STATISTICAL SCIENCE

Volume 35, Number 3 August 2020

Special Issue on Causal Inference

Editorial: Special Issue on "Causal Inference" ..... *Jason Roy and Dylan Small* 335
Matching Methods for Observational Studies Derived from Large Administrative Databases ..... *Ruoqi Yu, Jeffrey H. Silber and Paul R. Rosenbaum* 338
Comment: Matching Methods for Observational Studies Derived from Large Administrative Databases ..... *Fredrik Sävje* 356
Comment: Matching Methods for Observational Studies Derived from Large Administrative Databases.... *Mark M. Fredrickson, Josh Errickson and Ben B. Hansen* 361
Commentary on Yu et al.: Opportunities and Challenges for Matching Methods in Large Databases ..... *Elizabeth A. Stuart and Benjamin Ackerman* 367
Rejoinder: Matching Methods for Observational Studies Derived from Large Administrative Databases ..... *Ruoqi Yu, Jeffrey H. Silber and Paul R. Rosenbaum* 371
Linear Mixed Models with Endogenous Covariates: Modeling Sequential Treatment Effects with Application to a Mobile Health Study ..... *Tianchen Qian, Predrag Klasnja and Susan A. Murphy* 375
Comment: Clarifying Endogeneous Data Structures and Consequent Modelling Choices Using Causal Graphs ..... *Erica E. M. Moodie and David A. Stephens* 391
Moving Toward Rigorous Evaluation of Mobile Health Interventions..... *Kristin A. Linn* 394
Comment: Diagnostics and Kernel-based Extensions for Linear Mixed Effects Models with Endogenous Covariates ..... *Hunyong Cho, Joshua P. Zitovsky, Xinyi Li, Minxin Lu, Kushal Shah, John Sperger, Matthew C. B. Tsilimigras and Michael R. Kosorok* 396
Rejoinder: Linear Mixed Models with Endogenous Covariates: Modeling Sequential Treatment Effects with Application to a Mobile Health Study ..... *Tianchen Qian, Predrag Klasnja and Susan A. Murphy* 400
Invariance, Causality and Robustness ..... *Peter Bühlmann* 404
Comment: Invariance, Causality and Robustness by P. Bühlmann ..... *Vanessa Didelez* 427
Comment: Invariance and Causal Inference ..... *Stefan Wager* 430
Rejoinder: Invariance, Causality and Robustness..... *Peter Bühlmann* 434
Outcome-Wide Longitudinal Designs for Causal Inference: A New Template for Empirical Studies..... *Tyler J. VanderWeele, Maya B. Mathur and Ying Chen* 437
Comment: On the Potential for Misuse of Outcome-Wide Study Designs, and Ways to Prevent It..... *Stijn Vansteelandt and Oliver Dukes* 467
Comment: Outcome-Wide Individualized Treatment Strategies ..... *Ashkan Ertefaie and Brent A. Johnson* 472

continued

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 35, Number 3, August 2020. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, OH 44122, 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, 9650 Rockville Pike—Suite L2310, Bethesda, MD 20814-3998, USA.

Copyright © 2020 by the Institute of Mathematical Statistics
Printed in the United States of America

STATISTICAL SCIENCE

Volume 35, Number 3 August 2020

continued

A New Template for Empirical Studies: From positivity to Positivity . . . . . *Rhian Daniel* 476
Rejoinder: The Future of Outcome-Wide Studies
. . . . . *Tyler J. VanderWeele, Maya B. Mathur and Ying Chen* 479
A Nonparametric Super-Efficient Estimator of the Average Treatment Effect
. . . . . *David Benkeser, Weixin Cai and Mark J. van der Laan* 484
Comment: Increasing Real World Usage of Targeted Minimum Loss-Based Estimators
. . . . . *Mireille E. Schnitzer* 496
Comment: Automated Analyses: Because We Can, Does It Mean We Should?
. . . . . *Susan M. Shortreed and Erica E. M. Moodie* 499
Comment: Stabilizing the Doubly-Robust Estimators of the Average Treatment Effect under Positivity Violations . . . . . *Fan Li* 503
Rejoinder: A Nonparametric Superefficient Estimator of the Average Treatment Effect
. . . . . *David Benkeser, Weixin Cai and Mark J. van der Laan* 511
On Nearly Assumption-Free Tests of Nominal Confidence Interval Coverage for Causal Parameters Estimated by Machine Learning
. . . . . *Lin Liu, Rajarshi Mukherjee and James M. Robins* 518
Discussion of "On Nearly Assumption-Free Tests of Nominal Confidence Interval Coverage for Causal Parameters Estimated by Machine Learning"
. . . . . *Edward H. Kennedy, Sivaraman Balakrishnan and Larry Wasserman* 540
Rejoinder: On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning
. . . . . *Lin Liu, Rajarshi Mukherjee and James M. Robins* 545

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 35, Number 3, August 2020. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, OH 44122, 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, 9650 Rockville Pike—Suite L2310, Bethesda, MD 20814-3998, USA.

Copyright © 2020 by the Institute of Mathematical Statistics
Printed in the United States of America

Statistical Science **Volume 35, Number 3 (335–554) August 2020**

Volume 35 **Number 3** **August 2020**

Special Issue on Causal Inference

Editorial: Special Issue on "Causal Inference"
Jason Roy and Dylan Small

Matching Methods for Observational Studies Derived from Large Administrative Databases
Ruoqi Yu, Jeffrey H. Silber and Paul R. Rosenbaum

Linear Mixed Models with Endogenous Covariates: Modeling Sequential Treatment Effects with Application to a Mobile Health Study
Tianchen Qian, Predrag Klasnja and Susan A. Murphy

Invariance, Causality and Robustness
Peter Bühlmann

Outcome-Wide Longitudinal Designs for Causal Inference: A New Template for Empirical Studies
Tyler J. VanderWeele, Maya B. Mathur and Ying Chen

A Nonparametric Super-Efficient Estimator of the Average Treatment Effect
David Benkeser, Weixin Cai and Mark J. van der Laan

On Nearly Assumption-Free Tests of Nominal Confidence Interval Coverage for Causal Parameters Estimated by Machine Learning
Lin Liu, Rajarshi Mukherjee and James M. Robins

EDITOR
Sonia Petrone
Università Bocconi

DEPUTY EDITOR
Peter Green
University of Bristol and University of Technology Sydney

