.263	-0.147
	UMCU	15	0.733	14.5	54	-0.340	0.432	0.105	-0.317

Note that 171 out of 4808 trajectories at PG and 24 at UMCU concerned trajectories where two types of antidepressants were started on the same day. At PG, 106 concerned combinations of a tetracyclic antidepressant with another type; at UMCU this concerned 12 of the 24 cases. The remainder mainly consisted of combined prescriptions of tricyclic antidepressants, SSRIs and nSSRIs, possibly discontinuation schemes started at the beginning of the admission of the patient. For the outcome measure summaries in this table, if a patient started two types of antidepressants at the same day, this data is incorporated in the two separate corresponding rows in the table. This separation into two entries is offered here purely with the purpose of keeping this table concise. In the Bayesian network analyses in this manuscript, these types of trajectories are viewed as one trajectory with a combination of antidepressant types: the Bayesian network can handle learning such interactions between variables in the model.

Supplementary material for chapter 6

The supplementary material for Chapter 6 can be found online in the publication corresponding to this chapter in Psychiatry Research as: *Yuri van der Does, Rosanne J. Turner, Miel J.H. Bartels, Karin Hagoort, Aaron Metselaar, Floortje E. Scheepers, Peter D. Grünwald, Metten Somers and Edwin van Dellen. Outcome prediction of electroconvulsive therapy for depression. Psychiatry Res. 2023 Aug;326:115328. doi: 10.1016/j.psychres.2023.115328*

Supplementary material for chapter 7

Appendix section S7.A contains detailed proofs and section S7.B additional experiments and figures.

S7.A Proofs

Proof. (of theorem 7.2.1). First consider the basic case with $E^{(m)}$ as in (7.8). As we show below, we have, with $\mathbf{E} \equiv \mathbf{E}_{P_{\theta^*}}$,

$$\mathbf{E}\left[\log E^{(m)}\right] = \mathbf{E}\left[\sum_{j=1}^{m} \log S_j\right] = \mathbf{E}\left[\sum_{j=1..m}\sum_{x\in\{a,b\}}\sum_{i=1..n_x} \log \frac{p_{\breve{\theta}_{x,k_j}|Y^{(j-1)}}(Y_{j,x,i})}{p_{\breve{\theta}_{0,k_j}|Y^{(j-1)}}(Y_{j,x,i})}\right] \geq$$

$$\mathbf{E}\left[\sum_{j=1..m}\sum_{x\in\{a,b\}}\sum_{i=1..n_x} \log \frac{p_{\breve{\theta}_{x,k_j}|Y^{(j-1)}}(Y_{j,x,i})}{p_{\tilde{\theta}_{0,k_j}}(Y_{j,x,i})}\right] \geq$$

$$\mathbf{E}\left[\sum_{\substack{j=1..m\\ x\in\{a,b\}\\ i=1..n_x}} \log \frac{p_{\theta^*_{x,k_j}}(Y_{j,x,i})}{p_{\tilde{\theta}_{0,k_j}}(Y_{j,x,i})} - \sum_{\substack{k=1..K\\ x\in\{a,b\}}} \log(n_x m_k)\right] + O(1) =$$

$$\sum_{k=1..K} m_k \cdot D(P_{\theta^*_{a,k},\theta^*_{b,k}} \| P_{\tilde{\theta}_{0,k},\tilde{\theta}_{0,k}})) + O(\log m) \tag{A.8}$$

where we use notation $D(P_{\theta^*_a,\theta^*_b} \| P_{\theta_0,\theta_0})$ as in (7.4); and $\tilde{\theta}_{0,k}$ is defined as $\arg\min_{\theta\in[0,1]} D(P_{\theta^*_{a,k},\theta^*_{b,k}} \| P_{\theta,\theta})$ which by the same calculation as the one leading up to (7.4, is given by $\tilde{\theta}_{0,k} = (n_a/n)\theta^*_{a,k} + (n_b/n)\theta^*_{b,k}$, and m_k denotes the number of times that an instance of block k was observed in the first m blocks, and we remind the reader that $+O(\log m)$ may also indicate a negative difference of order $\log m$. (A.8) immediately implies the result, using (7.6).

