STATISTICAL SCIENCE

Volume 38, Number 4 November 2023

Special Issue on Reproducibility and Replicability

***Statistical Science* [ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 38, Number 4, November 2023. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage paid at Cleveland, Ohio and at additional mailing offices.

POSTMASTER: Send address changes to *Statistical Science*, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, Maryland 21769, USA.

Statistical Science Volume 38, Number 4 (525–671) November 2023

Volume 38 Number 4 November 2023

Special Issue on Reproducibility and Replicability

Statistical Science
2023, Vol. 38, No. 4, 525–526
https://doi.org/10.1214/23-STS909

Editorial: Special Issue on Reproducibility and Replicability

Alicia L. Carriquiry, Michael J. Daniels and Nancy Reid

REFERENCES

[1] HARRINGTON, D., D'AGOSTINO, R. B., GATSONIS, C., HOGAN, J. W., HUNTER, D. J., NORMAND, S.-L. T., DRAZEN, J. M. and HAMEL, M. B. (2019). New guidelines for statistical reporting in the. *N. Engl. J. Med.* **381** 285–286. https://doi.org/10.1056/NEJMe1906559

[2] IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *PLoS Med* **2** e124. https://doi.org/10.1371/journal.pmed.0020124

[3] JASA Reproducibility Guide (2020). https://jasa-acs.github.io/repro-guide/.

[4] LEEK, T. J. and JAGER, L. R. (2017). Is most published research really false? *Annu. Rev. Stat. Appl.* **4** 109–122. https://doi.org/10.1146/annurev-statistics-060116-054104

[5] National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC. https://doi.org/10.17226/25303

[6] STODDEN, V. (2020). Theme editor's introduction to reproducibility and replicability in science. *Harv. Data Sci. Rev.* **2** 4.

Alicia Carriquiry is Distinguished Professor and President's Chair and Director of CSAFE, Department of Statistics, Iowa State University, Ames, Iowa 50011, USA (e-mail: alicia@iastate.edu). Mike Daniels is Professor and Chair, Andrew Banks Family Endowed Chair, Department of Statistics, University of Florida, Gainesville, Florida 32603, USA (e-mail: daniels@ufl.edu). Nancy Reid is University Professor, Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5G 1X6, Canada (e-mail: nancym.reid@utoronto.ca).

Statistical Science
2023, Vol. 38, No. 4, 527–542
https://doi.org/10.1214/23-STS902
© Institute of Mathematical Statistics, 2023

Distributionally Robust and Generalizable Inference

Dominik Rothenhäusler and Peter Bühlmann

Abstract. We discuss recently developed methods that quantify the stability and generalizability of statistical findings under distributional changes. In many practical problems, the data is not drawn i.i.d. from the target population. For example, unobserved sampling bias, batch effects, or unknown associations might inflate the variance compared to i.i.d. sampling. For reliable statistical inference, it is thus necessary to account for these types of variation. We discuss and review two methods that allow to quantify distribution stability based on a single dataset. The first method computes the sensitivity of a parameter under worst-case distributional perturbations to understand which types of shift pose a threat to external validity. The second method treats distributional shifts as random which allows to assess average robustness (instead of worst-case). Based on a stability analysis of multiple estimators on a single dataset, it integrates both sampling and distributional uncertainty into a single confidence interval.

Key words and phrases: Distributional robustness, external validity, generalizability, stability, uncertainty quantification.

REFERENCES

[1] Angrist, J., Imbens, G. and Rubin, D. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

[2] Arjovsky, M., Bottou, L., Gulrajani, I. and Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint, arXiv:1907.02893.

[3] Baktashmotlagh, M., Harandi, M. T., Lovell, B. C. and Salzmann, M. (2013). Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision* 769–776.

[4] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York–Chichester–Brisbane. MR0576408

[5] Ben-Tal, A. and Nemirovski, A. (2002). Robust optimization—methodology and applications. *Math. Program.* **92** 453–480. MR1905762 https://doi.org/10.1007/s101070100286

[6] Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215–1222. MR2522270 https://doi.org/10.1111/j.1541-0420.2007.00984.x

[7] Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

[8] Bertsimas, D., Brown, D. B. and Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Rev.* **53** 464–501. MR2834084 https://doi.org/10.1137/080734510

[9] Bühlmann, P. (2014). Discussion of big Bayes stories and BayesBag. *Statist. Sci.* **29** 91–94. MR3201850 https://doi.org/10.1214/13-STS460

[10] Bühlmann, P. (2020). Invariance, causality and robustness: 2018 Neyman Lecture. *Statist. Sci.* **35** 404–426. MR4148216 https://doi.org/10.1214/19-STS721

[11] Bühlmann, P. and Meinshausen, N. (2015). Magging: Maximin aggregation for inhomogeneous large-scale data. *Proc. IEEE* **104** 126–135.

[12] Chen, Y. and Bühlmann, P. (2021). Domain adaptation under structural causal models. *J. Mach. Learn. Res.* **22** Paper No. [261], 80. MR4353040 https://doi.org/10.1007/s11081-020-09512-z

[13] Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 39–67. MR4060976

[14] Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.

[15] Dahabreh, I. J., Petito, L. C., Robertson, S. E., Hernán, M. A. and Steingrimsson, J. A. (2020). Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology* **31** 334–344.

[16] Denzen, N. (1978). Sociological methods: A sourcebook. New York.

[17] Devaux, M. and Egami, N. (2022). Quantifying robustness to external validity bias.

Dominik Rothenhäusler is Assistant Professor, Department of Statistics, Stanford University, Stanford, California 94305-4020, USA (e-mail: rdominik@stanford.edu). Peter Bühlmann is Professor, Seminar for Statistics, ETH Zürich, Switzerland (e-mail: buhlmann@stat.math.ethz.ch).

[18] Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840 https://doi.org/10.1214/15-STS527

[19] Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368.

[20] Donsker, M. D. and Varadhan, S. R. S. (1976). Asymptotic evaluation of certain Markov process expectations for large time. III. *Comm. Pure Appl. Math.* **29** 389–461. MR0428471 https://doi.org/10.1002/cpa.3160290405

[21] Dorn, J., Guo, K. and Kallus, N. (2021). Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. arXiv preprint, arXiv:2112.11449.

[22] Gerber, A. S., Green, D. P. and Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *Amer. Polit. Sci. Rev.* **102** 33–48.

[23] Gong, B., Shi, Y., Sha, F. and Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In 2012 *IEEE Conference on Computer Vision and Pattern Recognition* 2066–2073. IEEE.

[24] Gopalan, R., Li, R. and Chellappa, R. (2011). Domain adaptation for object recognition: An unsupervised approach. In 2011 *International Conference on Computer Vision* 999–1006. IEEE.

[25] Gupta, S. and Rothenhäusler, D. (2021). The s-value: Evaluating stability with respect to distributional shifts. To appear in *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

[26] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0829458

[27] Heinze-Deml, C. and Meinshausen, N. (2021). Conditional variance penalties and domain shift robustness. *Mach. Learn.* **110** 303–348. MR4207502 https://doi.org/10.1007/s10994-020-05924-1

[28] Heinze-Deml, C., Peters, J. and Meinshausen, N. (2018). Invariant causal prediction for nonlinear models. *J. Causal Inference* **6** Art. No. 20170016, 35. MR4335430 https://doi.org/10.1515/jci-2017-0016

[29] Heller, R., Golland, Y., Malach, R. and Benjamini, Y. (2007). Conjunction group analysis: An alternative to mixed/random effect analysis. *NeuroImage* **37** 1178–1185.

[30] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

[31] Huber, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Stat.* **36** 1753–1758. MR0185747 https://doi.org/10.1214/aoms/1177699803

[32] Huggins, J. H. and Miller, J. W. (2023). Reproducible model selection using bagged posteriors. *Bayesian Anal.* **18** 79–104. MR4515726 https://doi.org/10.1214/21-ba1301

[33] Imbens, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statist. Sci.* **29** 323–358. MR3264545 https://doi.org/10.1214/14-STS480

[34] Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference—for Statistics*, *Social*, *and Biomedical Sciences*: *An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

[35] Ioannidis, J. P. A. (2005). Why most published research findings are false. *Chance* **18** 40–47. MR2216666 https://doi.org/10.1080/09332480.2005.10722754

[36] Jeong, Y. and Rothenhäusler, D. (2022). Calibrated inference: Statistical inference that accounts for both sampling uncertainty and distributional uncertainty. arXiv preprint, arXiv:2202.11886.

[37] Jin, Y., Ren, Z. and Candès, E. J. (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proc. Natl. Acad. Sci. USA* **120** Paper No. e2214889120, 13. MR4575282

[38] Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

[39] Li, S., Song, S. and Huang, G. (2017). Prediction reweighting for domain adaption. *IEEE Trans. Neural Netw. Learn. Syst.* **28** 1682–1695. MR3666190 https://doi.org/10.1109/TNNLS.2016.2538282

[40] Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970 https://doi.org/10.1214/13-AOS1175

[41] Long, M., Wang, J., Ding, G., Sun, J. and Yu, P. S. (2014). Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1410–1417.

[42] Meinshausen, N. (2018). Causality from a distributional robustness point of view. In 2018 *IEEE Data Science Workshop* (*DSW*) 6–10. IEEE.

[43] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523 https://doi.org/10.1111/j.1467-9868.2010.00740.x

[44] Meinshausen, N. and Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. *Ann. Statist.* **43** 1801–1830. MR3357879 https://doi.org/10.1214/15-AOS1325

[45] Meinshausen, N., Meier, L. and Bühlmann, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584 https://doi.org/10.1198/jasa.2009.tm08647

[46] Modigliani, F. (1966). The life cycle hypothesis of saving, the demand for wealth and the supply of capital. *Soc. Res.* 160–217.

[47] Munafò, M. R. and Smith, G. D. (2018). Repeating experiments is not enough. *Nature* **553** 399–401.

[48] Neykov, M., Ning, Y., Liu, J. S. and Liu, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statist. Sci.* **33** 427–443. MR3843384 https://doi.org/10.1214/18-STS661

[49] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.

[50] Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Serv. Res.* **34** 1189.

[51] Pearl, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

[52] Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[53] Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K. and Wang, B. (2019). Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 1406–1415.

[54] Peters, J., Bühlmann, P. and Meinshausen, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012. With comments and a rejoinder. MR3557186 https://doi.org/10.1111/rssb.12167

[55] PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2017). *Elements of Causal Inference*: *Foundations and Learning Algorithms*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3822088

[56] QUINONERO-CANDELA, J., SUGIYAMA, M., SCHWAIGHOFER, A. and LAWRENCE, N. D. (2009). *Dataset Shift in Machine Learning*. Mit Press.

[57] ROJAS-CARULLA, M., SCHÖLKOPF, B., TURNER, R. and PETERS, J. (2018). Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **19** Paper No. 36, 34. MR3862443

[58] ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915 https://doi.org/10.1093/biomet/74.1.13

[59] ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2021). Anchor regression: Heterogeneous data meet causality. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 215–246. MR4250274 https://doi.org/10.1111/rssb.12398

[60] ROTHWELL, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* **365** 82–93.

[61] SAGAWA, S., KOH, P. W., HASHIMOTO, T. B. and LIANG, P. (2019). Distributionally robust neural networks. In *International Conference on Learning Representations*.

[62] SINHA, A., NAMKOONG, H. and DUCHI, J. (2017). Certifiable distributional robustness with principled adversarial training. arXiv preprint, arXiv:1710.10571, presented at Sixth International Conference on Learning Representations (ICLR 2018).

[63] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[64] VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. With a rejoinder by the authors. MR3724985 https://doi.org/10.1214/17-BA1065

[65] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[66] WANG, J. and OWEN, A. B. (2019). Admissibility in partial conjunction testing. *J. Amer. Statist. Assoc.* **114** 158–168. MR3941245 https://doi.org/10.1080/01621459.2017.1385465

[67] WITTEVEEN, E., WIESKE, L., SOMMERS, J., SPIJKSTRA, J.-J., DE WAARD, M. C., ENDEMAN, H., RIJKENBERG, S., DE RUIJTER, W., SLEESWIJK, M. et al. (2020). Early prediction of intensive care unit–acquired weakness: A multicenter external validation study. *J. Intens. Care Med.* **35** 595–605.

[68] YADLOWSKY, S., NAMKOONG, H., BASU, S., DUCHI, J. and TIAN, L. (2022). Bounds on the conditional and average treatment effect with unobserved confounding factors. *Ann. Statist.* **50** 2587–2615. MR4505372 https://doi.org/10.1214/22-aos2195

[69] YU, B. (2013). Stability. *Bernoulli* **19** 1484–1500. MR3102560 https://doi.org/10.3150/13-BEJSP14

[70] YU, B. and KUMBIER, K. (2020). Veridical data science. *Proc. Natl. Acad. Sci. USA* **117** 3920–3929. MR4075122 https://doi.org/10.1073/pnas.1901326117

[71] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

[72] ZHAO, Q., SMALL, D. S. and BHATTACHARYA, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 735–761. MR3997099

Statistical Science
2023, Vol. 38, No. 4, 543–556
https://doi.org/10.1214/23-STS891

Defining Replicability of Prediction Rules

Giovanni Parmigiani

Abstract. In this article, I propose an approach for defining replicability for prediction rules. Motivated by a recent report by the U.S.A. National Academy of Sciences, I start from the perspective that replicability is obtaining consistent results across studies suitable to address the same prediction question, each of which has obtained its own data. I then discuss concept and issues in defining key elements of this statement. I focus specifically on the meaning of "consistent results" in typical utilization contexts, and propose a multi-agent framework for defining replicability, in which agents are neither allied nor adversaries. I recover some of the prevalent practical approaches as special cases. I hope to provide guidance for a more systematic assessment of replicability in machine learning.

Key words and phrases: Replicability, prediction, decision theory.

REFERENCES

[1] Barba, L. A. Terminologies for reproducible research. Available at arXiv:1802.03311.

[2] Beckers, R., Kwade, Z. and Zanca, F. (2021). The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. *Phys. Med.* **83** 1–8. https://doi.org/10.1016/j.ejmp.2021.02.011

[3] Bernau, C., Riester, M., Boulesteix, A., Parmigiani, G., Huttenhower, C., Waldron, L. and Trippa, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30** i105–i112. https://doi.org/10.1093/bioinformatics/btu279.

[4] Breiman, L. (1996). Stacked regressions. *Mach. Learn.* **24** 49–64. https://doi.org/10.1007/BF00117832

[5] Breiman, L. (2001). Statistical modeling: The two cultures. *Statist. Sci.* **16** 199–231. https://doi.org/10.1214/ss/1009213726

[6] Broman, K., Cetinkaya-Rundel, M., Nussbaum, A., Paciorek, C., Peng, R., Turek, D. and Wickham, H. (2017). *Recommendations to funding agencies for supporting reproducible research*. Amer. Statist. Assoc., Alexandria, VA.

[7] Chang, L.-B. and Geman, D. (2015). Tracking cross-validated estimates of prediction error as studies accumulate. *J. Amer. Statist. Assoc.* **110** 1239–1247. https://doi.org/10.1080/01621459.2014.1002926

[8] Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.-M. et al. (2014). External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med. Res. Methodol.* **14** 40. https://doi.org/10.1186/1471-2288-14-40

[9] D'Alterio, C., Spina, A., Arenare, L. and Chiodini, P. (2022). Biological role of tumor/stromal CXCR4-CXCL12-CXCR7 in MITO16A/MaNGO-OV2 advanced ovarian cancer patients. *Cancers* **14** 1849.