ASSOCIATE EDITORS

Shankar Bhamidi
University of North Carolina
Peter Bühlmann
ETH Zürich
Rong Chen
Rutgers University
Robin Evans
University of Oxford
Stefano Favaro
Università di Torino
Peter Green
University of Bristol and University of Technology Sydney
Chris Holmes
University of Oxford
Susan Holmes
Stanford University
Tailen Hsing
University of Michigan
Michael I. Jordan
University of California, Berkeley
Ian McKeague
Columbia University
Vladimir Minin
University of California, Irvine
Peter Müller
University of Texas
Luc Pronzato
Université Nice
Nancy Reid
University of Toronto
Jason Roy
Rutgers University
Chiara Sabatti
Stanford University
Richard Samworth
University of Cambridge
Bodhisattva Sen
Columbia University
Glenn Shafer
Rutgers Business School–Newark and New Brunswick
Royal Holloway College, University of London
Michael Stein
University of Chicago
Jon Wellner
University of Washington
Yihong Wu
Yale University
Minge Xie
Rutgers University
Bin Yu
University of California, Berkeley
Giacomo Zanella
Università Bocconi
Cun-Hui Zhang
Rutgers University
Harrison Zhou
Yale University

MANAGING EDITOR
Robert Keener
University of Michigan

PRODUCTION EDITOR
Patrick Kelly

EDITORIAL COORDINATOR
Kristina Mattson

PAST EXECUTIVE EDITORS

Morris H. DeGroot, 1986–1988
Carl N. Morris, 1989–1991
Robert E. Kass, 1992–1994
Paul Switzer, 1995–1997
Leon J. Gleser, 1998–2000
Richard Tweedie, 2001
Morris Eaton, 2001
George Casella, 2002–2004
Edward I. George, 2005–2007
David Madigan, 2008–2010
Jon A. Wellner, 2011–2013
Peter Green, 2014–2016
Cun-Hui Zhang, 2017–2019

Statistical Science
2020, Vol. 35, No. 3, 335–337
https://doi.org/10.1214/20-STS800
© Institute of Mathematical Statistics, 2020

Editorial: Special Issue on "Causal Inference"

Jason Roy and Dylan Small

REFERENCES

FISHER, R. A. (1935). *The Design of Experiments*. Hafner, New York.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

Jason Roy is Professor of Biostatistics, Rutgers School of Public Health, Piscataway, New Jersey 08854, USA (e-mail: jason.roy@rutgers.edu). Dylan Small is Class of 1965 Wharton Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: dsmall@wharton.upenn.edu).

Statistical Science
2020, Vol. 35, No. 3, 338–355
https://doi.org/10.1214/19-STS699
© Institute of Mathematical Statistics, 2020

Matching Methods for Observational Studies Derived from Large Administrative Databases

Ruoqi Yu, Jeffrey H. Silber and Paul R. Rosenbaum

Abstract. We propose new optimal matching techniques for large administrative data sets. In current practice, very large matched samples are constructed by subdividing the population and solving a series of smaller problems, for instance, matching men to men and separately matching women to women. Without simplification of some kind, the time required to optimally match T treated individuals to T controls selected from $C \geq T$ potential controls grows much faster than linearly with the number of people to be matched—the required time is of order $O\{(T + C)^3\}$—so splitting one large problem into many small problems greatly accelerates the computations. This common practice has several disadvantages that we describe. In its place, we propose a single match, using everyone, that accelerates the computations in a different way. In particular, we use an iterative form of Glover's algorithm for a doubly convex bipartite graph to determine an optimal caliper for the propensity score, radically reducing the number of candidate matches; then we optimally match in a large but much sparser graph. In this graph, a modified form of near-fine balance can be used on a much larger scale, improving its effectiveness. We illustrate the method using data from US Medicaid, matching children receiving surgery at a children's hospital to similar children receiving surgery at a hospital that mostly treats adults. In the example, we form 38,841 matched pairs from 159,527 potential controls, controlling for 29 covariates plus 463 Principal Surgical Procedures, plus 973 Principal Diagnoses. The method is implemented in an `R` package `bigmatch` available from `CRAN`.

Key words and phrases: Causal inference, fine balance, Glover's algorithm, observational study, optimal caliper, optimal matching, propensity score.

REFERENCES

BERTSEKAS, D. P. (1981). A new algorithm for the assignment problem. *Math. Program.* **21** 152–171. MR0623835 https://doi.org/10.1007/BF01584237

BERTSEKAS, D. P. (1998). *Network Optimization*. Athena Scientific, Belmont, MA.

BERTSEKAS, D. P. and TSENG, P. (1988). The RELAX codes for linear minimum cost network flow problems. *Ann. Oper. Res.* **13** 125–190. MR0950991 https://doi.org/10.1007/BF02288322

COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā, A* **35** 417–446.

GLOVER, F. (1967). Maximum matching in a convex bipartite graph. *Naval Res. Logist.* **14** 313–316.

GOEMAN, J. J., SOLARI, A. and STIJNEN, T. (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Stat. Med.* **29** 2117–2125. MR2756559 https://doi.org/10.1002/sim.4002

HANSEN, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika* **95** 481–488. MR2521594 https://doi.org/10.1093/biomet/asn004

HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist.* **15** 609–627. MR2280151 https://doi.org/10.1198/106186006X137047

KORTE, B. and VYGEN, J. (2012). *Combinatorial Optimization: Theory and Algorithms*, 5th ed. *Algorithms and Combinatorics*

Ruoqi Yu is a PhD student, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: ruoqiyu@wharton.upenn.edu). Jeffrey H. Silber is Professor, Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: silber@email.chop.edu; URL: https://cor.research.chop.edu/). Paul R. Rosenbaum is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: rosenbaum@wharton.upenn.edu).

21. Springer, Heidelberg. MR2850465 https://doi.org/10.1007/978-3-642-24488-9

LIPSKI, W. JR. and PREPARATA, F. P. (1981). Efficient algorithms for finding maximum matchings in convex bipartite graphs and related problems. *Acta Inform*. **15** 329–346. MR0632418 https://doi.org/10.1007/BF00264533

LU, B., GREEVY, R., XU, X. and BECK, C. (2011). Optimal nonbipartite matching and its statistical applications. *Amer. Statist*. **65** 21–30. MR2899649 https://doi.org/10.1198/tast.2011.08294

PIMENTEL, S. D. (2016). Large, sparse optimal matching with R package rcbalance. *Obs*. *Stud*. **2** 4–23.