The first two equalities in (A.8) are immediate. The first inequality follows because $P_{\tilde{\theta}_{0,k_j},\tilde{\theta}_{0,k_j}}$ minimizes KL divergence to $P_{\theta^*_{a,k_j},\theta^*_{b,k_j}}$ among all $\theta \in [0,1]$, within each block j. The final equality follows by independence and basic calculus. It remains to show the second inequality. This one follows because we use a prior $W(\theta_{a,k}, \theta_{b,k}$ under which θ_a and θ_b are independently beta distributed with strictly positive densities on $(0,1)$. We can then use a standard Laplace approximation of the Bayesian marginal likelihood to obtain, for each fixed $k \in \{1,\ldots,K\}$, where the expectation $\mathbf{E}$ is over $Y'_{(1)}, \ldots, Y'_{(m')} \sim P_{\theta^*_{a,k}, \theta^*_{b,k}}$:

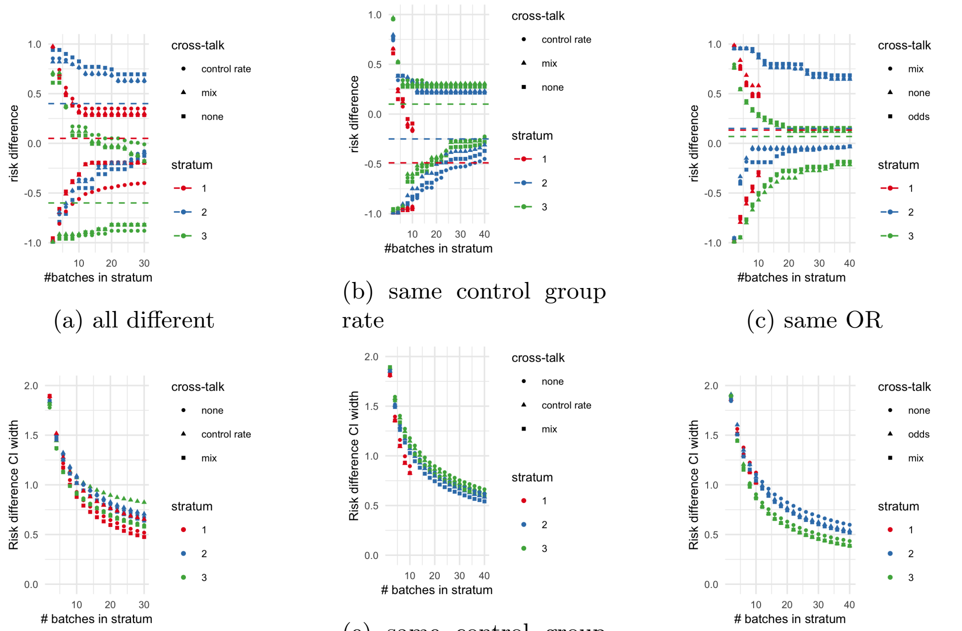
$$
\begin{aligned}
&\mathbf{E}\left[-\log \prod_{j=1}^{m'} \prod_{x \in \{a,b\}} \prod_{i=1}^{n_x} p_{\breve{\theta}_{x,k} | Y^{(j-1)}}(Y_{j,x,i}) \right] = \\
&\mathbf{E}\left[-\log \left(\int \prod_{j=1}^{m'} \prod_{x \in \{a,b\}} \prod_{i=1}^{n_x} p_{\theta_{x,k}}(Y_{j,x,i}) \right) dW(\theta_{a,k}, \theta_{b,k}) \right] \\
&\leq \mathbf{E}\left[\sum_{j=1}^{m'} -\log p_{\theta^*_{a,k}, \theta^*_{b,k}}(Y_{(j)}) \right] + \log(n_a + n_b) m' + O(1).
\end{aligned}
$$

Here the equality is standard telescoping of the Bayesian marginal likelihood, and the inequality is the Laplace approximation, i.e. the same calculation as the one leading up to the $(d/2) \log n$ BIC approximation of Bayesian marginal likelihood for a d-parameter exponential family; here $d = 2$ since we have two free parameters, $\theta^*_{a,k}$ and $\theta^*_{b,k}$; see [Grünwald, 2007, Chapter 8] for proof and detailed explanation).

This shows the result for the basic case that $E^{(m)}$ is arrived at by multiplication, (7.8). The case for $E^{(m)}_{\text{MIX}}$ follows similarly by noting that, by construction, $E^{(m)}_{\text{MIX}} \geq E^{(m)}_{\text{NONE}}/3$, where $E^{(m)}_{\text{NONE}}$ denotes the standard e-process with multiplication and without cross-talk, for which we have already (just) shown the result.