[10] Davison, C. A. and Hinkley, D. V. (1997). *Boostrap Methods and Their Applications*. Cambridge Univ. Press, New York.

[11] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the* 3rd *Innovations in Theoretical Computer Science Conference* 214–226. ACM, New York. MR3388391

[12] Ebrahimian, S., Kalra, M. K., Agarwal, S., Bizzo, B. C., Elkholy, M., Wald, C., Allen, B. and Dreyer, K. J. FDA-regulated AI algorithms: Trends, strengths, and gaps of validation studies. *Acad. Radiol.* **29** 559–566. https://doi.org/10.1016/j.acra.2021.09.002

[13] Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S. and Saria, S. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385** 283–286. https://doi.org/10.1056/NEJMc2104626

[14] Fisher, R. A. (1925). *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh.

[15] Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J. et al. (2013). curatedOvarianData: Clinically annotated data for the ovarian cancer transcriptome. *Database* (*Oxford*) **2013** bat013. https://doi.org/10.1093/database/bat013.

[16] Geisser, S. (1993). *Predictive Inference*: *An Introduction*, Chapman & Hall, New York.

[17] Goodman, S. N., Fanelli, D. and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Sci. Transl. Med.* **8** 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

[18] Heller, R., Bogomolov, M. and Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proc. Natl. Acad. Sci. USA* **111** 16262–16267. https://doi.org/10.1073/pnas.1314814111

[19] Jaljuli, I., Benjamini, Y., Shenhav, L., Panagiotou, O. A. and Heller, R., Quantifying replicability and consistency in systematic reviews. *Stat. Biopharm. Res.* **15** 372–385. https://doi.org/10.1080/19466315.2022.2050291

Giovanni Parmigiani is Professor, Department of Data Science, Dana Farber Cancer Institute & Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: gp@ds.dfci.harvard.edu).

[20] KEENEY, R. L., RAIFFA, H. and MEYER, R. F. (1976). *Decisions with Multiple Objectives*: *Preferences and Value Tradeoffs*, Wiley & Sons, New York.

[21] KELLY, C. J., KARTHIKESALINGAM, A., SULEYMAN, M., CORRADO, G. and KING, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. **17** 195. https://doi.org/10.1186/s12916-019-1426-2

[22] KENETT, R. S. and SHMUELI, G. (2015). Clarifying the terminology that describes scientific reproducibility. *Nat*. *Methods* **12** 699–699. https://doi.org/10.1038/nmeth.3489

[23] KOH, P. W., SAGAWA, S., MARKLUND, H., XIE, S. M., ZHANG, M., BALSUBRAMANI, A., HU, W., YASUNAGA, M., LANAS PHILLIPS, R. et al. WILDS: A benchmark of in-the-wild distribution shifts. Available at arXiv:2012.07421.

[24] KOUW, W. and LOOG, M. (2019). An introduction to domain adaptation and transfer learning. Available at arXiv:1812.11806.

[25] LEE, A. Y., YANAGIHARA, R. T., LEE, C. S., BLAZES, M., JUNG, H. C., CHEE, Y. E., GENCARELLA, M. D., GEE, H., MAA, A. Y. et al. (2021). Head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. *Diabetes Care* **44** 1168–1175. https://doi.org/10.2337/dc20-1877

[26] LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. and IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat*. *Rev*. *Genet*. **11** 733–739. https://doi.org/10.1038/nrg2825

[27] LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. **3** e161. https://doi.org/10.1371/journal.pgen.0030161

[28] LEMAY, A., HOEBEL, K., BRIDGE, C. P., BEFANO, B., SANJOSÉ, S. D., EGEMEN, D., RODRIGUEZ, A. C., SCHIFFMAN, M., CAMPBELL, J. P. et al. (2022). Improving the repeatability of deep learning models with Monte Carlo dropout. *npj Digit*. *Med*. **5** 174. https://doi.org/10.1038/s41746-022-00709-3

[29] LOEWINGER, G., PATIL, P. KISHIDA, K. T. and PARMIGIANI, G. (2022). Hierarchical resampling for bagging in multistudy prediction with applications to human neurochemical sensing. *Ann*. *Appl*. *Stat*. **16** 2145–2165. https://doi.org/10.1214/21-AOAS1574

[30] METZ, C. E. Basic principles of ROC analysis. *Semin*. *Nucl*. *Med* **8** 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

[31] MORENO-TORRES, J. G., RAEDER, T., ALAIZ-RODRÍGUEZ, R. and CHAWLA, N. V. (2012). A unifying view on dataset shift in classification. *Pattern Recognit*. **45** 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

[32] COMMITTEE ON REPRODUCIBILITY AND REPLICABILITY IN SCIENCE (2019). *Reproducibility and Replicability in Science*. National Academies Press, Washington, D.C. https://doi.org/10.17226/25303

[33] PATIL, P. and PARMIGIANI, G. (2018). Training replicable predictors in multiple studies. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **115** 2578–2583.

[34] QIN, Z. Z., SANDER, M. S., RAI, B., TITAHONG, C., SUDRUNGROT, S., LAAH, S. N., ADHIKARI, L. M., CARTER, E. J., PURI, L. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci*. *Rep*. **9** 15000. https://doi.org/10.1038/s41598-019-51503-3

[35] RAMSEY, F. (1926). *The Foundations of Mathematics*, Oxford University Press Oxford.

[36] RASHID, N. U., LI QUEFENG, Y., JEN, J. and IBRAHIM, J. G. (2020). Modeling between-study heterogeneity for improved replicability in gene signature selection and clinical prediction. *J*. *Amer*. *Statist*. *Assoc*. **115** 1125–1138. https://doi.org/10.1080/01621459.2019.1671197

[37] RIESTER, M., TAYLOR, J. M., FEIFER, A., KOPPIE, T., ROSENBERG, J. E., DOWNEY, R. J., BOCHNER, B. H. and MICHOR, F. (2012). Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. *Clin*. *Cancer Res*. **18** 1323–1333. https://doi.org/10.1158/1078-0432.CCR-11-2271

[38] RIESTER, M., WEI, W., WALDRON, L., CULHANE, A. C., TRIPPA, L., OLIVA, E., KIM, S.-H., MICHOR, F., HUTTENHOWER, C. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J*. *Natl*. *Cancer Inst*. **106** dju048–dju048. https://doi.org/10.1093/jnci/dju048

[39] RIESTER, M., WEI, W., WALDRON, L., CULHANE, A. C., TRIPPA, L., OLIVA, E., KIM, S.-H., MICHOR, F., HUTTENHOWER, C. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J*. *Natl*. *Cancer Inst*. https://doi.org/10.1093/jnci/dju048

[40] SAVAGE, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

[41] SISMONDO, S. (2004). *An Introduction to Science and Technology Studies*. Blackwell, Malden, MA.

[42] STEYERBERG, E. W. and VERGOUWE, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur*. *Heart J*. **35** 1925–1931. https://doi.org/10.1093/eurheartj/ehu207

[43] STIGLER, S. M. (1982). Thomas Bayes's Bayesian inference. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *A* **145** 250–258. MR0669120 https://doi.org/10.2307/2981538

[44] TRIPPA, L., WALDRON, L., HUTTENHOWER, C. and PARMIGIANI, G. (2015). Bayesian nonparametric cross-study validation of prediction methods. *Ann*. *Appl*. *Stat*. **9** 402–428.

[45] VENTZ, S., MAZUMDER, R. and TRIPPA, L. (2022). Integration of survival data from multiple studies. *Biometrics* **78** 1365–1376.

[46] VIJAYAKUMAR, R. and CHEUNG, M. W. L. Assessing replicability of machine learning results: An introduction to methods on predictive accuracy in social sciences. *Soc*. *Sci*. *Comput*. *Rev*. **39** 768–801.

[47] VIJAYAKUMAR, R. and CHEUNG, M. W. L. (2018). Replicability of machine learning models in the social sciences: A case study in variable selection. *Z*. *Psychol*. **226** 259–273.

[48] WALDRON, L., HAIBE-KAINS, B., CULHANE, A. C., RIESTER, M., DING, J., WANG, X. V., AHMADIFAR, M., TYEKUCHEVA, S., BERNAU, C. (2014). Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *J*. *Natl*. *Cancer Inst*. **106** dju049. https://doi.org/10.1093/jnci/dju049

[49] WANG, J., LAN, C., LIU, C., OUYANG, Y. and QIN, T. (2021). Generalizing to unseen domains: A survey on domain generalization. Available at arXiv:2103.03097.

[50] WONG, A., JIE, C., LYONS, P. G., DUTTA, S., MAJOR, V. J., ÖTLEŞ, E. and SINGH, K. (2021). Quantification of sepsis model alerts in 24 US hospitals before and during the Covid-19 pandemic. *JAMA Netw*. *Open* **4** e2135286–e2135286. https://doi.org/10.1001/jamanetworkopen.2021.35286

[51] WU, E., WU, K., DANESHJOU, R., OUYANG, D., HO, D. E. and ZOU, J. (2021). How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. *Nat*. *Med*. **27** 582–584. https://doi.org/10.1038/s41591-021-01312-x

[52] YU, B. and KUMBIER, K. (2020). Veridical data science. *Proc. Natl. Acad. Sci. USA* **117** 3920–3929. MR4075122 https://doi.org/10.1073/pnas.1901326117

[53] ZEMEL, R., SWERSKY, K. and PITASSI, T. (2013). Learning fair representations. In *Proceedings of the* 30*th International Conference on Machine Learning*.

[54] ZHANG, Y., PATIL PRASAD, J., EVAN, W. and PARMIGIANI, G. (2021). Robustifying genomic classifiers to batch effects via ensemble learning. *Bioinformatics* **37** 1521–1527. https://doi.org/10.1093/bioinformatics/btaa986

[55] ZHUANG, F., QI, Z., DUAN, K., XI, D., ZHU, Y., ZHU, H., XIONG, H. and HE, Q. (2020). A comprehensive survey on transfer learning. Available at arXiv:02685.

[56] INSTITUTE OF MEDICINE (2012). *Evolution of Translational Omics*. The National Academies Press, Washington, D.C.

Statistical Science
2023, Vol. 38, No. 4, 557–575
https://doi.org/10.1214/23-STS901
This research was funded, in whole or in part, by [UK Medical Research Council, MC_UU_00002/14]. A CC BY 4.0 license is applied to this article arising from this submission, in accordance with the grant's open access conditions

Online Multiple Hypothesis Testing

David S. Robertson, James M. S. Wason and Aaditya Ramdas

Abstract. Modern data analysis frequently involves large-scale hypothesis testing, which naturally gives rise to the problem of maintaining control of a suitable type I error rate, such as the false discovery rate (FDR). In many biomedical and technological applications, an additional complexity is that hypotheses are tested in an online manner, one-by-one over time. However, traditional procedures that control the FDR, such as the Benjamini–Hochberg procedure, assume that all p-values are available to be tested at a single time point. To address these challenges, a new field of methodology has developed over the past 15 years showing how to control error rates for online multiple hypothesis testing. In this framework, hypotheses arrive in a stream, and at each time point the analyst decides whether to reject the current hypothesis based both on the evidence against it, and on the previous rejection decisions. In this paper, we present a comprehensive exposition of the literature on online error rate control, with a review of key theory as well as a focus on applied examples. We also provide simulation results comparing different online testing algorithms and an up-to-date overview of the many methodological extensions that have been proposed.

Key words and phrases: A/B testing, data repositories, platform trials, type I error rate.

REFERENCES

1000 Genomes Project Consortium et al. (2015). A global reference for human genetic variation. *Nature* **526** 68–74.

Aharoni, E. and Rosset, S. (2014). Generalized α-investing: Definitions, optimality results and application to public databases. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 771–794. MR3248676 https://doi.org/10.1111/rssb.12048

Berman, R. and Van den Bulte, C. (2021). False discovery in A/B testing. *Manage. Sci.* **69** 6762–6782. https://doi.org/10.1287/mnsc.2021.4207

Bretz, F., Maurer, W. and Xi, D. (2019). Replicability, reproducibility, and multiplicity in drug development. *Chance* **32** 4–11.

Bretz, F. and Westfall, P. H. (2014). Multiplicity and replicability: Two sides of the same coin. *Pharm. Stat.* **13** 343–344. https://doi.org/10.1002/pst.1648

Burman, C.-F., Sonesson, C. and Guilbaud, O. (2009). A recycling framework for the construction of Bonferroni-based multiple tests. *Stat. Med.* **28** 739–761. MR2656960 https://doi.org/10.1002/sim.3513

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* **104** 1467–1481. MR2597000 https://doi.org/10.1198/jasa.2009.tm08415

Chen, S. and Arias-Castro, E. (2021). On the power of some sequential multiple testing procedures. *Ann. Inst. Statist. Math.* **73** 311–336. MR4233523 https://doi.org/10.1007/s10463-020-00752-5

Chen, S. and Kasiviswanathan, S. (2020). Contextual online false discovery rate control. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (S. Chiappa and R. Calandra, eds.). *Proceedings of Machine Learning Research* **108** 952–961. PMLR.

Cook, T., Dubey, H. T., Lee, J.-A., Zhu, G., Zhao, T. and Flaherty, P. (2022). Cost-aware generalized α-investing for multiple hypothesis testing. arXiv preprint. Available at arXiv:2210.17514.

Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S. and Hernán, M. A. (2019). Avoidable flaws in observational analyses: An application to statins and cancer. *Nat. Med.* **25** 1601–1606.

Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., Meehan, T. F., Weninger, W. J., Westerberg, H. et al. (2016). High-throughput discovery of novel developmental phenotypes. *Nature* **537** 508–514.

Döhler, S., Meah, I. and Roquain, E. (2021). Online multiple testing with super-uniformity reward. arXiv preprint. Available at arXiv:2110.01255.

David S. Robertson is a Senior Research Associate, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK (e-mail: david.robertson@mrc-bsu.cam.ac.uk). James M. S. Wason is a Professor, Biostatistics, Population Health Sciences Institute, Newcastle University, Newcastle, UK (e-mail: james.wason@newcastle.ac.uk). Aaditya Ramdas is an Assistant Professor, Departments of Statistics and Machine Learning, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: aramdas@stat.cmu.edu).

Farcomeni, A. and Finos, L. (2013). FDR control with pseudo-gatekeeping based on a possibly data driven order of the hypotheses. *Biometrics* **69** 606–613. MR3106588 https://doi.org/10.1111/biom.12058

Finos, L. and Farcomeni, A. (2011). k-FWER control without p-value adjustment, with application to detection of genetic determinants of multiple sclerosis in Italian twins. *Biometrics* **67** 174–181. MR2898829 https://doi.org/10.1111/j.1541-0420.2010.01443.x

Fischer, L., Roig, M. B. and Brannath, W. (2023a). An adaptive-discard-graph for online error control. arXiv preprint. Available at arXiv:2301.11711.

Fischer, L., Roig, M. B. and Brannath, W. (2023b). An exhaustive ADDIS principle for online FWER control. Available at arXiv:2308.13827.