PIMENTEL, S. D., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Amer. Statist*. *Assoc*. **110** 515–527. MR3367244 https://doi.org/10.1080/01621459.2014.997879

ROSENBAUM, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist*. *Assoc*. **84** 1024–1032.

ROSENBAUM, P. R. (2002). Attributing effects to treatment in matched observational studies. *J. Amer. Statist*. *Assoc*. **97** 183–192. MR1963391 https://doi.org/10.1198/016214502753479329

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist*. *Assoc*. **102** 75–83. MR2345534 https://doi.org/10.1198/016214506000001059

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist*. **39** 33–38.

ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist*. *Assoc*. **104** 1398–1405. MR2750570 https://doi.org/10.1198/jasa.2009.tm08470

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist*. *Assoc*. **74** 318–328.

RUBIN, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics* **36** 293–298.

SILBER, J. H., ROSENBAUM, P. R., MCHUGH, M. D., LUDWIG, J. M., SMITH, H. L., NIKNAM, B. A., EVEN-SHOSHAN, O., FLEISHER, L. A., KELZ, R. R. et al. (2016). Comparison of the value of nursing work environments in hospitals across different levels of patient risk. *JAMA Surg*. **151** 527–536.

SILBER, J. H., ROSENBAUM, P. R., WANG, W., CALHOUN, S. R., REITER, J. G., EVEN-SHOSHAN, O. and GREELEY, W. J. (2018). Practice style variation in medicaid and non-medicaid children with complex chronic conditions undergoing surgery. *Ann*. *Surg*. **267** 392–400.

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist*. *Sci*. **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

YANG, D., SMALL, D. S., SILBER, J. H. and ROSENBAUM, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* **68** 628–636. MR2959630 https://doi.org/10.1111/j.1541-0420.2011.01691.x

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist*. *Assoc*. **107** 1360–1371. MR3036400 https://doi.org/10.1080/01621459.2012.703874

ZUBIZARRETA, J. R., REINKE, C. E., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2011). Matching for several sparse nominal variables in a case-control study of readmission following surgery. *Amer. Statist*. **65** 229–238. MR2867507 https://doi.org/10.1198/tas.2011.11072

Statistical Science
2020, Vol. 35, No. 3, 356–360
https://doi.org/10.1214/19-STS739
© Institute of Mathematical Statistics, 2020

Comment: Matching Methods for Observational Studies Derived from Large Administrative Databases

Fredrik Sävje

REFERENCES

BENNETT, M., VIELMA, J. P. and ZUBIZARRETA, J. R. (2018). Building representative matched samples with multi-valued treatments in large observational studies: Analysis of the impact of an earthquake on educational attainment. Available at arXiv:1810.06707.

EFRAT, A., ITAI, A. and KATZ, M. J. (2001). Geometry helps in bottleneck matching and related problems. *Algorithmica* **31** 1–28. MR1840188 https://doi.org/10.1007/s00453-001-0016-8

FULKERSON, R., GLICKSBERG, I. and GROSS, O. (1953). A production line assignment problem Technical Report No. RM-1102 The Rand Corporation Santa Monica.

GARFINKEL, R. S. (1971). An improved algorithm for the bottleneck assignment problem. *Oper. Res*. **19** 1747–1751.

PFERSCHY, U. (1996). The random linear bottleneck assignment problem. *RAIRO Oper. Res*. **30** 127–142. MR1424230 https://doi.org/10.1051/ro/1996300201271

PUNNEN, A. P. and NAIR, K. P. K. (1994). Improved complexity bound for the maximum cardinality bottleneck bipartite matching problem. *Discrete Appl. Math*. **55** 91–93. MR1298517 https://doi.org/10.1016/0166-218X(94)90039-6

RINNOOY KAN, A. H. G. (1976). *Machine Scheduling Problems*: *Classification*, *Complexity and Computations*. Martinus Nijhoff, The Hague.

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist*. **39** 33–38.

SÄVJE, F., HIGGINS, M. J. and SEKHON, J. S. (2017). Generalized full matching. Available at arXiv:1703.03882.

YU, R., SILBER, J. H. and ROSENBAUM, P. R. (2020). Matching methods for observational studies derived from large administrative databases. *Statist. Sci*. **35** 338–355.

Fredrik Sävje is Assistant Professor, Department of Political Science and Department of Statistics and Data Science, Yale University, Rosenkranz Hall, 115 Prospect Street, New Haven, Connecticut 06520, USA (e-mail: fredrik.savje@yale.edu).

Statistical Science
2020, Vol. 35, No. 3, 361–366
https://doi.org/10.1214/19-STS740
© Institute of Mathematical Statistics, 2020

Comment: Matching Methods for Observational Studies Derived from Large Administrative Databases

Mark M. Fredrickson, Josh Errickson and Ben B. Hansen

REFERENCES

BERTSEKAS, D. (1998). *Network Optimization*: *Continuous and Discrete Models*. Athena Scientific, Belmont, MA.

COHEN, N. and BRASSIL, J. (2000). A parallel pruning technique for highly asymmetric assignment problems. *IEEE Trans*. *Parallel Distrib. Syst*. **11** 550–558. https://doi.org/10.1109/71.862206

COLANNINO, J., DAMIAN, M., HURTADO, F., IACONO, J., MEIJER, H., RAMASWAMI, S. and TOUSSAINT, G. (2006). An $O(n \log n)$-time algorithm for the restriction scaffold assignment problem. *J. Comput*. *Biol*. **13** 979–989. MR2231777 https://doi.org/10.1089/cmb.2006.13.979

COLANNINO, J., DAMIAN, M., HURTADO, F., LANGERMAN, S., MEIJER, H., RAMASWAMI, S., SOUVAINE, D. and TOUSSAINT, G. (2007). Efficient many-to-many point matching in one dimension. *Graphs Combin*. **23** 169–178. MR2320626 https://doi.org/10.1007/s00373-007-0714-3

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144 https://doi.org/10.1093/biomet/asn055

FOGARTY, C. B., MIKKELSEN, M. E., GAIESKI, D. F. and SMALL, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Amer. Statist*. *Assoc*. **111** 447–458. MR3538678 https://doi.org/10.1080/01621459.2015.1112802

GU, X. S. and ROSENBAUM, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *J. Comput*. *Graph*. *Statist*. **2** 405–420.

GURM, H. S., HOSMAN, C., SHARE, D., MOSCUCCI, M. and HANSEN, B. B. (2013). Comparative safety of vascular closure devices and manual closure among patients having percutaneous coronary intervention. *Ann*. *Intern*. *Med*. **159** 660–666.

HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist*. *Assoc*. **99** 609–618. MR2086387 https://doi.org/10.1198/016214504000000647

HANSEN, B. B. (2011). Propensity score matching to extract latent experiments from nonexperimental data: A case study. In *Looking Back*: *Proceedings of a Conference in Honor of Paul W. Holland* (N. Dorans and S. Sinharay, eds.) 149–181 9. Springer.

HANSEN, B. B. and BOWERS, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statist*. *Sci*. **23** 219–236. MR2516821 https://doi.org/10.1214/08-STS254

HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput*. *Graph*. *Statist*. **15** 609–627. MR2280151 https://doi.org/10.1198/106186006X137047

HILL, J. and SU, Y.-S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann*. *Appl*. *Stat*. **7** 1386–1420. MR3127952 https://doi.org/10.1214/13-AOAS630

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

KALTON, G. (1968). Standardization: A technique to control for extraneous variables. *Appl*. *Statist*. **17** 118–136. MR0234599 https://doi.org/10.2307/2985676

KARP, R. M. and LI, S.-Y. R. (1975). Two special cases of the assignment problem. *Discrete Math*. **13** 129–142. MR0384086 https://doi.org/10.1016/0012-365X(75)90014-X

KILCIOGLU, C. and ZUBIZARRETA, J. R. (2016). Maximizing the information content of a balanced matched sample in a study of the economic performance of green buildings. *Ann*. *Appl*. *Stat*. **10** 1997–2020. MR3592046 https://doi.org/10.1214/16-AOAS962

MOHAMAD, M., RAPPAPORT, D. and TOUSSAINT, G. (2015). Minimum many-to-many matchings for computing the distance between two sequences. *Graphs Combin*. **31** 1637–1648. MR3386035 https://doi.org/10.1007/s00373-014-1467-4

PIMENTEL, S. D., YOON, F. and KEELE, L. (2015). Variable-ratio matching with fine balance in a study of the Peer Health Exchange. *Stat*. *Med*. **34** 4070–4082. MR3431322 https://doi.org/10.1002/sim.6593

PIMENTEL, S. D., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Amer. Statist*. *Assoc*. **110** 515–527. MR3367244 https://doi.org/10.1080/01621459.2014.997879

PIMENTEL, S. D., PAGE, L. C., LENARD, M. and KEELE, L. (2018). Optimal multilevel matching using network flows: An application to a summer reading intervention. *Ann*. *Appl*. *Stat*. **12** 1479–1505. MR3852685 https://doi.org/10.1214/17-AOAS1118

RAJABI-ALNI, F. and BAGHERI, A. (2016). An $O(n^2)$ algorithm for the limited-capacity many-to-many point matching in one dimension. *Algorithmica* **76** 381–400. MR3537023 https://doi.org/10.1007/s00453-015-0044-4

ROSENBAUM, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist*. *Assoc*. **84** 1024–1032.

Mark M. Fredrickson is Postdoctoral Research Fellow, Department of Statistics, University of Michigan, 311 West Hall, 1085 South University Ave, Ann Arbor, MI 48109 (e-mail: mfredric@umich.edu). Josh Errickson is Statistician Senior, Consulting for Statistics, Computing and Analytics Research (CSCAR), University of Michigan 3550 Rackham Bldg, Ann Arbor MI 48109 (e-mail: jerrick@umich.edu). Ben B. Hansen is Associate Professor, Department of Statistics, University of Michigan, 323 West Hall, 1085 South University Ave, Ann Arbor, MI 48109 (e-mail: bbh@umich.edu).

ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610. MR1125717

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2

ROSENBAUM, P. R. (2017). Imposing minimax and quantile constraints on optimal matching in observational studies. *J. Comput. Graph. Statist.* **26** 66–78. MR3610408 https://doi.org/10.1080/10618600.2016.1152971

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

RUZANKIN, P. S. (2019). A fast algorithm for maximal propensity score matching. *Methodol. Comput. Appl. Probab.* https://doi.org/10.1007/s11009-019-09718-4

SÄVJE, F., HIGGINS, M. J. and SEKHON, J. S. (2017). Generalized full matching. arXiv e-prints, arXiv:1703.03882.

TRASKIN, M. and SMALL, D. S. (2011). Defining the study population for an observational study to ensure sufficient overlap: A tree approach. *Stat. Biosci.* **3** 94–118. https://doi.org/10.1007/s12561-011-9036-3

VASCONCELOS, C. and ROSENHAHN, B. (2009). Bipartite graph matching computation on GPU. In *Energy Minimization Methods in Computer Vision and Pattern Recognition* (D. Cremers, Y. Boykov, A. Blake and F. R. Schmidt, eds.). *Lecture Notes in Computer Science* **5681** 42–55. Springer, Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03641-5_4

YANG, D., SMALL, D. S., SILBER, J. H. and ROSENBAUM, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* **68** 628–636. MR2959630 https://doi.org/10.1111/j.1541-0420.2011.01691.x

YU, R. and ROSENBAUM, P. R. (2019). Directional penalties for optimal matching in observational studies. *Biometrics* **75** 1380–1390. MR4041838 https://doi.org/10.1111/biom.13098

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400 https://doi.org/10.1080/01621459.2012.703874

Statistical Science
2020, Vol. 35, No. 3, 367–370
https://doi.org/10.1214/19-STS741
© Institute of Mathematical Statistics, 2020

Commentary on Yu et al.: Opportunities and Challenges for Matching Methods in Large Databases

Elizabeth A. Stuart and Benjamin Ackerman

REFERENCES

[1] AUSTIN, P. C. and STUART, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* **34** 3661–3679. MR3422140 https://doi.org/10.1002/sim.6607

[2] DAW, J. R. and HATFIELD, L. A. (2018). Matching and regression to the mean in difference-in-differences analysis. *Health Serv. Res.* **53** 4138–4156. https://doi.org/10.1111/1475-6773.12993

[3] ERLANGSEN, A., LIND, B. D., STUART, E. A., QIN, P., STENAGER, E., LARSEN, K. J., WANG, A. G., HVID, M., NIELSEN, A. C. et al. (2015). Short-term and long-term effects of psychosocial therapy for people after deliberate self-harm: A register-based, nationwide multicentre study using propensity score matching. *The Lancet Psychiatry* **2** 49–58.

[4] GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.* **22** 153–164. MR2408951 https://doi.org/10.1214/088342306000000691

[5] GERUSO, M. and LAYTON, T. (2015). Upcoding: Evidence from Medicare on squishy risk adjustment. Technical Report, National Bureau of Economic Research.