□

S7.B Additional experiments



(a) all different

(b) same control group rate

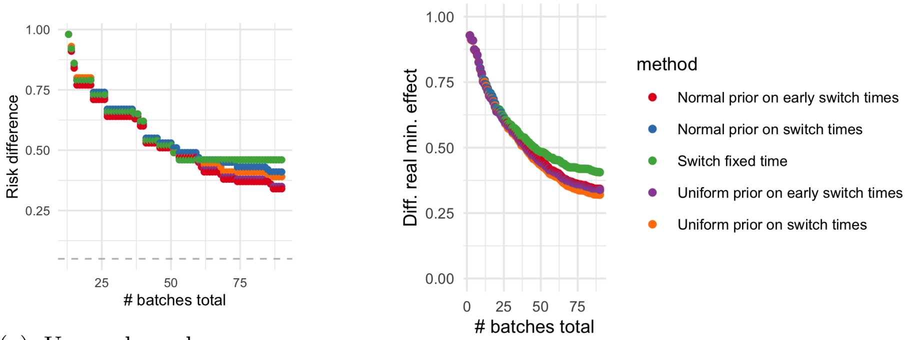
(c) same OR

(d) all different

(e) same control group rate

(f) same OR

Figure S7.1: Examples of 95% stratified confidence intervals ((a), (b) and (c)) and mean confidence interval widths estimated over 100 runs ((d), (e) and (f)) with different types of cross-talk, including mixing different types of cross-talk. In (a), (b) and (c) the true risk difference of the data generating distribution in each stratum is indicated by a dashed line. For (a) and (d), the data were generated by distributions with different control group success rates (0.1, 0.2 and 0.8) and risk differences (0.05, 0.4 and −0.6) in each stratum. For (b) and (e), strata sizes were unbalanced: as can be seen for stratum 1, the red points, data collection stopped after 10 batches. Control group success rates were all 0.5 and risk differences were different (−0.49, −0.25 and 0.1). For (c) and (f), strata sizes were unbalanced as well, and now odds ratios were the same in each stratum (2), but control group rates differed again (0.2, 0.25 and 0.85).



(a) Upper bound sequence example

(b) Average difference with true minimum

Figure S7.2: Example of a confidence sequence and average difference from upper bound to true minimal effect size value through 100 simulations, for different switch priors on j^*. 30 observations were made in each stratum, and the real differences were 0.5, 0.4 and 0.05. For the priors on early switch times, all prior mass was distributed between batch numbers 5 up to 10.α was set to 0.05.

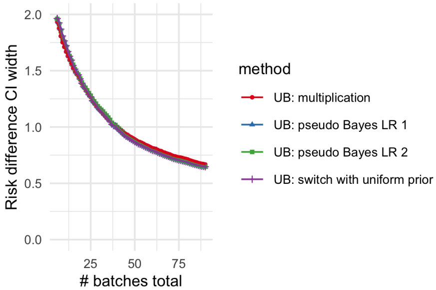
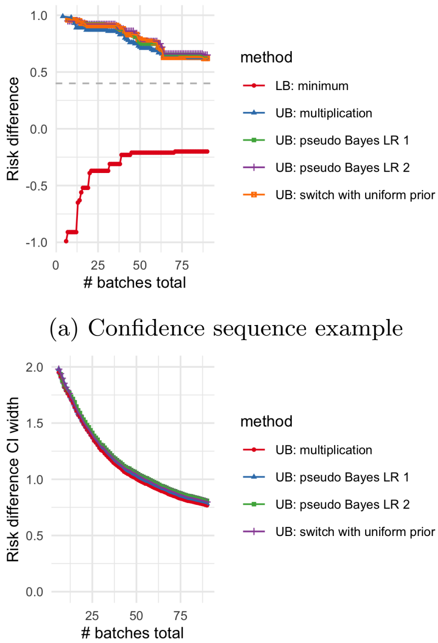


Figure S7.3: Average interval width (upper bound for the respective methods minus lower bound estimated with the minimum method) of confidence sequences for the lower- (LB) and upper (UB) bounds of the minimum effect and estimated through 100 simulations. 30 observations were made in each stratum, and the real differences were 0.5, 0.4 and 0.05. With the switch method, a uniform prior ranging from $j^* = 5$ until 30 was applied. With the pseudo-Bayesian approach, the learning rate η was set to 1 and 2. α was set to 0.05.



(a) Confidence sequence example

(b) Average width

Figure S7.4: Example of confidence sequences for the lower- (LB) and upper (UB) bounds of the minimum effect, and average interval width (upper bound for the respective methods minus lower bound estimated with the minimum method). 30 observations were made in each stratum, and the real differences were 0.4, 0.4 and 0.5. With the switch method, a uniform prior ranging from $m_{\text{switch}} = 5$ until 30 was applied. With the pseudo-Bayesian approach, the learning rate η was set to 1 and 2. α was set to 0.05.

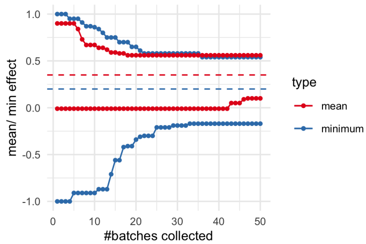


Figure S7.5: Simulated example of a confidence sequence for the mean effect across subpopulations. 25 observations were made in each stratum, and the real risk differences were 0.2 and 0.5. The confidence sequence for the mean difference is plotted alongside the confidence sequence for the minimum of the differences, estimated with pseudo-Bayesian averaging and a uniform switch prior. α was set to 0.05.