Fisher, A. (2021). SAFFRON and LORD ensure online control of the false discovery rate under positive, local dependence. arXiv preprint. Available at arXiv:2110.08161.

Fisher, A. J. (2022). Online control of the false discovery rate under "decision deadlines". In *International Conference on Artificial Intelligence and Statistics*, 8340–8359. PMLR.

Foster, D. P. and Stine, R. A. (2008). α-Investing: A procedure for sequential control of expected false discoveries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 429–444. MR2424761 https://doi.org/10.1111/j.1467-9868.2007.00643.x

Gang, B., Sun, W. and Wang, W. (2023). Structure-adaptive sequential testing for online false discovery rate control. *J. Amer. Statist. Assoc.* **118** 732–745. MR4571154 https://doi.org/10.1080/01621459.2021.1955688

Goodman, S. N., Fanelli, D. and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Sci. Transl. Med.* **8** 341ps12–341ps12.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biol.* **13** e1002106. https://doi.org/10.1371/journal.pbio.1002106

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.* **49** 1055–1080. MR4255119 https://doi.org/10.1214/20-aos1991

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* **2** e124.

James, N. D., de Bono, J. S., Spears, M. R., Clarke, N. W., Mason, M. D., Dearnaley, D. P., Ritchie, A. W. S., Amos, C. L., Gilson, C. et al. (2017). Abiraterone for prostate cancer not previously treated with hormone therapy. *N. Engl. J. Med.* **377** 338–351. https://doi.org/10.1056/NEJMoa1702900

James, N. D., Sydes, M. R., Clarke, N. W., Mason, M. D., Dearnaley, D. P., Anderson, J., Popert, R. J., Sanders, K., Morgan, R. C. et al. (2008). STAMPEDE: Systemic therapy for advancing or metastatic prostate cancer—a multi-arm multi-stage randomised controlled trial. *Clin. Oncol.* **20** 577–581. https://doi.org/10.1016/j.clon.2008.07.002

James, N. D., Sydes, M. R., Clarke, N. W., Mason, M. D., Dearnaley, D. P., Spears, M. R., Ritchie, A. W. S., Parker, C. C., Russell, J. M. et al. (2016). Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): Survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. *Lancet* **387** 1163–1177. https://doi.org/10.1016/S0140-6736(15)01037-5

Javanmard, A. and Montanari, A. (2015). On online control of false discovery rate. arXiv preprint. Available at arXiv:1502.06197.

Javanmard, A. and Montanari, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *Ann. Statist.* **46** 526–554. MR3782376 https://doi.org/10.1214/17-AOS1559

Johari, R., Koomen, P., Pekelis, L. and Walsh, D. (2022). Always valid inference: Continuous monitoring of A/B tests. *Oper. Res.* **70** 1806–1821. MR4451064 https://doi.org/10.1287/opre.2021.2135

Karp, N. A., Mason, J., Beaudet, A. L., Benjamini, Y., Bower, L., Braun, R. E., Brown, S. D. M., Chesler, E. J., Dickinson, M. E. et al. (2017). Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nat. Commun.* **8** 15475.

Katsevich, E. and Ramdas, A. (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *Ann. Statist.* **48** 3465–3487. MR4185816 https://doi.org/10.1214/19-AOS1938

Kohavi, R., Tang, D., Xu, Y., Hemkens, L. G. and Ioannidis, J. (2020). Online randomized controlled experiments at scale: Lessons and extensions to medicine. *Trials* **21** 1–9.

Koscielny, G., Yaikhom, G., Iyer, V., Meehan, T. F., Morgan, H., Atienza-Herrero, J., Blake, A., Chen, C.-K., Easty, R. et al. (2013). The international mouse phenotyping consortium web portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res.* **42** D802–D809.

Liou, L., Hornburg, M. and Robertson, D. S. (2023). Global FDR control across multiple RNAseq experiments. *Bioinformatics* **39**. https://doi.org/10.1093/bioinformatics/btac718

Liou, L. and Robertson, D. S. (2021). OnlineFDRexplore. Available at http://shiny.mrc-bsu.cam.ac.uk/apps/onlineFDRexplore/. Accessed: 2022-06-15.

Mason, M. D., Clarke, N. W., James, N. D., Dearnaley, D. P., Spears, M. R., Ritchie, A. W. S., Attard, G., Cross, W., Jones, R. J. et al. (2017). Adding celecoxib with or without zoledronic acid for hormone-naïve prostate cancer: Long-term survival results from an adaptive, multiarm, multistage, platform, randomized controlled trial. *J. Clin. Oncol.* **35** 1530–1541. https://doi.org/10.1200/JCO.2016.69.0677

Parker, C. C., James, N. D., Brawley, C. D., Clarke, N. W., Hoyle, A. P., Ali, A., Ritchie, A. W. S., Attard, G., Chowdhury, S. et al. (2018). Radiotherapy to the primary tumour for newly diagnosed, metastatic prostate cancer (STAMPEDE): A randomised controlled phase 3 trial. *Lancet* **392** 2353–2366. https://doi.org/10.1016/S0140-6736(18)32486-3

Ramdas, A., Yang, F., Wainwright, M. J. and Jordan, M. I. (2017). Online control of the false discovery rate with decaying memory. In *Advances in Neural Information Processing Systems* **30** 5650–5659.

Ramdas, A., Zrnic, T., Wainwright, M. and Jordan, M. (2018). SAFFRON: An adaptive algorithm for online control of the false discovery rate. In *Proceedings of the* 35*th International Conference on Machine Learning. Proceedings of Machine Learning Research* **80** 4286–4294.

Rebjock, Q., Kurt, B., Januschowski, T. and Callot, L. (2021). Online false discovery rate control for anomaly detection in time series. *Adv. Neural Inf. Process. Syst.* **34** 26487–26498.

Robertson, D. S., Liou, L., Ramdas, A. and Karp, N. A. (2021). onlineFDR: Online error rate control. https://doi.org/10.18129/B9.bioc.onlineFDR

Robertson, D. S. and Wason, J. M. S. (2018). Online control of the false discovery rate in biomedical research. arXiv preprint. Available at arXiv:1809.07292.

Robertson, D. S., Wason, J. M. S., König, F., Posch, M. and Jaki, T. (2023). Online error rate control for platform trials. *Stat. Med.* **42** 2475–2495. MR4596806 https://doi.org/10.1002/sim.9733

ROBERTSON, D. S., WILDENHAIN, J., JAVANMARD, A. and KARP, N. A. (2019). onlineFDR: An R package to control the false discovery rate for growing data repositories. *Bioinformatics* **35** 4196–4199. https://doi.org/10.1093/bioinformatics/btz191

SAVILLE, B. R. and BERRY, S. M. (2016). Efficiencies of platform clinical trials: A vision of the future. *Clin. Trials* **13** 358–366. https://doi.org/10.1177/1740774515626362

ŠIDÁK, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **62** 626–633. MR0216666

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302 https://doi.org/10.1111/1467-9868.00346

TIAN, J. and RAMDAS, A. (2019). ADDIS: An adaptive discarding algorithm for online FDR control with conservative nulls. *Adv. Neural Inf. Process. Syst.* **32**.

TIAN, J. and RAMDAS, A. (2021). Online control of the familywise error rate. *Stat. Methods Med. Res.* **30** 976–993. MR4259882 https://doi.org/10.1177/0962280220983381

TUKEY, J. W. (1953). *The Collected Works of John W. Tukey, Vol. III. Multiple comparisons*: 1948–1983. Chapman & Hall, London.

WEINSTEIN, A. and RAMDAS, A. (2020). Online control of the false coverage rate and false sign rate. In *International Conference on Machine Learning* 10193–10202. PMLR.

WELLCOME TRUST CASE CONTROL CONSORTIUM et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **447** 661–678.

XU, Z. and RAMDAS, A. (2022). Dynamic algorithms for online multiple testing. In *Mathematical and Scientific Machine Learning* 955–986. PMLR.

YANG, F., RAMDAS, A., JAMIESON, K. and WAINWRIGHT, M. (2017). A framework for multi-A(rmed)/B(anit) testing with online FDR control. In *Advances in Neural Information Processing Systems* **30** 5959–5968.

ZEEVI, Y., ASTASHENKO, S. and BENJAMINI, Y. (2020). Ignored evident multiplicity harms replicability—adjusting for it offers a remedy. arXiv preprint. Available at arXiv:2006.11585.

ZEHETMAYER, S., POSCH, M. and KOENIG, F. (2022). Online control of the False Discovery Rate in group-sequential platform trials. *Stat. Methods Med. Res.* **31** 2470–2485. MR4513312 https://doi.org/10.1177/09622802221129051

ZHAO, Q., SMALL, D. S. and SU, W. (2019). Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *J. Amer. Statist. Assoc.* **114** 1291–1304. MR4011780 https://doi.org/10.1080/01621459.2018.1497499

ZRNIC, T., JIANG, D., RAMDAS, A. and JORDAN, M. (2020). The power of batching in multiple hypothesis testing. In *International Conference on Artificial Intelligence and Statistics* 3806–3815. PMLR.

ZRNIC, T., RAMDAS, A. and JORDAN, M. I. (2021). Asynchronous online testing of multiple hypotheses. *J. Mach. Learn. Res.* **22** Paper No. 33, 39. MR4253726 https://doi.org/10.1515/ijnsns-2019-0210

Statistical Science
2023, Vol. 38, No. 4, 576–601
https://doi.org/10.1214/23-STS894
© Institute of Mathematical Statistics, 2023

Game-Theoretic Statistics and Safe Anytime-Valid Inference

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk and Glenn Shafer

Abstract. Safe anytime-valid inference (SAVI) provides measures of statistical evidence and certainty—e-processes for testing and confidence sequences for estimation—that remain valid at all stopping times, accommodating continuous monitoring and analysis of accumulating data and optional stopping or continuation for any reason. These measures crucially rely on test martingales, which are nonnegative martingales starting at one. Since a test martingale is the wealth process of a player in a betting game, SAVI centrally employs game-theoretic intuition, language and mathematics. We summarize the SAVI goals and philosophy, and report recent advances in testing composite hypotheses and estimating functionals in nonparametric settings.

Key words and phrases: Test martingales, Ville's inequality, universal inference, reverse information projection, e-process, optional stopping, confidence sequence, nonparametric composite hypothesis testing.

REFERENCES

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems* **24**.

Anscombe, F. J. (1954). Fixed-sample size analysis of sequential observations. *Biometrics* **10** 89–100.

Barnard, G. A. (1947). Review of Abraham Wald's *Sequential Analysis*. *J. Amer. Statist. Assoc.* **42** 658–665.

Barron, A., Rissanen, J. and Yu, B. (1998). The Minimum Description Length principle in coding and modeling. *IEEE Trans. Inf. Theory* **44** 2743–2760. Special Commemorative Issue: Information Theory: 1948–1998.

Bates, S., Jordan, M. I., Sklar, M. and Soloff, J. (2022). Principal-agent hypothesis testing. Available at arXiv:2205.06812.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.* **100** 71–81.

Berger, J. O., Pericchi, L. R. and Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya, Ser. A* **60** 307–321. MR1718789

Breiman, L. (1961). Optimal gambling systems for favorable games. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 65–78. Univ. California Press, Berkeley, CA. MR0135630

Candès, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

Carney, D. R. My position on "Power Poses". Accessed 5 June 2022. Available at http://faculty.haas.berkeley.edu/dana_carney/pdf_my position on power poses.pdf.

Carney, D. R., Cuddy, A. J. C. and Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychol. Sci.* **21** 1363–1368. https://doi.org/10.1177/0956797610383437

Casgrain, P., Larsson, M. and Ziegel, J. (2022). Anytime-valid sequential testing for elicitable functionals via supermartingales. Available at arXiv:2204.05680.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

Choe, Y. J. and Ramdas, A. (2023). Comparing sequential forecasters. *Oper. Res.* To appear. Available at arXiv:2110.00115.

Chowdhury, S. R. and Gopalan, A. (2017). On kernelized multi-armed bandits. In *International Conference on Machine Learning* 844–853. PMLR.

Cover, T. M. (1974). Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin. Technical Report, No. 12. Stanford Univ., Stanford, CA.

Aaditya Ramdas is Assistant Professor, Statistics and Data Science, and Machine Learning, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: aramdas@cmu.edu). Peter Grünwald is Head, Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, and Professor of Statistics at Leiden University, The Netherlands (e-mail: pdg@cwi.nl). Vladimir Vovk is Professor, Computer Science, Royal Holloway, University of London, UK (e-mail: v.vovk@rhul.ac.uk). Glenn Shafer is University Professor, Rutgers University, Piscataway, New Jersey 08854-8019, USA (e-mail: gshafer@business.rutgers.edu).

Cox, D. R. (1952). Sequential tests for composite hypotheses. *Proc. Camb. Philos. Soc.* **48** 290–299. MR0047292 https://doi.org/10.1017/s030500410002764x

Crane, H. and Shafer, G. (2020). Risk is random: The magic of the d'Alembert. Available at: http://www.probabilityandfinance.com/articles/57.pdf.

Darling, D. A. and Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proc. Natl. Acad. Sci. USA* **58** 66–68. MR0215406 https://doi.org/10.1073/pnas.58.1.66

Darling, D. A. and Robbins, H. (1968). Some nonparametric sequential tests with power one. *Proc. Natl. Acad. Sci. USA* **61** 804–809. MR0238437 https://doi.org/10.1073/pnas.61.3.804

Dawid, A. P. (1984). Statistical theory. The prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292. MR0763811 https://doi.org/10.2307/2981683

Dawid, A. P., de Rooij, S., Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Insuring against loss of evidence in game-theoretic probability. *Statist. Probab. Lett.* **81** 157–162. MR2740080 https://doi.org/10.1016/j.spl.2010.10.013

de Heide, R. and Grünwald, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychon. Bull. Rev.* **28** 795–812. https://doi.org/10.3758/s13423-020-01803-x

Delyon, B. (2009). Exponential inequalities for sums of weakly dependent variables. *Electron. J. Probab.* **14** 752–779. MR2495559 https://doi.org/10.1214/EJP.v14-636

de la Peña, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Probab.* **27** 537–564. MR1681153 https://doi.org/10.1214/aop/1022677271

Dimitrov, V., Shafer, G. and Zhang, T. (2022). The martingale index. Available at: http://www.probabilityandfinance.com/articles/61.pdf.

Duan, B., Ramdas, A., Balakrishnan, S. and Wasserman, L. (2020). Interactive martingale tests for the global null. *Electron. J. Stat.* **14** 4489–4551. MR4194269 https://doi.org/10.1214/20-EJS1790

Duan, B., Ramdas, A. and Wasserman, L. (2022). Interactive rank testing by betting. In *Proceedings First Conference on Causal Learning and Reasoning* PMLR.