[6] GOLDSTEIN, B. A., BHAVSAR, N. A., PHELAN, M. and PENCINA, M. J. (2016). Controlling for informed presence bias due to the number of health encounters in an electronic health record. *Am. J. Epidemiol.* **184** 847–855.

[7] ROSE, S. (2016). A machine learning framework for plan payment risk adjustment. *Health Serv. Res.* **51** 2358–2374. https://doi.org/10.1111/1475-6773.12464

[8] ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* **102** 75–83. MR2345534 https://doi.org/10.1198/016214506000001059

[9] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

[10] ROUSSEL, R., TRAVERT, F., PASQUET, B., WILSON, P. W., SMITH, S. C., GOTO, S., RAVAUD, P., MARRE, M., PORATH, A. et al. (2010). Metformin use and mortality among patients with diabetes and atherothrombosis. *Arch. Intern. Med.* **170** 1892–1899.

[11] STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

[12] STUART, E. A., LEE, B. K. and LEACY, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J. Clin. Epidemiol.* **66** S84–S90.

[13] YU, R., SILBER, J. H. and ROSENBAUM, P. R. (2020). Matching methods for observational studies derived from large administrative databases. *Statist. Sci.* **35** 338-355.

Elizabeth A. Stuart is Professor of Mental Health, Biostatistics, and Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205, USA (e-mail: estuart@jhu.edu). Benjamin Ackerman is a Doctoral Candidate in the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205, USA.

Statistical Science
2020, Vol. 35, No. 3, 371–374
https://doi.org/10.1214/20-STS790
© Institute of Mathematical Statistics, 2020

Rejoinder: Matching Methods for Observational Studies Derived from Large Administrative Databases

Ruoqi Yu, Jeffrey H. Silber and Paul R. Rosenbaum

REFERENCES

BERTSEKAS, D. P. (1998). *Network Optimization*. Athena Scientific, Belmont, MA.

FOGARTY, C. B., MIKKELSEN, M. E., GAIESKI, D. F. and SMALL, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Amer. Statist. Assoc.* **111** 447–458. MR3538678 https://doi.org/10.1080/01621459.2015.1112802

GARFINKEL, R. S. (1971). An improved algorithm for the bottleneck assignment problem. *Oper. Res.* **19** 1747–1751.

GLOVER, F. (1967). Maximum matching in a convex bipartite graph. *Nav. Res. Logist. Q.* **14** 313–316.

HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2009). Split samples and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **104** 1090–1101. MR2750238 https://doi.org/10.1198/jasa.2009.tm08338

HSU, J. Y., ZUBIZARRETA, J. R., SMALL, D. S. and ROSENBAUM, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* **102** 767–782. MR3431552 https://doi.org/10.1093/biomet/asv034

KARMAKAR, B., SMALL, D. S. and ROSENBAUM, P. R. (2019). Using approximation algorithms to build evidence factors and related designs for observational studies. *J. Comput. Graph. Statist.* **28** 698–709. MR4007751 https://doi.org/10.1080/10618600.2019.1584900

KORTE, B. and VYGEN, J. (2012). *Combinatorial Optimization: Theory and Algorithms*, 5th ed. *Algorithms and Combinatorics* **21**. Springer, Heidelberg. MR2850465 https://doi.org/10.1007/978-3-642-24488-9

LEE, K., SMALL, D. S. and ROSENBAUM, P. R. (2018). A powerful approach to the study of moderate effect modification in observational studies. *Biometrics* **74** 1161–1170. MR3908134

PIMENTEL, S. D. and KELZ, R. R. (2020). Optimal tradeoffs in matched designs comparing US-trained and internationally trained surgeons. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1720693

PIMENTEL, S. D., YOON, F. and KEELE, L. (2015). Variable-ratio matching with fine balance in a study of the Peer Health Exchange. *Stat. Med.* **34** 4070–4082. MR3431322 https://doi.org/10.1002/sim.6593

ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152. MR2133562 https://doi.org/10.1198/000313005X42831

ROSENBAUM, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *J. Comput. Graph. Statist.* **21** 57–71. MR2913356 https://doi.org/10.1198/jcgs.2011.09219

ROSENBAUM, P. R. (2017). Imposing minimax and quantile constraints on optimal matching in observational studies. *J. Comput. Graph. Statist.* **26** 66–78. MR3610408 https://doi.org/10.1080/10618600.2016.1152971

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat. Med.* **26** 20–36. MR2312697 https://doi.org/10.1002/sim.2739

TUKEY, J. W. (1980). We need both exploratory and confirmatory. *Amer. Statist.* **34** 23–25.

YU, R. and ROSENBAUM, P. R. (2019). Directional penalties for optimal matching in observational studies. *Biometrics* **75** 1380–1390. MR4041838 https://doi.org/10.1111/biom.13098

ZHANG, K., SMALL, D. S., LORCH, S., SRINIVAS, S. and ROSENBAUM, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *J. Amer. Statist. Assoc.* **106** 511–524. MR2847966 https://doi.org/10.1198/jasa.2011.ap10604

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400 https://doi.org/10.1080/01621459.2012.703874

ZUBIZARRETA, J. R., PAREDES, R. D. and ROSENBAUM, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Appl. Stat.* **8** 204–231. MR3191988 https://doi.org/10.1214/13-AOAS713

Ruoqi Yu is PhD student, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: ruoqiyu@wharton.upenn.edu). Jeffrey H. Silber is Professor, Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: silber@email.chop.edu; URL: https://cor.research.chop.edu/). Paul R. Rosenbaum is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: rosenbaum@wharton.upenn.edu).

Statistical Science
2020, Vol. 35, No. 3, 375–390
https://doi.org/10.1214/19-STS720
© Institute of Mathematical Statistics, 2020

Linear Mixed Models with Endogenous Covariates: Modeling Sequential Treatment Effects with Application to a Mobile Health Study

Tianchen Qian, Predrag Klasnja and Susan A. Murphy

Abstract. Mobile health is a rapidly developing field in which behavioral treatments are delivered to individuals via wearables or smartphones to facilitate health-related behavior change. Micro-randomized trials (MRT) are an experimental design for developing mobile health interventions. In an MRT, the treatments are randomized numerous times for each individual over course of the trial. Along with assessing treatment effects, behavioral scientists aim to understand between-person heterogeneity in the treatment effect. A natural approach is the familiar linear mixed model. However, directly applying linear mixed models is problematic because potential moderators of the treatment effect are frequently endogenous—that is, may depend on prior treatment. We discuss model interpretation and biases that arise in the absence of additional assumptions when endogenous covariates are included in a linear mixed model. In particular, when there are endogenous covariates, the coefficients no longer have the customary marginal interpretation. However, these coefficients still have a conditional-on-the-random-effect interpretation. We provide an additional assumption that, if true, allows scientists to use standard software to fit linear mixed model with endogenous covariates, and person-specific predictions of effects can be provided. As an illustration, we assess the effect of activity suggestion in the HeartSteps MRT and analyze the between-person treatment effect heterogeneity.

Key words and phrases: Linear mixed model, endogenous covariates, micro-randomized trial, causal inference.

REFERENCES

AGRESTI, A., BOOTH, J. G., HOBERT, J. P. and CAFFO, B. (2000). Random-effects modeling of categorical response data. *Sociol. Method.* **30** 27–80.

AMEMIYA, T. and MACURDY, T. E. (1986). Instrumental-variable estimation of an error-components model. *Econometrica* **54** 869–880. MR0853986 https://doi.org/10.2307/1912840

ARELLANO, M. and BOND, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Econ. Stud.* **58** 277–297.

ARELLANO, M. and BOVER, O. (1995). Another look at the instrumental variable estimation of error-components models. *J. Econometrics* **68** 29–51.

BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48.

BERGER, M. P. F. and TAN, F. E. S. (2004). Robust designs for linear mixed effects models. *J. Roy. Statist. Soc. Ser. C* **53** 569–581. MR2087773 https://doi.org/10.1111/j.1467-9876.2004.05152.x

BOLGER, N. and LAURENCEAU, J.-P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. Guilford Press.

BORUVKA, A., ALMIRALL, D., WITKIEWITZ, K. and MURPHY, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.* **113** 1112–1121. MR3862343 https://doi.org/10.1080/01621459.2017.1305274

BRUMBACK, B., GREENLAND, S., REDMAN, M., KIVIAT, N. and DIEHR, P. (2003). The intensity-score approach to adjusting for confounding. *Biometrics* **59** 274–285. MR1987394 https://doi.org/10.1111/1541-0420.00034

Tianchen Qian is Postdoctoral Fellow, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: qiantianchen@fas.harvard.edu). Predrag Klasnja is Assistant Professor, School of Information, University of Michigan, Ann Arbor, Massachusetts 48109, USA (e-mail: klasnja@umich.edu). Susan A. Murphy is Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: samurphy@fas.harvard.edu).

CHEUNG, M. W.-L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling *Psychol. Methods* **13** 182.

CRAINICEANU, C. M. and RUPPERT, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 165–185. MR2035765 https://doi.org/10.1111/j.1467-9868.2004.00438.x

DANIEL, R. M., COUSENS, S. N., DE STAVOLA, B. L., KENWARD, M. G. and STERNE, J. A. C. (2013). Methods for dealing with time-dependent confounding. *Stat. Med.* **32** 1584–1618. MR3060620 https://doi.org/10.1002/sim.5686

DEMPFLE, L. (1977). Comparison of several sire evaluation methods in dairy cattle breeding. *Livest. Prod. Sci.* **4** 129–139.

DEMPSEY, W., LIAO, P., KLASNJA, P., NAHUM-SHANI, I. and MURPHY, S. A. (2015). Randomised trials for the Fitbit generation. *Significance* **12** 20–23. https://doi.org/10.1111/j.1740-9713.2015.00863.x

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

EBBES, P., BÖCKENHOLT, U. and WEDEL, M. (2004). Regressor and random-effects dependencies in multilevel models. *Stat. Neerl.* **58** 161–178. MR2086353 https://doi.org/10.1046/j.0039-0402.2003.00254.x

GARCIA, T. P. and MA, Y. (2016). Optimal estimator for logistic model with distribution-free random intercept. *Scand. J. Stat.* **43** 156–171. MR3466999 https://doi.org/10.1111/sjos.12170

GOETGELUK, S. and VANSTEELANDT, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics* **64** 772–780. MR2526627 https://doi.org/10.1111/j.1541-0420.2007.00944.x

GRILLI, L. and RAMPICHINI, C. (2011). The role of sample cluster means in multilevel models. *Methodol.* **7** 121–133.

HANCHANE, S. and MOSTAFA, T. (2012). Solving endogeneity problems in multilevel estimation: An example using education production functions. *J. Appl. Stat.* **39** 1101–1114. MR2914514 https://doi.org/10.1080/02664763.2011.638705

HAUSMAN, J. A. and TAYLOR, W. E. (1981). Panel data and unobservable individual effects. *Econometrica* **49** 1377–1398. MR0636158 https://doi.org/10.2307/1911406

HEAGERTY, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55** 688–698.

HEAGERTY, P. J. and ZEGER, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statist. Sci.* **15** 1–26. MR1842235 https://doi.org/10.1214/ss/1009212670

HENDERSON, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 423–447.

HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J. Amer. Statist. Assoc.* **96** 440–448. MR1939347 https://doi.org/10.1198/016214501753168154

HERNÁN, M. A. and ROBINS, J. M. (2019). *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL.

KEOGH, R. H., DANIEL, R. M., VANDERWEELE, T. J. and VANSTEELANDT, S. (2017). Analysis of longitudinal studies with repeated outcome measures: Adjusting for time-dependent confounding using conventional methods. *Am. J. Epidemiol.* **187** 1085–1092.

KIM, J.-S. and FREES, E. W. (2006). Omitted variables in multilevel models. *Psychometrika* **71** 659–690. MR2312237 https://doi.org/10.1007/s11336-005-1283-0

KIM, J.-S. and FREES, E. W. (2007). Multilevel modeling with correlated effects. *Psychometrika* **72** 505–533. MR2377551 https://doi.org/10.1007/s11336-007-9008-1

KLASNJA, P., SMITH, S., SEEWALD, N. J., LEE, A., HALL, K., LUERS, B., HEKLER, E. B. and MURPHY, S. A. (2018). Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of HeartSteps. *Annals Behav. Med.* **53** 573–582.

KUZNETSOVA, A., BROCKHOFF, P. B. and CHRISTENSEN, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* **82** 1–26.

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LARSEN, K., PETERSEN, J. H., BUDTZ-JØRGENSEN, E. and ENDAHL, L. (2000). Interpreting parameters in the logistic regression model with random effects. *Biometrics* **56** 909–914.