Dubins, L. E. and Savage, L. J. (1965). A Tchebycheff-like inequality for stochastic processes. *Proc. Natl. Acad. Sci. USA* **53** 274–275. MR0182042 https://doi.org/10.1073/pnas.53.2.274

Dunn, R., Ramdas, A., Balakrishnan, S. and Wasserman, L. (2023). Gaussian universal likelihood ratio testing. *Biometrika* **110** 319–337. MR4588356 https://doi.org/10.1093/biomet/asac064

Edwards, A. W. F. (1992). *Likelihood*, Expanded ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1191161

Efron, B. (1969). Student's t-test under symmetry conditions. *J. Amer. Statist. Assoc.* **64** 1278–1302. MR0251826

Fan, X., Grama, I. and Liu, Q. (2015). Exponential inequalities for martingales with applications. *Electron. J. Probab.* **20** 1–22. MR3311214 https://doi.org/10.1214/EJP.v20-3496

Feller, W. K. (1940). Statistical aspects of ESP. *J. Parapsychol.* **4** 271–298. MR0004461

Gangrade, A., Rinaldo, A. and Ramdas, A. (2023). A sequential test for log-concavity. ArXiv preprint. Available at arXiv:2301.03542.

Grünwald, P. (2022). Beyond Neyman–Pearson. Available at arXiv:2205.00901.

Grünwald, P., De Heide, R. and Koolen, W. (2023). Safe testing. *J. Roy. Statist. Soc. Ser. B*. To appear, with discussion.

Grünwald, P., Henzi, A. and Lardy, T. (2023). Anytime-valid tests of conditional independence under model-X. *J. Amer. Statist. Assoc.*

Grünwald, P. and Roos, T. (2020). Minimum description length revisited. *Int. J. Math. Ind.* **11**.

Grünwald, P. D. (2023). The e-posterior. *Philos. Trans. R. Soc. A* **381** 20220146. MR4590499

Hao, Y., Grünwald, P., Lardy, T., Long, L. and Adams, R. (2023). E-values for k-sample tests with exponential families. Available at arXiv:2303.0047.

Hendriks, H. (2018). Test martingales for bounded random variables. Available at arXiv:1801.09418.

Henzi, A., Arnold, S. and Ziegel, J. F. (2023). Sequentially valid tests for forecast calibration. *Ann. Appl. Stat.*

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

Hildreth, C. (1963). Bayesian statisticians and remote clients. *Econometrica* **31** 422–438.

Howard, S. R. and Ramdas, A. (2022). Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli* **28** 1704–1728. MR4411508 https://doi.org/10.3150/21-bej1388

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probab. Surv.* **17** 257–317. MR4100718 https://doi.org/10.1214/18-PS321

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021a). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.* **49** 1055–1080. MR4255119 https://doi.org/10.1214/20-aos1991

Ignatiadis, N., Wang, R. and Ramdas, A. (2022). E-values as unnormalized weights in multiple testing. *Biometrika*. To appear. https://doi.org/10.1093/biomet/asad057

Jamieson, K., Malloy, M., Nowak, R. and Bubeck, S. (2014). Lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory* 423–439. PMLR.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. MR0187257

Johari, R., Koomen, P., Pekelis, L. and Walsh, D. (2022). Always valid inference: Continuous monitoring of A/B tests. *Oper. Res.* **70** 1806–1821. MR4451064 https://doi.org/10.1287/opre.2021.2135

John, L. K., Loewenstein, G. and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23** 524–532. https://doi.org/10.1177/0956797611430953

Karampatziakis, N., Mineiro, P. and Ramdas, A. (2021). Off-policy confidence sequences. In *International Conference on Machine Learning* 5301–5310. PMLR.

Kaufmann, E. and Koolen, W. M. (2021). Mixture martingales revisited with applications to sequential tests and confidence intervals. *J. Mach. Learn. Res.* **22** 246. MR4353025

Kelly, J. L. Jr. (1956). A new interpretation of information rate. *Bell Syst. Tech. J.* **35** 917–926. MR0090494 https://doi.org/10.1002/j.1538-7305.1956.tb03809.x

Lai, T. L. (1976a). On confidence sequences. *Ann. Statist.* **4** 265–280. MR0395103

Lhéritier, A. and Cazals, F. (2018). A sequential non-parametric multivariate two-sample test. *IEEE Trans. Inf. Theory* **64** 3361–3370. MR3798382 https://doi.org/10.1109/TIT.2018.2800658

Li, J. Q. (1999). Estimation of Mixture Models Ph.D. thesis Yale Univ. New Haven, CT.

Li, J. Q. and Barron, A. R. (2000). Mixture density estimation. In *Advances in Neural Information Processing Systems* **12** 279–285.

Maclean, L. C., Thorp, E. O. and Ziemba, W. T. (2010). Long-term capital growth: The good and bad properties of the Kelly and fractional Kelly capital growth criteria. *Quant. Finance* **10** 681–687. MR2741943 https://doi.org/10.1080/14697688.2010.506108

MANOLE, T. and RAMDAS, A. (2023). Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Trans. Inform. Theory*. **69** 4641–4658. MR4613565 https://doi.org/10.1109/TIT.2023.3250099

MINGXIU, H., CAPPELLERI, J. C. and GORDON LAN, K. K. (2007). Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clin. Trials* **4** 329–340.

NEISWANGER, W. and RAMDAS, A. (2021). Uncertainty quantification using martingales for misspecified Gaussian processes. In *Algorithmic Learning Theory* 963–982. PMLR.

ORABONA, F. and JUN, K.-S. (2021). Tight concentrations and confidence sequences from the regret of universal portfolio. Available at arXiv:2110.14099.

PACE, L. and SALVAN, A. (2020). Likelihood, replicability and Robbins' confidence sequences. *Int. Stat. Rev.* **88** 599–615. MR4180669 https://doi.org/10.1111/insr.12355

PANDEVA, T., BAKKER, T., NAESSETH, C. A. and FORRÉ, P. (2022). E-Valuating Classifier Two-Sample Tests.

PAWEL, S., LY, A. and WAGENMAKERS, E.-J. (2022). Evidential calibration of confidence intervals. Available at arXiv:2206.12290.

PÉREZ-ORTIZ, M. F., LARDY, T., DE HEIDE, R. and GRÜNWALD, P. (2022). E-statistics, group invariance and anytime valid testing. Available at arXiv:2208.07610.

PODKOPAEV, A., BLOEBAUM, P., KASIVISWANATHAN, S. and RAMDAS, A. (2023). Sequential kernelized independence testing. In *International Conference on Machine Learning*.

RAMDAS, A. and MANOLE, T. (2023). Randomized and exchangeable improvements of Markov's, Chebyshev's and Chernoff's inequalities. ArXiv preprint. Available at arXiv:2304.02611.

RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. Available at arXiv:2009.03167.

RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *Internat. J. Approx. Reason.* **141** 83–109. MR4364897 https://doi.org/10.1016/j.ijar.2021.06.017

RISSANEN, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Inf. Theory* **30** 629–636. MR0755791 https://doi.org/10.1109/TIT.1984.1056936

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535. MR0050246 https://doi.org/10.1090/S0002-9904-1952-09620-8

ROBBINS, H. (1970). Statistical methods related to the law of the iterated logarithm. *Ann. Math. Stat.* **41** 1397–1409. MR0277063 https://doi.org/10.1214/aoms/1177696786

ROBBINS, H. and SIEGMUND, D. (1974). The expected sample size of some tests of power one. *Ann. Statist.* **2** 415–436. MR0448750

ROUDER, J. N., SPECKMAN, P. L., SUN, D., MOREY, R. D. and IVERSON, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16** 225–237.

ROYALL, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm. Monographs on Statistics and Applied Probability* **71**. CRC Press, London. MR1629481

RUF, J., LARSSON, M., KOOLEN, W. M. and RAMDAS, A. (2022). A composite generalization of Ville's martingale theorem. Available at arXiv:2203.04485.

RUSHTON, S. (1950). On a sequential t-test. *Biometrika* **37** 326–333. MR0044080 https://doi.org/10.1093/biomet/37.3-4.326

SHAER, S., MAMAN, G. and ROMANO, Y. (2023). Model-free sequential testing for conditional independence via testing by betting. In *International Conference on Artificial Intelligence and Statistics*.

SHAFER, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *J. Roy. Statist. Soc. Ser. A* **184** 407–478. MR4255905 https://doi.org/10.1111/rssa.12647

SHAFER, G., SHEN, A., VERESHCHAGIN, N. and VOVK, V. (2011). Test martingales, Bayes factors and p-values. *Statist. Sci.* **26** 84–101. MR2849911 https://doi.org/10.1214/10-STS347

SHAFER, G. and VOVK, V. (2001a). *Probability and Finance: It's Only a Game! Wiley Series in Probability and Statistics. Financial Engineering Section*. Wiley Interscience, New York. MR1852450 https://doi.org/10.1002/0471249696

SHAFER, G. and VOVK, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ.

SHEKHAR, S. and RAMDAS, A. (2023a). Nonparametric two sample testing by betting. *IEEE Trans. Inform. Theory*. To appear. https://doi.org/10.1109/TIT.2023.3305867

SHEKHAR, S. and RAMDAS, A. (2023b). Sequential change detection via backward confidence sequences. In *International Conference on Machine Learning*.

SHIN, J., RAMDAS, A. and RINALDO, A. (2022). E-detectors: A nonparametric framework for online changepoint detection. Available at arXiv:2203.03532.

SPERTUS, J. V. and STARK, P. B. (2022). Sweeter than SUITE: Supermartingale stratified union-intersection tests of elections. In *International Joint Conference on Electronic Voting*.

TER SCHURE, J. and GRÜNWALD, P. (2022). ALL-IN meta-analysis: Breathing life into living systematic reviews. *F1000Res.* **11** 549. https://doi.org/10.12688/f1000research.74223.1

TER SCHURE, J., GRÜNWALD, P. and LY, A. (2021). Pandemic preparedness in data sharing; lessons learned from collaborating in a live meta-analysis. *STAtOR* **24** 47–52.

TER SCHURE, J., PEREZ-ORTIZ, M. F., LY, A. and GRÜNWALD, P. (2021). The safe log rank test: Error control under continuous monitoring with unlimited horizon. Available at arXiv:1906.07801.

TURING, A. M. (1941). The Applications of Probability to Cryptography. UK National Archives, HW 25/37. See arXiv:1505.04714 for a version set in Latex.

TURNER, R. and GRÜNWALD, P. (2023a). Anytime-valid confidence intervals for contingency tables and beyond. *Statist. Probab. Lett* **198**.

TURNER, R. and GRÜNWALD, P. (2023b). Safe sequential testing and effect estimation in stratified count data. In *Annual AI and Statistics Conference*. PMLR.

TURNER, R., LY, A. and GRÜNWALD, P. (2021). Generic E-variables for exact sequential k-sample tests that allow for optional stopping. Available at arXiv:2106.02693.

TURNER, R., LY, A., ORTIZ-PEREZ, M.-F., TER SCHURE, J. and GRÜNWALD, P. (2022). R-package `safestats`. CRAN.

VILLE, J. (1939). *Etude Critique de la Notion de Collectif*. Gauthier-Villars, Paris.

VOLKHONSKIY, D., BURNAEV, E., NOURETDINOV, I., GAMMERMAN, A. and VOVK, V. (2017). Inductive conformal martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications* 132–153. PMLR.

VOVK, V. (2021). Testing randomness online. *Statist. Sci.* **36** 595–611. MR4323055 https://doi.org/10.1214/20-sts817

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2022). *Algorithmic Learning in a Random World*. Springer, Cham.

VOVK, V., NOURETDINOV, I. and GAMMERMAN, A. (2021). Conformal testing: Binary case with Markov alternatives. Available at arXiv:2111.01885.

VOVK, V. and WANG, R. (2021). E-values: Calibration, combination and applications. *Ann. Statist.* **49** 1736–1754. MR4298879 https://doi.org/10.1214/20-aos2020

WAGENMAKERS, E.-J., GRONAU, Q. F., DABLANDER, F. and ETZ, A. (2020). The support interval. *Erkenntnis* 1–13.

WALD, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat.* **16** 117–186. MR0013275 https://doi.org/10.1214/aoms/1177731118

WALD, A. (1947). *Sequential Analysis*. Wiley, New York. MR0020764

WANG, H. and RAMDAS, A. (2023a). The extended Ville's inequality for nonintegrable nonnegative supermartingales. ArXiv preprint. Available at arXiv:2304.01163.

WANG, H. and RAMDAS, A. (2023b). Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Process. Appl.* **163** 168–202. MR4610125 https://doi.org/10.1016/j.spa.2023.05.007

WANG, H. and RAMDAS, A. (2023c). Huber-robust confidence sequences. *26th International Conference on Artificial Intelligence and Statistics*.

WANG, R. and RAMDAS, A. (2022). False discovery rate control with e-values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 822–852. MR4460577

WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference. *Proc. Natl. Acad. Sci. USA* **117** 16880–16890. MR4242731 https://doi.org/10.1073/pnas.1922664117

WAUDBY-SMITH, I., ARBOUR, D., SINHA, R., KENNEDY, E. H. and RAMDAS, A. (2021). Time-uniform central limit theory and asymptotic confidence sequences. Available at arXiv:2103.06476.

WAUDBY-SMITH, I. and RAMDAS, A. (2020). Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems* **33** 20204–20214.

WAUDBY-SMITH, I. and RAMDAS, A. (2023). Estimating means of bounded random variables by betting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear, with discussion.

WAUDBY-SMITH, I., STARK, P. B. and RAMDAS, A. (2021). RiLACS: Risk limiting audits via confidence sequences. In *International Joint Conference on Electronic Voting* 124–139. Springer, Berlin.

WAUDBY-SMITH, I., WU, L., RAMDAS, A., KARAMPATZIAKIS, N. and MINEIRO, P. (2022). Anytime-valid off-policy inference for contextual bandits. ArXiv preprint. Available at arXiv:2210.10768.

XU, Z., WANG, R. and RAMDAS, A. (2021). A unified framework for bandit multiple testing. In *Advances in Neural Information Processing Systems* **34**.

XU, Z., WANG, R. and RAMDAS, A. (2022). Post-selection inference for e-value based confidence intervals. Available at arXiv:2203.12572.

ZHANG, Z., RAMDAS, A. and WANG, R. (2023). On the existence of powerful p-values and e-values for composite hypotheses. ArXiv preprint. Available at arXiv:2305.16539.

Statistical Science
2023, Vol. 38, No. 4, 602–620
https://doi.org/10.1214/23-STS892
© Institute of Mathematical Statistics, 2023

Replicability Across Multiple Studies

Marina Bogomolov and Ruth Heller

Abstract. Meta-analysis is routinely performed in many scientific disciplines. This analysis is attractive since discoveries are possible even when all the individual studies are underpowered. However, the meta-analytic discoveries may be entirely driven by signal in a single study, and thus nonreplicable. Although the great majority of meta-analyses carried out to date do not infer on the replicability of their findings, it is possible to do so. We provide a selective overview of analyses that can be carried out towards establishing replicability of the scientific findings. We describe methods for the setting where a single outcome is examined in multiple studies (as is common in systematic reviews of medical interventions), as well as for the setting where multiple studies each examine multiple features (as in genomics applications). We also discuss some of the current shortcomings and future directions.

Key words and phrases: Composite null, false discovery rate, meta-analysis, multiple hypothesis testing, replicability analysis.

REFERENCES

AMAR, D., VIZEL, A., LEVY, C. and SHAMIR, R. (2018). ADEPTUS: A discovery tool for disease prediction, enrichment and network analysis based on profiles from many diseases. *Bioinformatics* **34** 1959–1961.