LIAO, P., KLASNJA, P., TEWARI, A. and MURPHY, S. A. (2016). Sample size calculations for micro-randomized trials in mHealth. *Stat. Med.* **35** 1944–1971. MR3513494 https://doi.org/10.1002/sim.6847

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41. MR0415861

LIU, L. and XIANG, L. (2014). Semiparametric estimation in generalized linear mixed models with auxiliary covariates: A pairwise likelihood approach. *Biometrics* **70** 910–919. MR3295752 https://doi.org/10.1111/biom.12208

LUGER, T. M., SULS, J. and VANDER WEG, M. W. (2014). How robust is the association between smoking and depression in adults? A meta-analysis using linear mixed-effects models. *Addict. Behav.* **39** 1418–1429.

MIGLIORETTI, D. L. and HEAGERTY, P. J. (2004). Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* **5** 381–398.

MUNDLAK, Y. (1978). On the pooling of time series and cross section data. *Econometrica* **46** 69–85. MR0478489 https://doi.org/10.2307/1913646

NEUGEBAUER, R., VAN DER LAAN, M. J., JOFFE, M. M. and TAGER, I. B. (2007). Causal inference in longitudinal studies with history-restricted marginal structural models. *Electron. J. Stat.* **1** 119–154. MR2312147 https://doi.org/10.1214/07-EJS050

NEUHAUS, J. M. and MCCULLOCH, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 859–872. MR2301298 https://doi.org/10.1111/j.1467-9868.2006.00570.x

PAN, W., LOUIS, T. A. and CONNETT, J. E. (2000). A note on marginal linear regression with correlated response data. *Amer. Statist.* **54** 191–195. MR1802670 https://doi.org/10.2307/2685589

PEPE, M. S. and ANDERSON, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Comm. Statist. Simulation Comput.* **23** 939–951.

RAUDENBUSH, S. W. and BRYK, A. S. (2002). *Hierarchical Linear Models*: *Applications and Data Analysis Methods*. Sage.

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758 https://doi.org/10.1016/0270-0255(86)90088-6

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185 https://doi.org/10.1080/03610929408831393

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality* (*Los Angeles*, *CA*, 1994). *Lect. Notes Stat.* **120** 69–117. Springer, New York. MR1601279 https://doi.org/10.1007/978-1-4612-1842-5_4

ROBINS, J. M. (1998). Correction for non-compliance in equivalence trials. *Stat. Med.* **17** 269–302.

ROBINS, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (*Minneapolis, MN*, 1997). *IMA Vol. Math. Appl.* **116** 95–133. Springer, New York. MR1731682 https://doi.org/10.1007/978-1-4612-1284-3_2

ROY, J., ALDERSON, D., HOGAN, J. W. and TASHIMA, K. T. (2006). Conditional inference methods for incomplete Poisson data with endogenous time-varying covariates: Emergency department use among HIV-infected women. *J. Amer. Statist. Assoc.* **101** 424–434. MR2256164 https://doi.org/10.1198/016214505000001203

SATTERTHWAITE, F. E. (1941). Synthesis of variance. *Psychometrika* **6** 309–316. MR0005564 https://doi.org/10.1007/BF02288586

SCHILDCROUT, J. S. and HEAGERTY, P. J. (2005). Regression analysis of longitudinal binary data with time-dependent environmental covariates: Bias and efficiency. *Biostatistics* **6** 633–652.

SCHWARTZ, J. E. and STONE, A. A. (2007). The analysis of real-time momentary data: A practical guide. In *The Science of Real-Time Data Capture: Self-Reports in Health Research* (A. A. Stone, A. Atienza, L. Nebeling and S. Shiffman, eds.) 76–113. Oxford Univ. Press, New York.

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. MR1190470 https://doi.org/10.1002/9780470316856

SELF, S. G. and LIANG, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82** 605–610. MR0898365

SEMYKINA, A. and WOOLDRIDGE, J. M. (2010). Estimating panel data models in the presence of endogeneity and selection. *J. Econometrics* **157** 375–380. MR2661609 https://doi.org/10.1016/j.jeconom.2010.03.039

SHARDELL, M. and FERRUCCI, L. (2018). Joint mixed-effects models for causal inference with longitudinal data. *Stat. Med.* **37** 829–846. MR3760452 https://doi.org/10.1002/sim.7567

SITLANI, C. M., HEAGERTY, P. J., BLOOD, E. A. and TOSTESON, T. D. (2012). Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Stat. Med.* **31** 1738–1760. MR2947521 https://doi.org/10.1002/sim.4510

STRAM, D. O. and LEE, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50** 1171–1177.

TCHETGEN, E. J. T., GLYMOUR, M. M., WEUVE, J. and ROBINS, J. (2012). Specifying the correlation structure in inverse-probability-weighting estimation for repeated measures. *Epidemiology* **23** 644–646.

VANSTEELANDT, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scand. J. Stat.* **34** 478–498. MR2368794 https://doi.org/10.1111/j.1467-9469.2006.00555.x

WANG, Z. and LOUIS, T. A. (2004). Marginalized binary mixed-effects models with covariate-dependent random effects and likelihood inference. *Biometrics* **60** 884–891. MR2133540 https://doi.org/10.1111/j.0006-341X.2004.00243.x

WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA. MR2768559

ZEGER, S. L. and LIANG, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.

ZEGER, S. L. and LIANG, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Stat. Med.* **11** 1825–1839.

ZEGER, S. L., LIANG, K.-Y. and ALBERT, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44** 1049–1060. MR0980999 https://doi.org/10.2307/2531734

Statistical Science
2020, Vol. 35, No. 3, 391–393
https://doi.org/10.1214/20-STS777
© Institute of Mathematical Statistics, 2020

Comment: Clarifying Endogeneous Data Structures and Consequent Modelling Choices Using Causal Graphs

Erica E. M. Moodie and David A. Stephens

REFERENCES

FORAITA, R., SPALLEK, J. and ZEEB, H. (2014). Directed acyclic graphs. In *Handbook of Epidemiology*, 2nd ed. (W. Ahrens and I. Pigeot, eds.) 1481–1517. Springer, New York.

GREENLAND, S., PEARL, J. and ROBINS, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.

LOPEZ, P. M., SUBRAMANIAN, S. V. and SCHOOLING, C. M. (2019). Effect measure modification conceptualized using selection diagrams as mediation by mechanisms of varying population-level relevance. *J. Clin. Epidemiol.* **113** 123–128. https://doi.org/10.1016/j.jclinepi.2019.05.005

MOODIE, E. E. M. and STEPHENS, D. A. (2010). Using Directed Acyclic Graphs to detect limitations of traditional regression in longitudinal studies. *Int. J. Public Health* **55** 701–703.