ANDREASSEN, O. A., THOMPSON, W. K., SCHORK, A. J., RIPKE, S., MATTINGSDAL, M., KELSOE, J. R., KENDLER, K. S., O'DONOVAN, M. C., RUJESCU, D. et al. (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet*. **9**.

BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist*. **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

BENJAMIN, D., BEGER, J., JOHANNESSON, M. et al. (2018). Redefine statistical significance. *Nat. Hum. Behav*. **2** 6–10.

BENJAMINI, Y. and HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215–1222. MR2522270 https://doi.org/10.1111/j.1541-0420.2007.00984.x

BENJAMINI, Y., HELLER, R. and YEKUTIELI, D. (2009). Selective inference in complex research. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci*. **367** 4255–4271. MR2546387 https://doi.org/10.1098/rsta.2009.0127

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and YEKUTIELI, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc*. **100** 71–93. MR2156820 https://doi.org/10.1198/016214504000001907

BLANCHARD, G., NEUVIAL, P. and ROQUAIN, E. (2020). Post hoc confidence bounds on false positives using reference families. *Ann. Statist*. **48** 1281–1303. MR4124323 https://doi.org/10.1214/19-AOS1847

BOGOMOLOV, M. (2023). Testing partial conjunction hypotheses under dependency, with applications to meta-analysis. *Electron. J. Stat*. **17** 102–155. MR4533743 https://doi.org/10.1214/22-ejs2100

BOGOMOLOV, M. and HELLER, R. (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *J. Amer. Statist. Assoc*. **108** 1480–1492. MR3174723 https://doi.org/10.1080/01621459.2013.829002

BOGOMOLOV, M. and HELLER, R. (2018). Assessing replicability of findings across two studies of multiple features. *Biometrika* **105** 505–516. MR3842881 https://doi.org/10.1093/biomet/asy029

BOGOMOLOV, M. and HELLER, R. (2023). Supplement to "Replicability across multiple studies." https://doi.org/10.1214/23-STS892SUPP

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

CHUNG, D., YANG, C., LI, C., GELERNTER, J. and ZHAO, H. (2014). GPA: A statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet*. **10** e1004787.

DJORDJILOVIĆ, V., PAGE, C. M., GRAN, J. M., NØST, T. H., SANDANGER, T. M., VEIERØD, M. B. and THORESEN, M. (2019). Global test for high-dimensional mediation: Testing groups of potential mediators. *Stat. Med*. **38** 3346–3360. MR3979813 https://doi.org/10.1002/sim.8199

EFRON, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics (IMS) Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758 https://doi.org/10.1017/CBO9780511761362

Marina Bogomolov is Associate Professor, Faculty of Data and Decision Sciences, Technion—Israel Institute of Technology, Haifa, Israel (e-mail: marinabo@technion.ac.il). Ruth Heller is Professor, Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel (e-mail: ruheller@gmail.com).

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571 https://doi.org/10.1198/016214501753382129

FISHER, R. A. (1934). *Statistical Methods for Research Workers*, 5th ed.

FITHIAN, W., SUN, D. and TAYLOR, J. (2017). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.

FRANKE, A., MCGOVERN, D. P., BARRETT, J. C., WANG, K., RADFORD-SMITH, G. L., AHMAD, T., LEES, C. W., BALSCHUN, T., LEE, J. et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42** 1118–1125.

GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 499–517. MR1924303 https://doi.org/10.1111/1467-9868.00347

GOEMAN, J. J. and SOLARI, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26** 584–597. MR2951390 https://doi.org/10.1214/11-STS356

GOODMAN, S. N., FANELLI, D. and IOANNIDIS, J. P. (2016). What does research reproducibility mean? *Sci. Transl. Med.* **8**: 341ps12–341ps12.

HEDGES, L. V. and SCHAUER, J. M. (2019a). Consistency of effects is important in replication: Rejoinder to Mathur and VanderWeele (2019) reply. *Psychol. Methods* **24** 576–577.

HEDGES, L. V. and SCHAUER, J. M. (2019b). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychol. Methods* **24** 557–570.

HELD, L., MICHELOUD, C. and BALABDAOUI, F. (2022). A statistical framework for replicability. arXiv preprint. Available at arXiv:2207.00464.

HELLER, R., BOGOMOLOV, M. and BENJAMINI, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proc. Natl. Acad. Sci. USA* **111** 16262–16267.

HELLER, R., GOLLAND, Y., MALACH, R. and BENJAMINI, Y. (2007). Conjunction group analysis: An alternative to mixed/random effect analysis. *NeuroImage* **37** 1178–1185.

HELLER, R. and ROSSET, S. (2021). Optimal control of false discovery criteria in the two-group model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 133–155. MR4220987 https://doi.org/10.1111/rssb.12403

HELLER, R. and SOLARI, A. (2023). Simultaneous directional inference. arXiv preprint. Available at arXiv:2301.01653.

HELLER, R. and YEKUTIELI, D. (2014). Replicability analysis for genome-wide association studies. *Ann. Appl. Stat.* **8** 481–498. MR3191999 https://doi.org/10.1214/13-AOAS697

HIGGINS, J., THOMAS, J., CHANDLER, J., CUMPSTON, M., LI, T., PAGE, M. and WELCH, V. (2022). *Cochrane Handbook for Systematic Reviews of Interventions*, *Version* 6.3, (updated February 2022).

HOANG, A.-T. and DICKHAUS, T. (2022). Combining independent p-values in replicability analysis: A comparative study. *J. Stat. Comput. Simul.* **92** 2184–2204. MR4437521 https://doi.org/10.1080/00949655.2021.2022678

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

HOMMEL, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75** 383–386.

HUGHES, D., JUDGE, C., MURPHY, R., LOUGHLIN, E., COSTELLO, M., WHITELEY, W., BOSCH, J., O'DONNELL, M. J. and CANAVAN, M. (2020). Association of blood pressure lowering with incident dementia or cognitive impairment: A systematic review and meta-analysis. *JAMA* **323** 1934–1944.

HUNG, K. and FITHIAN, W. (2020). Statistical methods for replicability assessment. *Ann. Appl. Stat.* **14** 1063–1087. MR4152124 https://doi.org/10.1214/20-AOAS1336

IOANNIDIS, J. (2005). Why most published research findings are false. *PLoS Med.* **2** 696–701.

JALJULI, I., BENJAMINI, Y., SHENHAV, L., PANAGIOTOU, O. A. and HELLER, R. (2022). Quantifying replicability and consistency in systematic reviews. *Stat. Biopharm. Res.* 1–14.

JONES, L. V. and TUKEY, J. W. (2000). A sensible formulation of the significance test. *Psychol. Methods* **5** 411.

KARMAKAR, B. and SMALL, D. S. (2020). Assessment of the extent of corroboration of an elaborate theory of a causal hypothesis using partial conjunctions of evidence factors. *Ann. Statist.* **48** 3283–3311. MR4185809 https://doi.org/10.1214/19-AOS1929

KIDD, K. K., PAKSTIS, A. J., SPEED, W. C. and KIDD, J. R. (2004). Understanding human DNA sequence variation. *J. Hered.* **95** 406–420. https://doi.org/10.1093/jhered/esh060

LAWLOR, D. A., TILLING, K. and DAVEY SMITH, G. (2017). Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* **45** 1866–1886.

LI, S., SESIA, M., ROMANO, Y., CANDÈS, E. and SABATTI, C. (2022). Searching for robust associations with a multi-environment knockoff filter. *Biometrika* **109** 611–629. MR4472838 https://doi.org/10.1093/biomet/asab055

LIU, Z., SHEN, J., BARFIELD, R., SCHWARTZ, J., BACCARELLI, A. A. and LIN, X. (2022). Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *J. Amer. Statist. Assoc.* **117** 67–81. MR4399068 https://doi.org/10.1080/01621459.2021.1914634

MARIGORTA, U. M., RODRIGUEZ, J. A., GIBSON, G. and NAVARRO, A. (2018). Replicability and prediction: Lessons and challenges from gwas. *Trends Genet.* **34** 504–517.

MATHUR, M. B. and VANDERWEELE, T. J. (2019). New metrics for meta-analyses of heterogeneous effects. *Stat. Med.* **38** 1336–1342. MR3920618 https://doi.org/10.1002/sim.8057

NAKAGOME, S., MANO, S., KOZLOWSKI, L., BUJNICKI, J. M., SHIBATA, H., FUKUMAKI, Y., KIDD, J. R., KIDD, K. K., KAWAMURA, S. et al. (2012). Crohn's disease risk alleles on the NOD2 locus have been maintained by natural selection on standing variation. *Mol. Biol. Evol.* **29** 1569–1585.

NOSEK, B. A., HARDWICKE, T. E., MOSHONTZ, H., ALLARD, A., CORKER, K. S., DREBER, A., FIDLER, F., HILGARD, J., STRUHL, M. K. et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **73** 719–748.

OWEN, A. B. (2009). Karl Pearson's meta-analysis revisited. *Ann. Statist.* **37** 3867–3892. MR2572446 https://doi.org/10.1214/09-AOS697

PANAGIOTOU, O. A., JALJULI, I. and HELLER, R. (2020). Replicability of treatment effect in study of blood pressure lowering with dementia. *JAMA* **324** 1465–1466.

PATIL, P., PENG, R. D. and LEEK, J. T. (2019). A visual tool for defining reproducibility and replicability. *Nat. Hum. Behav.* **3** 650–652. https://doi.org/10.1038/s41562-019-0629-z

PAWEL, S. and HELD, L. (2022). The sceptical Bayes factor for the assessment of replication success. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 879–911. MR4460579

ROSENBAUM, P. (2022). *Replication and Evidence Factors in Observational Studies*. Taylor & Francis, London.

ROSENBAUM, P. R. (2001). Replicating effects and biases. *Amer. Statist.* **55** 223–227. MR1963397 https://doi.org/10.1198/000313001317098220

ROSENBAUM, P. R. (2010). Evidence factors in observational studies. *Biometrika* **97** 333–345. MR2650742 https://doi.org/10.1093/biomet/asq019

Roy, S., Bogomolov, M., Heller, R., Claridge, A. M., Beeson, T. and Small, D. S. (2022). Protocol for an observational study on the effects of giving births from unintended pregnancies on later life physical and mental health. arXiv preprint. Available at arXiv:2210.05169.

Saad, A., Yekutieli, D., Lev-Ran, S., Gross, R. and Guyatt, G. (2019). Getting more out of meta-analyses: A new approach to meta-analysis in light of unexplained heterogeneity. *J. Clin. Epidemiol.* **107** 101–106. https://doi.org/10.1016/j.jclinepi.2018.11.023

Sampson, J. N., Boca, S. M., Moore, S. C. and Heller, R. (2018). FWER and FDR control when testing multiple mediators. *Bioinformatics* **34** 2418–2424.

Sesia, M., Sabatti, C. and Candès, E. J. (2019). Gene hunting with hidden Markov model knockoffs. *Biometrika* **106** 1–18. MR3912377 https://doi.org/10.1093/biomet/asy033

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* **26** 559–569.

Sofer, T., Heller, R., Bogomolov, M., Avery, C. L., Graff, M., North, K. E., Reiner, A. P., Thornton, T. A., Rice, K. et al. (2017). A powerful statistical framework for generalization testing in GWAS, with application to the HCHS/SOL. *Genet. Epidemiol.* **41** 251–258.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398 https://doi.org/10.1214/aos/1074290335

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657 https://doi.org/10.1198/016214507000000545

Sun, W. and Wei, Z. (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *J. Amer. Statist. Assoc.* **106** 73–88. MR2816703 https://doi.org/10.1198/jasa.2011.ap09587

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statist. Sci.* 100–116.

Wang, J., Gui, L., Su, W. J., Sabatti, C. and Owen, A. B. (2022). Detecting multiple replicating signals using adaptive filtering procedures. *Ann. Statist.* **50** 1890–1909. MR4474476 https://doi.org/10.1214/21-aos2139

Wang, J. and Owen, A. B. (2019). Admissibility in partial conjunction testing. *J. Amer. Statist. Assoc.* **114** 158–168. MR3941245 https://doi.org/10.1080/01621459.2017.1385465

Wang, P. and Zhu, W. (2019). Replicability analysis in genome-wide association studies via Cartesian hidden Markov models. *BMC Bioinform.* **20** 146.

Xiang, D., Zhao, S. D. and Cai, T. T. (2019). Signal classification for the integrative analysis of multiple sequences of large-scale multiple tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 707–734. MR3997098

Xie, J., Cai, T. T., Maris, J. and Li, H. (2011). Optimal false discovery rate control for dependent data. *Stat. Interface* **4** 417–430. MR2868825 https://doi.org/10.4310/SII.2011.v4.n4.a1

Zhao, Q., Small, D. S. and Rosenbaum, P. R. (2018). Cross-screening in observational studies that test many hypotheses. *J. Amer. Statist. Assoc.* **113** 1070–1084. MR3862340 https://doi.org/10.1080/01621459.2017.1407770

Zhao, Q., Small, D. S. and Su, W. (2019). Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *J. Amer. Statist. Assoc.* **114** 1291–1304. MR4011780 https://doi.org/10.1080/01621459.2018.1497499

Statistical Science
2023, Vol. 38, No. 4, 621–639
https://doi.org/10.1214/23-STS904
© Institute of Mathematical Statistics, 2023

Replication Success Under Questionable Research Practices—a Simulation Study

Francesca Freuli, Leonhard Held and Rachel Heyard

Abstract. Increasing evidence suggests that the reproducibility and replicability of scientific findings is threatened by researchers employing questionable research practices (QRPs) in order to achieve statistically significant results. Numerous metrics have been developed to determine replication success but it has not yet been investigated how well those metrics perform in the presence of QRPs. This paper aims to compare the performance of different metrics quantifying replication success in the presence of four types of QRPs: cherry picking of outcomes, questionable interim analyses, questionable inclusion of covariates, and questionable subgroup analyses. Our results show that the metric based on the version of the sceptical p-value that is recalibrated in terms of effect size performs better in maintaining low values of overall type-I error rate, but often requires larger replication sample sizes compared to metrics based on significance, the controlled version of the sceptical p-value, meta-analysis or Bayes factors, especially when severe QRPs are employed.

Key words and phrases: Questionable research practices, replication success, simulation study, type-I error rate, power, rejection ratio.

REFERENCES

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P. and Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS ONE* **12** e0172792. https://doi.org/10.1371/journal.pone.0172792

Anderson, S. F. and Kelley, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychol. Methods*. https://doi.org/10.1037/met0000520

Anderson, S. F. and Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychol. Methods* **21** 1–12. https://doi.org/10.1037/met0000051

Bayarri, M. J., Benjamin, D. J., Berger, J. O. and Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *J. Math. Psych.* **72** 90–103. MR3506028 https://doi.org/10.1016/j.jmp.2015.12.007

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature* **568** 435–435.

Boulesteix, A.-L., Lauer, S. and Eugster, M. J. A. (2013). A plea for neutral comparison studies in computational sciences. *PLoS ONE* **8** e61562. https://doi.org/10.1371/journal.pone.0061562

Brookes, S. T., Whitely, E., Egger, M., Smith, G. D., Mulheran, P. A. and Peters, T. J. (2004). Subgroup analyses in randomized trials: Risks of subgroup-specific analyses; power and sample size for the interaction test. *J. Clin. Epidemiol.* **57** 229–236. https://doi.org/10.1016/j.jclinepi.2003.08.009

Burton, A., Altman, D. G., Royston, P. and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Stat. Med.* **25** 4279–4292. MR2307592 https://doi.org/10.1002/sim.2673

Christian, K., Johnstone, C., Larkins, J.-A., Wright, W. and Doran, M. R. (2021). A survey of early-career researchers in Australia. *eLife* **10**. https://doi.org/10.7554/eLife.60613

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* **349**.

Cousins, R. D. (2007). Annotated bibliography of some papers on combining significances or p-values.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E. and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife* **10**. https://doi.org/10.7554/eLife.71601

Freuli, F., Held, L. and Heyard, R. (2023). Supplement to "Replication success under questionable research practices—a simulation study." https://doi.org/10.1214/23-STS904SUPP

Gopalakrishna, G., Riet, G. T., Vink, G., Stoop, I., Wicherts, J. M. and Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in

Francesca Freuli is Ph.D. Student, Department of Psychology and Cognitive Science, University of Trento, Trento, Italy (e-mail: francesca.freuli@unitn.it). Leonhard Held is Professor, Center for Reproducible Science, Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland (e-mail: leonhard.held@uzh.ch). Rachel Heyard is Postdoctoral Fellow, Center for Reproducible Science, Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland (e-mail: rachel.heyard@uzh.ch).

The Netherlands. *PLoS ONE* **17** e0263023. https://doi.org/10.1371/journal.pone.0263023

GRIEVE, A. P. (2016). Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. *Pharm. Stat.* **15** 96–108. https://doi.org/10.1002/pst.1736

HEAD, M. L., HOLMAN, L., LANFEAR, R., KAHN, A. T. and JENNIONS, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biol.* **13** e1002106. https://doi.org/10.1371/journal.pbio.1002106

HEDGES, L. V. and SCHAUER, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. *J. Educ. Behav. Stat.* **44** 543–570.

HELD, L. (2020). A new standard for the analysis and design of replication studies. *J. Roy. Statist. Soc. Ser. A* **183** 431–448. MR4052785

HELD, L., MATTHEWS, R., OTT, M. and PAWEL, S. (2022). Reverse-Bayes methods for evidence assessment and research synthesis. *Res. Synth. Methods* **13** 295–314. https://doi.org/10.1002/jrsm.1538

HELD, L., MICHELOUD, C. and PAWEL, S. (2022). The assessment of replication success based on relative effect size. *Ann. Appl. Stat.* **16** 706–720. MR4438808 https://doi.org/10.1214/21-aoas1502

HELD, L. and OTT, M. (2018). On p-values and Bayes factors. *Annu. Rev. Stat. Appl.* **5** 393–422. MR3774753 https://doi.org/10.1146/annurev-statistics-031017-100307

JOHN, L. K., LOEWENSTEIN, G. and PRELEC, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23** 524–532. https://doi.org/10.1177/0956797611430953

KIRKHAM, J. J., ALTMAN, D. G., CHAN, A.-W., GAMBLE, C., DWAN, K. M. and WILLIAMSON, P. R. (2018). Outcome reporting bias in trials: A methodological approach for assessment and adjustment in systematic reviews. *BMJ* **362** k3802. https://doi.org/10.1136/bmj.k3802

KIRKHAM, J. J., DWAN, K. M., ALTMAN, D. G., GAMBLE, C., DODD, S., SMYTH, R. and WILLIAMSON, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* **340** c365. https://doi.org/10.1136/bmj.c365

KLEIN, R. A., RATLIFF, K. A., VIANELLO, M., ADAMS, R. B., BAHNÍK, Š., BERNSTEIN, M. J., BOCIAN, K., BRANDT, M. J., BROOKS, B. et al. (2014). Investigating variation in replicability. *Soc. Psychol.* **45** 142–152.

LY, A., ETZ, A., MARSMAN, M. and WAGENMAKERS, E.-J. (2018). Replication Bayes factors from evidence updating. *Behav. Res. Methods* **51** 2498–2508.

MATTHEWS, J. N. S. (2006). *Introduction to Randomized Controlled Clinical Trials*, 2nd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. MR2261274 https://doi.org/10.1201/9781420011302

MAYO-WILSON, E., LI, T., FUSCO, N., BERTIZZOLO, L., CANNER, J. K., COWLEY, T., DOSHI, P., EHMSEN, J., GRESHAM, G. et al. (2017). Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *J. Clin. Epidemiol.* **91** 95–110.

MICHELOUD, C., BALABDAOUI, F. and HELD, L. (2023). Assessing replicability with the sceptical p-value: Type-I error control and sample size planning. *Stat. Neerl.*.

MICHELOUD, C. and HELD, L. (2022). Power calculations for replication studies. *Statist. Sci.* **37** 369–379. MR4444372 https://doi.org/10.1214/21-sts828

MORAN, C., RICHARD, A., WILSON, K., TWOMEY, R. and COROIU, A. (2022). I know it's bad, but I have been pressured into it: Questionable research practices among psychology students in Canada. *Can. Psychol.*

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. MR3937487 https://doi.org/10.1002/sim.8086

MURADCHANIAN, J., HOEKSTRA, R., KIERS, H. and VAN RAVENZWAAIJ, D. (2021). How best to quantify replication success? A simulation study on the comparison of replication success metrics. *R. Soc. Open Sci.* **8** 201697.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, MEDICINE (2019). *Reproducibility and Replicability in Science*. The National Academies Press. Washington, DC.

NOSEK, B. A., SPIES, J. R. and MOTYL, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* **7** 615–631. https://doi.org/10.1177/1745691612459058

PAWEL, S., CONSONNI, G. and HELD, L. (2023). Bayesian approaches to designing replication studies. *Psychol. Methods*, Accepted.

PAWEL, S. and HELD, L. (2022). The sceptical Bayes factor for the assessment of replication success. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 879–911. MR4460579

PAWEL, S., KOOK, L. and REEVE, K. (2023). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biom. J.* e2200091. https://doi.org/10.1002/bimj.202200091

POCOCK, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64** 191–199.

RABELO, A. L., FARIAS, J. E., SARMET, M. M., JOAQUIM, T. C., HOERSTING, R. C., VICTORINO, L., MODESTO, J. G. and PILATI, R. (2020). Questionable research practices among Brazilian psychological researchers: Results from a replication study and an international comparison. *Int. J. Psychol.* **55** 674–683.

ROETTGER, T. B. (2019). Researcher degrees of freedom in phonetic research. *Lab. Phonol.* **10**.

ROSENKRANZ, G. (2019). *Exploratory Subgroup Analyses in Clinical Research*. Wiley, New York.

ROSENKRANZ, G. K. (2023). A generalization of the two trials paradigm. *Ther. Innov. Regul. Sci.* **57** 316–320. https://doi.org/10.1007/s43441-022-00471-4

SAGARIN, B. J., AMBLER, J. K. and LEE, E. M. (2014). An ethical approach to peeking at data. *Perspect. Psychol. Sci.* **9** 293–304.

SCHMIDT, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* **13** 90–100.

SENN, S. (2021). *Statistical Issues in Drug Development*, 3rd ed. Wiley, New York.

SIMMONS, J. P., NELSON, L. D. and SIMONSOHN, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22** 1359–1366.

SIMONSOHN, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* **26** 559–569. https://doi.org/10.1177/0956797614567341

STEFAN, A. M. and SCHÖNBRODT, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *R. Soc. Open Sci.* **10**.

ULRICH, R. and MILLER, J. (2020). Questionable research practices may have little effect on replicability. *eLife* **9**. https://doi.org/10.7554/eLife.58237

VAN ZWET, E. W. and CATOR, E. A. (2021). The significance filter, the winner's curse and the need to shrink. *Stat. Neerl.* **75** 437–452. MR4374073 https://doi.org/10.1111/stan.12241

VERHAGEN, J. and WAGENMAKERS, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* **143** 1457–1475. https://doi.org/10.1037/a0036731

WANG, Y. A., SPARKS, J., GONZALES, J. E., HESS, Y. D. and LEDGERWOOD, A. (2017). Using independent covariates in experimental designs: Quantifying the trade-off between power boost and type I error inflation. *J. Exp. Soc. Psychol.* **72** 118–124.

WICHERTS, J. M., VELDKAMP, C. L. S., AUGUSTEIJN, H. E. M., BAKKER, M., VAN AERT, R. C. M. and VAN ASSEN, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid. *Front. Psychol.* **7** 1832. https://doi.org/10.3389/fpsyg.2016.01832

WOLFF, W., BAUMANN, L. and ENGLERT, C. (2018). Self-reports from behind the scenes: Questionable research practices and rates of replication in ego depletion research. *PLoS ONE* **13** e0199554. https://doi.org/10.1371/journal.pone.0199554

Statistical Science
2023, Vol. 38, No. 4, 640–654
https://doi.org/10.1214/23-STS890

Methods for Integrating Trials and Non-experimental Data to Examine Treatment Effect Heterogeneity

Carly Lupton Brantner, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano and Elizabeth A. Stuart

Abstract. Estimating treatment effects conditional on observed covariates can improve the ability to tailor treatments to particular individuals. Doing so effectively requires dealing with potential confounding, and also enough data to adequately estimate effect moderation. A recent influx of work has looked into estimating treatment effect heterogeneity using data from multiple randomized controlled trials and/or observational datasets. With many new methods available for assessing treatment effect heterogeneity using multiple studies, it is important to understand which methods are best used in which setting, how the methods compare to one another, and what needs to be done to continue progress in this field. This paper reviews these methods broken down by data setting: aggregate-level data, federated learning, and individual participant-level data. We define the conditional average treatment effect and discuss differences between parametric and nonparametric estimators, and we list key assumptions, both those that are required within a single study and those that are necessary for data combination. After describing existing approaches, we compare and contrast them and reveal open areas for future research. This review demonstrates that there are many possible approaches for estimating treatment effect heterogeneity through the combination of datasets, but that there is substantial work to be done to compare these methods through case studies and simulations, extend them to different settings, and refine them to account for various challenges present in real data.

Key words and phrases: Treatment effect heterogeneity, combining data, generalizability and reproducibility.

REFERENCES

ABREVAYA, J., HSU, Y.-C. and LIELI, R. P. (2015). Estimating conditional average treatment effects. *J. Bus. Econom. Statist.* **33** 485–505. MR3416596 https://doi.org/10.1080/07350015.2014.975555

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

AUDIGIER, V., WHITE, I. R., JOLANI, S., DEBRAY, T. P. A., QUARTAGNO, M., CARPENTER, J., VAN BUUREN, S. and RESCHE-RIGON, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statist. Sci.* **33** 160–183. MR3797708 https://doi.org/10.1214/18-STS646

BARON, R. M. and KENNY, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

BERLIN, J. A., SANTANNA, J., SCHMID, C. H., SZCZECH, L. A.,

Carly Lupton Brantner is a PhD Candidate, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA (e-mail: clupton1@jhu.edu). Ting-Hsuan Chang is a PhD Student, Department of Biostatistics, Columbia Mailman School of Public Health, New York, New York 10032, USA (e-mail: tc3255@cumc.columbia.edu). Trang Quynh Nguyen is an Associate Scientist, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA (e-mail: trang.nguyen@jhu.edu). Hwanhee Hong is an Associate Professor, Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina 27710, USA (e-mail: hwanhee.hong@duke.edu). Leon Di Stefano is a PhD Candidate, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA (e-mail: lds@jhu.edu). Elizabeth A. Stuart is Professor, Departments of Biostatistics, Mental Health, and Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA (e-mail: estuart@jhu.edu).

Feldman, H. I. and Anti-Lymphocyte Antibody Induction Therapy Study Group (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Stat. Med.* **21** 371–387. https://doi.org/10.1002/sim.1023

Brantner, C. L., Nguyen, T. Q., Tang, T., Zhao, C., Hong, H. and Stuart, E. A. (2023a). Comparing machine learning methods for estimating heterogeneous treatment effects by combining data from multiple randomized controlled trials. arXiv preprint. Available at arXiv:2303.16299.

Brantner, C. L., Chang, T.-H., Nguyen, T. Q., Hong, H., Di Stefano, L. and Stuart, E. A. (2023b). Supplement to "Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity." https://doi.org/10.1214/23-STS890SUPP

Brown, C. H., Sloboda, Z., Faggiano, F., Teasdale, B., Keller, F., Burkhart, G., Vigna-Taglianti, F., Howe, G., Masyn, K. et al. (2013). Methods for synthesizing findings on moderation effects across multiple randomized trials. *Prev. Sci.* **14** 144–156. https://doi.org/10.1007/s11121-011-0207-8

Burke, D. L., Ensor, J. and Riley, R. D. (2017). Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ. *Stat. Med.* **36** 855–875. MR3597661 https://doi.org/10.1002/sim.7141

Cheng, D. and Cai, T. (2021). Adaptive combination of randomized and observational data. Available at arXiv:2111.15012.

Colnet, B., Josse, J., Varoquaux, G. and Scornet, E. (2022). Causal effect on a target population: A sensitivity analysis to handle missing covariates. *J. Causal Inference* **10** 372–414. MR4512969 https://doi.org/10.1515/jci-2021-0059

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J., Josse, J. and Yang, S. (2021a). Causal inference methods for combining randomized trials and observational studies: A review. Available at arXiv:2011.08047.

Dagne, G. A., Brown, C. H., Howe, G., Kellam, S. G. and Liu, L. (2016). Testing moderation in network meta-analysis with individual participant data. *Stat. Med.* **35** 2485–2502. MR3513700 https://doi.org/10.1002/sim.6883

Dahabreh, I. J., Petito, L. C., Robertson, S. E., Hernán, M. A. and Steingrimsson, J. A. (2020). Towards causally interpretable meta-analysis: Transporting inferences from multiple studies to a target population. Available at arXiv:1903.11455.

Debray, T. P. A., Moons, K. G. M., Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. H. and Reitsma, J. B. (2015). Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Res. Synth. Methods* **6** 293–309. https://doi.org/10.1002/jrsm.1160

Debray, T. P. A., Schuit, E., Efthimiou, O., Reitsma, J. B., Ioannidis, J. P. A., Salanti, G., Moons, K. G. M. and Workpackage, G. (2018). An overview of methods for network meta-analysis using individual participant data: When do benefits arise? *Stat. Methods Med. Res.* **27** 1351–1364. MR3777761 https://doi.org/10.1177/0962280216660741

Donegan, S., Williamson, P., D'Alessandro, U. and Tudur Smith, C. (2012). Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: Individual patient-level covariates versus aggregate trial-level covariates. *Stat. Med.* **31** 3840–3857. MR3041777 https://doi.org/10.1002/sim.5470

Efthimiou, O., Debray, T. P. A., van Valkenhoef, G., Trelle, S., Panayidou, K., Moons, K., Reitsma, J. B., Shang, A. and Salanti, G. (2016). GetReal in network meta-analysis: A review of the methodology. *Res. Synth. Methods* **7** 236–263. https://doi.org/10.1002/jrsm.1195

Enderlein, G. (1988). Fleiss, J. L.: The design and analysis of clinical experiments. *Biom. J.* **30** 304–304. https://doi.org/10.1002/bimj.4710300308

Gelman, A., Hill, J. and Vehtari, A. (2020). *Regression and Other Stories*. Cambridge Univ. Press, Cambridge.

Godolphin, P. J., White, I. R., Tierney, J. F. and Fisher, D. J. (2023). Estimating interactions and subgroup-specific treatment effects in meta-analysis without aggregation bias: A within-trial framework. *Res. Synth. Methods* **14** 68–78. https://doi.org/10.1002/jrsm.1590

Green, A. K., Trivedi, N., Hsu, J. J., Yu, N. L., Bach, P. B. and Chimonas, S. (2022). Despite the FDA's five-year plan, black patients remain inadequately represented in clinical trials for drugs: Study examines FDA's five-year action plan aimed at improving diversity in and transparency of pivotal clinical trials for newly-approved drugs. *Health Aff.* **41** 368–374. https://doi.org/10.1377/hlthaff.2021.01432

Han, L., Hou, J., Cho, K., Duan, R. and Cai, T. (2021). Federated Adaptive Causal Estimation (FACE) of target treatment effects. Available at arXiv:2112.09313.

Hatt, T., Berrevoets, J., Curth, A., Feuerriegel, S. and van der Schaar, M. (2022). Combining observational and randomized data for estimating heterogeneous treatment effects. arXiv preprint. Available at arXiv:2202.12891.

Hayward, R. A., Gagnier, J. J., Borenstein, M., VanDerHeijden, G. J. M. G., Dahabreh, I. J., Sun, X., Sauerbrei, W., Walsh, M., Ioannidis, J. P. A. et al. (2020). Instrument for the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses: Manual version 1.0.

Hong, H., Fu, H. and Carlin, B. P. (2018). Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1047–1069. MR3832263 https://doi.org/10.1111/rssc.12275

Hong, H., Fu, H., Price, K. L. and Carlin, B. P. (2015). Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment. *Stat. Med.* **34** 2794–2819. MR3375982 https://doi.org/10.1002/sim.6519

Hua, H., Burke, D. L., Crowther, M. J., Ensor, J., Tudur Smith, C. and Riley, R. D. (2017). One-stage individual participant data meta-analysis models: Estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information. *Stat. Med.* **36** 772–789. MR3597655 https://doi.org/10.1002/sim.7171

Jolani, S., Debray, T. P. A., Koffijberg, H., van Buuren, S. and Moons, K. G. M. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: A generalized approach using MICE. *Stat. Med.* **34** 1841–1863. MR3334696 https://doi.org/10.1002/sim.6451

Kallus, N., Puli, A. M. and Shalit, U. (2018). Removing hidden confounding by experimental grounding. Available at arXiv:1810.11646.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. Available at arXiv:2004.14497.

Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P. A., Patrick-Lake, B., Morton, S. et al. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. *Ann. Intern. Med.* **172** 35–45.

Kent, D. M., Rothwell, P. M., Ioannidis, J. P. A., Altman, D. G. and Hayward, R. A. (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. *Trials* **11** 85. https://doi.org/10.1186/1745-6215-11-85

KOVALCHIK, S. A. (2013). Aggregate-data estimation of an individual patient data linear random effects meta-analysis with a patient covariate-treatment interaction term. *Biostatistics* **14** 273–283. https://doi.org/10.1093/biostatistics/kxs035

KÜNZEL, S. R., SEKHON, J. S., BICKEL, P. J. and YU, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. In *Proceedings of the National Academy of Sciences* **116** 4156–4165.

LAMBERT, P. C., SUTTON, A. J., ABRAMS, K. R. and JONES, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J. Clin. Epidemiol.* **55** 86–94. https://doi.org/10.1016/S0895-4356(01)00414-0

MCCANDLESS, L. (2009). *Bayesian Methods for Data Analysis*, 3rd ed. Bradley P. Carlin and Thomas A. Louis, Chapman & Hall/CRC, Boca Raton, 2008. ISBN 9781584886976.

NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133 https://doi.org/10.1093/biomet/asaa076

PETRELLI, F. and BARNI, S. (2012). Surgery of primary tumors in stage IV breast cancer: An updated meta-analysis of published studies with meta-regression. *Med. Oncol.* **29** 3282–3290. https://doi.org/10.1007/s12032-012-0310-0

RILEY, R. D., LAMBERT, P. C., STAESSEN, J. A., WANG, J., GUEYFFIER, F., THIJS, L. and BOUTITIE, F. (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat. Med.* **27** 1870–1893. MR2420350 https://doi.org/10.1002/sim.3165

RILEY, R. D., STEWART, L. A. and TIERNEY, J. F. (2021). Individual participant data meta-analysis for healthcare research. *Individual Participant Data Meta-Analysis*: *A Handbook for Healthcare Research* **1–6**.

ROSENMAN, E., BASSE, G., OWEN, A. and BAIOCCHI, M. (2020). Combining observational and experimental datasets using shrinkage estimators. Available at arXiv:2002.06708.

ROSENMAN, E. T. R., OWEN, A. B., BAIOCCHI, M. and BANACK, H. R. (2022). Propensity score methods for merging observational and experimental datasets. *Stat. Med.* **41** 65–86. MR4376789 https://doi.org/10.1002/sim.9223

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701. https://doi.org/10.1037/h0037350

SAMARA, M. T., NIKOLAKOPOULOU, A., SALANTI, G. and LEUCHT, S. (2019). How many patients with schizophrenia do not respond to antipsychotic drugs in the short term? An analysis based on individual patient data from randomized controlled trials. *Schizophr. Bull.* **45** 639–646. https://doi.org/10.1093/schbul/sby095

SARAMAGO, P., SUTTON, A. J., COOPER, N. J. and MANCA, A. (2012). Mixed treatment comparisons using aggregate and individual participant level data. *Stat. Med.* **31** 3516–3536. MR3041828 https://doi.org/10.1002/sim.5442

SEO, M., WHITE, I. R., FURUKAWA, T. A., IMAI, H., VALGIMIGLI, M., EGGER, M., ZWAHLEN, M. and EFTHIMIOU, O. (2021). Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis. *Stat. Med.* **40** 1553–1573. MR4212329 https://doi.org/10.1002/sim.8859

SILVA, S., GUTMAN, B. A., ROMERO, E., THOMPSON, P. A., ALTMANN, A. and LORENZI, M. (2019). Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In 2019 *IEEE* 16*th International Symposium on Biomedical Imaging* (*ISBI* 2019) 270–274. IEEE, Los Alamitos, CA.

SIMMONDS, M. C. and HIGGINS, J. P. T. (2007). Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Stat. Med.* **26** 2982–2999. MR2370988 https://doi.org/10.1002/sim.2768

TAN, X., CHANG, C.-C. H. and TANG, L. (2021). A tree-based federated learning approach for personalized treatment effect estimation from heterogeneous data sources. Available at arXiv:2103.06261.

TERAMUKAI, S., MATSUYAMA, Y., MIZUNO, S. and SAKAMOTO, J. (2004). Individual patient-level and study-level meta-analysis for investigating modifiers of treatment effect. *Jpn. J. Clin. Oncol.* **34** 717–721. https://doi.org/10.1093/jjco/hyh138

THOMAS, D., RADJI, S. and BENEDETTI, A. (2014). Systematic review of methods for individual patient data meta-analysis with binary outcomes. *BMC Med. Res. Methodol.* **14**. https://doi.org/10.1186/1471-2288-14-79

TIERNEY, J. F., VALE, C., RILEY, R., SMITH, C. T., STEWART, L., CLARKE, M. and ROVERS, M. (2015). Individual participant data (IPD) meta-analyses of randomised controlled trials: Guidance on their use. *PLoS Med.* **12** e1001855. https://doi.org/10.1371/journal.pmed.1001855

TRIVEDI, M. H., RUSH, A. J., WISNIEWSKI, S. R., NIERENBERG, A. A., WARDEN, D., RITZ, L., NORQUIST, G., HOWLAND, R. H., LEBOWITZ, B. et al. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. *Am. J. Psychiatr.* **163** 28–40. https://doi.org/10.1176/appi.ajp.163.1.28

VO, T. V., HOANG, T. N., LEE, Y. and LEONG, T.-Y. (2021). Federated estimation of causal effects from observational data. Available at arXiv:2106.00456.

WU, L. and YANG, S. (2021). Integrative R-learner of heterogeneous treatment effects combining experimental and observational studies. In *First Conference on Causal Learning and Reasoning*.

XIE, F., CHAN, J. C. and MA, R. C. (2018). Precision medicine in diabetes prevention, classification and management. *J. Diabetes Investig.* **9** 998–1015. https://doi.org/10.1111/jdi.12830

YANG, Q., LIU, Y., CHENG, Y., KANG, Y., CHEN, T. and YU, H. (2022). *Federated Learning*. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **43**. Springer, Cham. Reprint of the 2020 original. MR4592510 https://doi.org/10.1007/978-3-031-01585-4

YANG, S., ZENG, D. and WANG, X. (2020). Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. Available at arXiv:2005.10579.

YANG, S., ZENG, D. and WANG, X. (2022). Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. Available at arXiv:2007.12922.

Statistical Science
2023, Vol. 38, No. 4, 655–671
https://doi.org/10.1214/23-STS899

Tracking Truth Through Measurement and the Spyglass of Statistics

Antonio Possolo

Abstract. The measurement of a quantity is reproducible when mutually independent, multiple measurements made of it yield mutually consistent measurement results, that is, when the measured values, after due allowance for their associated uncertainties, do not differ significantly from one another. Interlaboratory comparisons organized deliberately for the purpose, and meta-analyses that are structured so as to be fit for the same purpose, are procedures of choice to ascertain measurement reproducibility.

The realistic evaluation of measurement uncertainty is a key preliminary to the assessment of reproducibility because lack of reproducibility manifests itself as dispersion or variability of measured values in excess of what their associated uncertainties suggest that they should exhibit. For this reason, we review the distinctive traits of measurement in the physical sciences and technologies, including medicine, and discuss the meaning and expression of measurement uncertainty.

This contribution illustrates the application of statistical models and methods to quantify measurement uncertainty and to assess reproducibility in four concrete, real-life examples, in the process revealing that lack of reproducibility can be a consequence of one or more of the following: intrinsic differences between laboratories making measurements; choice of statistical model and of procedure for data reduction or of causes yet to be identified.

Despite the instances of lack of reproducibility that we review, and many others like them, the outlook is optimistic. First, because “lack of reproducibility is not necessarily bad news; it may herald new discoveries and signal scientific progress” (*Nat. Phys.* **16** (2020) 117–119). Second, and as the example about the measurement of the Newtonian constant of gravitation, G, illustrates, when faced with a reproducibility crisis the scientific community often engages in cooperative efforts to understand the root causes of the lack of reproducibility, leading to advances in scientific knowledge.

Key words and phrases: Avandia, common mean, fixed effect, COVID-19, Newtonian constant of gravitation, Rosiglitazone, dark uncertainty, heterogeneity, interlaboratory study, meta-analysis, random effects, repeatability, replicability, reproducibility, reproduction number, W boson.

REFERENCES

[1] Azzalini, A. (2014). *The Skew-Normal and Related Families*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **3**. Cambridge Univ. Press, Cambridge. With the collaboration of Antonella Capitanio. MR3468021 https://doi.org/10.1017/cbo9781139248891

[2] Baker, R. and Jackson, D. (2015). New models for describing outliers in meta-analysis. *Res. Synth. Methods* **7** 314–328. https://doi.org/10.1002/jrsm.1191

[3] Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48. https://doi.org/10.18637/jss.v067.i01

[4] Beauchamp, C. R., Camara, J. E., Carney, J., Choquette, S. J., Cole, K. D., DeRose, P. C., Duewer, D. L., Epstein, M. S., Kline, M. C. et al. (2021). *Metrological Tools for the Reference Materials and Reference Instruments of the NIST Materials Measurement Laboratory*. *NIST Special Publication* 260-136 (2021 *Edition*). National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.260-136-2021

Antonio Possolo is NIST Fellow and Chief Statistician, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, U.S.A. (e-mail: antonio.possolo@nist.gov).

[5] Bell, S. (1999). *A Beginner's Guide to Uncertainty of Measurement. Measurement Good Practice Guide* 11 (*Issue* 2). National Physical Laboratory, Teddington, Middlesex, United Kingdom. Amendments March 2001.

[6] BIPM (2019). *The International System of Units* (*SI*), 9th ed. International Bureau of Weights and Measures (BIPM), Sèvres, France.

[7] Birge, R. T. (1932). The calculation of errors by the method of least squares. *Phys. Rev.* **40** 207–227. https://doi.org/10.1103/PhysRev.40.207

[8] Blackman, R. B. and Tukey, J. W. (1958). The measurement of power spectra from the point of view of communications engineering. I. *Bell Syst. Tech. J.* **37** 185–282. MR0102897 https://doi.org/10.1002/j.1538-7305.1958.tb03874.x

[9] Blackwell, T., Brown, C. and Mosteller, F. (1991). Which denominator? In *Fundamentals of Exploratory Analysis of Variance* (D. C. Hoaglin, F. Mosteller and J. W. Tukey, eds.) **10** 252–294. Wiley, New York, NY.

[10] Bodnar, O. and Elster, C. (2014). On the adjustment of inconsistent data using the Birge ratio. *Metrologia* **51** 516–521. https://doi.org/10.1088/0026-1394/51/5/516

[11] Borenstein, M., Hedges, L. V., Higgins, J. P. T. and Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* **1** 97–111. https://doi.org/10.1002/jrsm.12

[12] Bradburn, M. J., Deeks, J. J., Berlin, J. A. and Localio, A. R. (2007). Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. *Stat. Med.* **26** 53–77. MR2312699 https://doi.org/10.1002/sim.2528

[13] Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80** 1–28. https://doi.org/10.18637/jss.v080.i01

[14] Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* **10** 395–411. https://doi.org/10.32614/RJ-2018-017

[15] Campagnari, C. and Mulders, M. (2022). An upset to the standard model. *Science* **376** 136–136. https://doi.org/10.1126/science.abm0101

[16] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32. https://doi.org/10.18637/jss.v076.i01

[17] Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10** 101–129. https://doi.org/10.2307/3001666

[18] ATLAS Collaboration, Aaboud, M. (2018). Measurement of the W-boson mass in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *European Physical Journal C* **78** 110. https://doi.org/10.1140/epjc/s10052-017-5475-4

[19] CDF Collaboration (2022). High-precision measurement of the *W* boson mass with the CDF II detector. *Science* **376** 170–176. https://doi.org/10.1126/science.abk1781

[20] L3 Collaboration (2006). Measurement of the mass and the width of the W boson at LEP. *Eur. Phys. J. C* **45** 569–587. https://doi.org/10.1140/epjc/s2005-02459-6

[21] Analytical Methods Committee (1989a). Robust statistics—how not to reject outliers. Part 1. Basic concepts. *Analyst* **114** 1693–1697. https://doi.org/10.1039/AN9891401693

[22] Analytical Methods Committee (1989b). Robust statistics—how not to reject outliers. Part 2. Inter-laboratory trials. *Analyst* **114** 1699–1702.

[23] Cooper, H., Hedges, L. V. and Valentine, J. C., eds. (2019) *The Handbook of Research Synthesis and Meta-Analysis*, 3rd ed. Russell Sage Foundation Publications, New York, NY.

[24] Cox, M. G. (2007). The evaluation of key comparison data: Determining the largest consistent subset. *Metrologia* **44** 187–200. https://doi.org/10.1088/0026-1394/44/3/005

[25] Dai, D. C. (2021). Variance of Newtonian constant from local gravitational acceleration measurements. *Phys. Rev. D* **103** 064059. https://doi.org/10.1103/PhysRevD.103.064059

[26] De Bièvre, P. (2007). Statistics and measurement results in chemistry. *Accredit. Qual. Assur.* **12** 333–334. https://doi.org/10.1007/s00769-007-0294-1

[27] DELPHI Collaboration Abdallah, J. et al. Measurement of the mass and width of the W boson in e^+e^- collisions at $\sqrt{s} =$ 161-209 GeV. *Eur. Phys. J. C* **55** 1. https://doi.org/10.1140/epjc/s10052-008-0585-7

[28] DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7** 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

[29] Diamond, G. A., Bax, L. and Kaul, S. (2007). Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Ann. Intern. Med.* **147** 578–581. https://doi.org/10.7326/0003-4819-147-8-200710160-00182

[30] Fineberg, H. V., Allison, D. B., Barba, L. A., Chong, D., Donoho, D., Freire, J., Gabrielse, G., Gatsonis, C., Hall, E. et al. (2019). *Reproducibility and Replicability in Science. Committee on Reproducibility and Replicability in Science, the National Academies of Sciences, Engineering, and Medicine*. The National Academies Press, Washington, DC. https://doi.org/10.17226/25303

[31] Gaiser, C., Fellmuth, B., Haft, N., Kuhn, A., Thiele-Krivoi, B., Zandt, T., Fischer, J., Jusko, O. and Sabuga, W. (2017). Final determination of the Boltzmann constant by dielectric-constant gas thermometry. *Metrologia* **54** 280–289. https://doi.org/10.1088/1681-7575/aa62e3

[32] Particle Data Group, Zyla, P. A. et al. (2020). Review of Particle Physics. Progress of Theoretical and Experimental Physics 083C01. https://doi.org/10.1093/ptep/ptaa104

[33] Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **379** 20200210. https://doi.org/10.1098/rsta.2020.0210

[34] Michell, J. (2005). The logic of measurement: A realist overview. *Measurement* **38** 285–294. https://doi.org/10.1016/j.measurement.2005.09.004

[35] Harris, D. C. and Lucy, C. A. (2020). *Quantitative Chemical Analysis*, 10th ed. Macmillan Learning, New York, NY.

[36] Herschel, J. F. W. (1866). Familiar Lectures on Scientific Subjects X. The Yard, the Pendulum, and the Metre 419–451, London Alexander Strahan.

[37] Home, P. D., Pocock, S. J., Beck-Nielsen, H., Curtis, P. S., Gomis, R., Hanefeld, M., Jones, N. P., Komajda, M. and McMurray, J. J. V. (2009). Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): A multicentre, randomised, open-label trial. *Lancet* **373** 2125–2135. https://doi.org/10.1016/S0140-6736(09)60953-3

[38] Jewell, N. P. (2004). *Statistics for Epidemiology*. CRC Press/CRC, Boca Raton, FL.

[39] Kahn, S. E., Haffner, S. M., Heise, M. A., Herman, W. H., Holman, R. R., Jones, N. P., Kravitz, B. G., Lachin, J. M., O'Neill, M. C. et al. (2006). Glycemic durability of rosiglitazone, metformin, or glyburide monotherapy. *N. Engl. J. Med.* **355** 2427–2443. https://doi.org/10.1056/NEJMoa066224

[40] Klein, N. (2020). Evidence for modified Newtonian dynamics from Cavendish-type gravitational constant experiments. *Classical Quantum Gravity* **37** 065002, 21. MR4086686 https://doi.org/10.1088/1361-6382/ab6cab

[41] Koepke, A., Lafarge, T., Possolo, A. and Toman, B. (2017). Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia* **54** S34–S62. https://doi.org/10.1088/1681-7575/aa6c0e

[42] Koetse, M. J., Florax, R. J. G. M. and de Groot, H. L. F. (2010). Consequences of effect size heterogeneity for meta-analysis: A Monte Carlo study. *Stat. Methods Appl.* **19** 217–236. MR2651450 https://doi.org/10.1007/s10260-009-0125-0

[43] Langan, D., Higgins, J. P. T., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W. and Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res. Synth. Methods* **10** 83–98. https://doi.org/10.1002/jrsm.1316

[44] Langan, D., Higgins, J. P. T. and Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Res. Synth. Methods* **8** 181–198. https://doi.org/10.1002/jrsm.1198

[45] Maishman, T., Schaap, S., Silk, D. S., Nevitt, S. J., Woods, D. C. and Bowman, V. E. (2022). Statistical methods used to combine the effective reproduction number, $R(t)$, and other related measures of COVID-19 in the UK. *Stat. Methods Med. Res.* **31** 1757–1777. MR4478307 https://doi.org/10.1177/09622802221109506

[46] Mandel, J. (1972). Repeatability and reproducibility. *J. Qual. Technol.* **4** 74–85. https://doi.org/10.1080/00224065.1972.11980520

[47] Mandel, J. (1991). The validation of measurement through interlaboratory studies. *Chemom. Intell. Lab. Syst.* **11** 109–119. https://doi.org/10.1016/0169-7439(91)80058-X

[48] Mandel, J. and Paule, R. (1970). Interlaboratory evaluation of a material with unequal numbers of replicates. *Anal. Chem.* **42** 1194–1197. https://doi.org/10.1021/ac60293a019

[49] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22** 719–748. https://doi.org/10.1093/jnci/22.4.719

[50] McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2431553

[51] Merkatas, C., Toman, B., Possolo, A. and Schlamminger, S. (2019). Shades of dark uncertainty and consensus value for the Newtonian constant of gravitation. *Metrologia* **56** 054001. https://doi.org/10.1088/1681-7575/ab3365

[52] Milton, M. J. T. and Possolo, A. (2020). Trustworthy data underpin reproducible research. *Nat. Phys.* **16** 117–119. https://doi.org/10.1038/s41567-019-0780-5

[53] Misner, C. W., Thorne, K. S. and Wheeler, J. A. (2017). *Gravitation*. Princeton University Press, Princeton, NJ.

[54] Mohr, P. (2014). Newtonian constant of gravitation international consortium. https://www.nist.gov/programs-projects/newtonian-constant-gravitation-international-consortium. NIST Physical Measurement Laboratory.

[55] Mohr, P. J., Newell, D. B. and Taylor, B. N. (2015). *CODATA Recommended Values of the Fundamental Physical Constants*: 2014. CODATA Zenodo Collection. https://doi.org/10.5281/zenodo.22826

[56] Mohr, P. J., Newell, D. B. and Taylor, B. N. (2016). CODATA recommended values of the fundamental physical constants: 2014. *Rev. Modern Phys.* **88** 035009. https://doi.org/10.1103/RevModPhys.88.035009

[57] Moldover, M. R., Trusler, J. P. M., Edwards, T. J., Mehl, J. B. and Davis, R. S. (1988). Measurement of the universal gas constant R using a spherical acoustic resonator. *J. Res. Natl. Bur. Stand.* **93** 85–144. https://doi.org/10.6028/jres.093.010

[58] Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley Company, Reading, MA.

[59] Mould, J. and Uddin, S. A. (2014). Constraining a possible variation of G with type ia supernovae. *Publ. Astron. Soc. Austral.* **31** e015. https://doi.org/10.1017/pasa.2014.9

[60] Munafò, M. R., Chambers, C., Collins, A., Fortunato, L. and Macleod, M. (2022). The Reproducibility Debate Is an Opportunity, Not a Crisis. BMC Research Notes 15 43. https://doi.org/10.1186/s13104-022-05942-3

[61] Newell, D. B. (2014). A more fundamental international system of units. *Phys. Today* **67** 35–41. https://doi.org/10.1063/PT.3.2448

[62] Nissen, S. E. and Wolski, K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N. Engl. J. Med.* **356** 2457–2471. https://doi.org/10.1056/NEJMoa072761

[63] NIST/SEMATECH (2012). *NIST/SEMATECH E-Handbook of Statistical Methods*. National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD. https://doi.org/10.18434/M32189

[64] Nozick, R. (1981). *Philosophical Explanations*. Harvard Univ. Press, Cambridge, MA.

[65] Olver, F. W. J., Lozier, D. W., Boisvert, R. F. and Clark, C. W., eds. (2010) *NIST Handbook of Mathematical Functions*. Cambridge Univ. Press, Cambridge. MR2723248

[66] Pinheiro, J. C., Liu, C. and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *J. Comput. Graph. Statist.* **10** 249–276. MR1939700 https://doi.org/10.1198/10618600152628059

[67] Pinheiro, L. and Emslie, K. R. (2018). Basic concepts and validation of digital PCR measurements. In *Digital PCR: Methods and Protocols* 11–24 Springer, New York, New York, NY. https://doi.org/10.1007/978-1-4939-7778-9_2

[68] Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Front. Neuroinform.* **11** 76. https://doi.org/10.3389/fninf.2017.00076

[69] Pontius, P. E. (1966). Measurement philosophy of the pilot program for mass calibration. National Bureau of Standards, Washington, DC. NBS Technical Note 288, Reprinted 1968, with minor corrections.

[70] Possolo, A. (2018). Measurement. In *Advanced Mathematical and Computational Tools in Metrology and Testing: AMCTM XI* (A. B. Forbes, N. F. Zhang, A. Chunovkina, S. Eichstädt and F. Pavese, eds.). *Series on Advances in Mathematics for Applied Sciences* **89** 273–285. World Scientific Company, Singapore. https://doi.org/10.1142/9789813274303\protect\T1\textunderscore0027

[71] Possolo, A. (2021). Concepts, methods, and tools enabling measurement quality. In *Frontiers in Statistical Quality Control* 13 (S. Knoth and W. Schmid, eds.) **19** 339–357. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-030-67856-2\protect\T1\textunderscore19

[72] Possolo, A. (2023). Supplement to "Tracking truth through measurement and the spyglass of statistics." https://doi.org/10.1214/23-STS899SUPP

[73] Possolo, A., Bruce, S. S. and Watters, R. L. Jr. (2021). *Metrological Traceability Frequently Asked Questions and NIST Policy*. National Institute of Standards and Technology, Gaithersburg, MD. NIST Technical Note 2156. https://doi.org/10.6028/NIST.TN.2156

[74] Possolo, A., Koepke, A., Newton, D. and Winchester, M. R. (2021). Decision tree for key comparisons. *J. Res. Natl. Inst. Stand. Technol.* **126** 126007. https://doi.org/10.6028/jres.126.007

[75] Possolo, A. and Meija, J. (2022). *Measurement Uncertainty: A Reintroduction*, 2nd ed. Sistema Interamericano de Metrologia (SIM), Montevideo, Uruguay. https://doi.org/10.4224/1tqz-b038

[76] Qu, J., Benz, S. P., Coakley, K., Rogalla, H., Tew, W. L., White, R., Zhou, K. and Zhou, Z. (2017). An improved electronic determination of the Boltzmann constant by Johnson noise thermometry. *Metrologia* **54** 549–558. https://doi.org/10.1088/1681-7575/aa781e

[77] Quinn, T., Parks, H., Speake, C. and Davis, R. (2013). Improved determination of *G* using two methods. *Phys. Rev. Lett.* **111** 101102. https://doi.org/10.1103/PhysRevLett.111.101102

[78] Quinn, T., Speake, C., Parks, H. and Davis, R. (2014). The BIPM measurements of the Newtonian constant of gravitation, *G*. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **372** 0032. https://doi.org/10.1098/rsta.2014.0032

[79] Roush, S. (2005). *Tracking Truth: Knowledge, Evidence, and Science*. Oxford Univ. Press, New York, NY.

[80] Rukhin, A. L. (2009). Weighted means statistics in interlaboratory studies. *Metrologia* **46** 323–331. https://doi.org/10.1088/0026-1394/46/3/021

[81] Rukhin, A. L., Biggerstaff, B. J. and Vangel, M. G. (2000). Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. *J. Statist. Plann. Inference* **83** 319–330. MR1748018 https://doi.org/10.1016/S0378-3758(99)00098-1

[82] Schlamminger, S. (2014). A cool way to measure big *G*. *Nature* **510** 478–480. https://doi.org/10.1038/nature13507

[83] Schlamminger, S., Chao, L. S., Lee, V., Speake, C. C. and Newell, D. B. (2022). Measurement of Newton's gravitational constant with the BIPM torsion balance. In *American Physical Society April Meeting* 2022 Session S16: Lab Experiments and Detector Characterization *S*16.00002.

[84] Schlamminger, S., Holzschuh, E., Kündig, W., Nolting, F., Pixley, R. E., Schurr, J. and Straumann, U. (2006). Measurement of Newton's gravitational constant. *Phys. Rev. D* **74** 082001. https://doi.org/10.1103/PhysRevD.74.082001

[85] Strain, M. C., Lada, S. M., Luong, T., Rought, S. E., Gianella, S., Terry, V. H., Spina, C. A., Woelk, C. H. and Richman, D. D. (2013). Highly precise measurement of HIV DNA by droplet digital PCR. *PLoS ONE* **8** 1–8. https://doi.org/10.1371/journal.pone.0055943

[86] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[87] Stan Development Team (2022). RStan: the R interface to Stan. R package version 2.21.7.

[88] Thomas, K. and Schmidt, M. S. (2012). Glaxo Agrees to Pay $3 Billion in Fraud Settlement. The New York Times July 2.

[89] Thompson, M. and Ellison, S. L. R. (2011). Dark uncertainty. *Accredit. Qual. Assur.* **16** 483–487. https://doi.org/10.1007/s00769-011-0803-0

[90] Tiesinga, E., Mohr, P. J., Newell, D. B. and Taylor, B. N. (2021). CODATA recommended values of the fundamental physical constants: 2018. *Rev. Modern Phys.* **93** 025010. https://doi.org/10.1103/RevModPhys.93.025010

[91] Viberti, G., Kahn, S. E., Greene, D. A., Herman, W. H., Zinman, B., Holman, R. R., Haffner, S. M., Levy, D., Lachin, J. M. et al. (2002). A Diabetes Outcome Progression Trial (ADOPT): An international multicenter study of the comparative efficacy of rosiglitazone, glyburide, and metformin in recently diagnosed type 2 diabetes. *Diabetes Care* **25** 1737–1743. https://doi.org/10.2337/diacare.25.10.1737

[92] Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36** 1–48. https://doi.org/10.18637/jss.v036.i03

[93] White, R. (2011). The meaning of measurement in metrology. *Accredit. Qual. Assur.* **16** 31–41. https://doi.org/10.1007/s00769-010-0698-1

[94] Wilson, E. O. (1998). *Consilience: The Unity of Knowledge*. Alfred A. Knopf, New York, NY.

[95] Yusuf, S., Peto, R., Lewis, J., Collins, R. and Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Prog. Cardiovasc. Dis.* **27** 335–371. https://doi.org/10.1016/s0033-0620(85)80003-7

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.