MOODIE, E. E. M. and STEPHENS, D. A. (2011). Marginal Structural Models: Unbiased estimation for longitudinal studies. *Int. J. Public Health* **56** 117–119.

QIAN, T., KLASNJA, P. and MURPHY, S. A. (2020). Linear mixed models with endogenous covariates: Modeling sequential treatment effects with application to a mobile health study. *Statist. Sci.* **35** 375–390.

VANDERWEELE, T. J. and ROBINS, J. M. (2007). Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am. J. Epidemiol.* **166** 1096–1104.

WEINBERG, C. R. (2007). Can DAGs clarify effect modification? *Epidemiology* **18** 569–572.

Erica E. M. Moodie is Associate Professor, Biostatistics, McGill University, 1020 Pine Ave W., Montreal, Quebec, Canada H3A 1A2 (e-mail: erica.moodie@mcgill.ca). David A. Stephens is Professor, Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Ave W., Montreal, Quebec, Canada H3A 0B9 (e-mail: david.stephens@mcgill.ca).

Statistical Science
2020, Vol. 35, No. 3, 394–395
https://doi.org/10.1214/20-STS781
© Institute of Mathematical Statistics, 2020

Moving Toward Rigorous Evaluation of Mobile Health Interventions

Kristin A. Linn

Abstract. Qian, Klasnja and Murphy provide an assumption that allows for unbiased estimation of treatment effects in microrandomized trials when the data are modeled using linear mixed models with endogenous covariates. In this discussion, the validity of the assumption in the context of the HeartSteps microrandomized trial is reassessed. The utility of a marginal interpretation, versus the proposed conditional-on-the-random-effect interpretation, is also discussed.

Key words and phrases: Linear mixed model, endogenous covariates, microrandomized trial, causal inference.

REFERENCES

COONEY, G. M., DWAN, K., GREIG, C. A., LAWLOR, D. A., RIMER, J., WAUGH, F. R., MCMURDO, M. and MEAD, G. E. (2013). Exercise for depression. *Cochrane Database Syst. Rev.* **9** CD004366.

COTTON, V. and PATEL, M. S. (2019). Gamification use and design in popular health and fitness mobile applications. *Am. J. Health Promot.* **33** 448–451. https://doi.org/10.1177/0890117118790394

DE MOOR, M. H., BEEM, A., STUBBE, J. H., BOOMSMA, D. I. and DE GEUS, E. J. (2006). Regular exercise, anxiety, depression and personality: A population-based study. *Prev. Med.* **42** 273–279.

MICHAELIDES, A., RABY, C., WOOD, M., FARR, K. and TORO-RAMOS, T. (2016). Weight loss efficacy of a novel mobile diabetes prevention program delivery platform with human coaching. *BMJ Open Diabetes Res. Care* **4** e000264.

MILLER, A. S., CAFAZZO, J. A. and SETO, E. (2016). A game plan: Gamification design principles in mHealth applications for chronic disease management. *Health Inform. J.* **22** 184–193.

PATEL, M. S., ASCH, D. A. and VOLPP, K. G. (2015). Wearable devices as facilitators, not drivers, of health behavior change. *JAMA* **313** 459–460. https://doi.org/10.1001/jama.2014.14781

TORO-RAMOS, T., KIM, Y., WOOD, M., RAJDA, J., NIEJADLIK, K., HONCZ, J., MARRERO, D., FAWER, A. and MICHAELIDES, A. (2017). Efficacy of a mobile hypertension prevention delivery platform with human coaching. *J. Hum. Hypertens.* **31** 795–800.

Kristin A. Linn is Assistant Professor of Biostatistics, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Drive, Philadelphia, Pennsylvania 19104, USA (e-mail: klinn@upenn.edu).

Statistical Science
2020, Vol. 35, No. 3, 396–399
https://doi.org/10.1214/20-STS782
© Institute of Mathematical Statistics, 2020

Comment: Diagnostics and Kernel-based Extensions for Linear Mixed Effects Models with Endogenous Covariates

Hunyong Cho, Joshua P. Zitovsky, Xinyi Li, Minxin Lu, Kushal Shah, John Sperger, Matthew C. B. Tsilimigras and Michael R. Kosorok

Abstract. We discuss "Linear mixed models with endogenous covariates: modeling sequential treatment effects with application to a mobile health study" by Qian, Klasnja and Murphy. In this discussion, we study when the linear mixed effects models with endogenous covariates are feasible to use by providing examples and diagnostic tools as well as discussing potential extensions. This includes evaluating feasibility of partial likelihood-based inference, checking the conditional independence assumption, estimation of marginal effects, and kernel extensions of the model.

Key words and phrases: Linear mixed models, partial likelihood, conditional independence test, marginal effects, kernel mixed models.

REFERENCES

ARTIGAS, F., BORTOLOZZI, A. and CELADA, P. (2018). Can we increase speed and efficacy of antidepressant treatments? Part I: General aspects and monoamine-based strategies. *Eur. Neuropsychopharmacol*. **28** 445–456. https://doi.org/10.1016/j.euroneuro.2017.10.032

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

CHO, H., ZITOVSKY, J. P., LI, X., LU, M., SHAH, K., SPERGER, J., TSILIMIGRAS, M. C. B. and KOSOROK, M. R. (2020). Supplement to "Comment: Diagnostics and Kernel-based Extensions for Linear Mixed Effects Models with Endogenous Covariates." https://doi.org/10.1214/20-STS782SUPP

HAJJEM, A., BELLAVANCE, F. and LAROCQUE, D. (2014). Mixed-effects random forest for clustered data. *J. Stat. Comput. Simul*. **84** 1313–1328. MR3169395 https://doi.org/10.1080/00949655.2012.741599

WANG, X., PAN, W., HU, W., TIAN, Y. and ZHANG, H. (2015). Conditional distance correlation. *J. Amer. Statist. Assoc*. **110** 1726–1734. MR3449068 https://doi.org/10.1080/01621459.2014.993081

WONG, W. H. (1986). Theory of partial likelihood. *Ann. Statist*. **14** 88–123. MR0829557 https://doi.org/10.1214/aos/1176349844

Hunyong Cho is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: hunycho@live.unc.edu). Joshua P. Zitovsky is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: joshz@live.unc.edu). Xinyi Li is a postdoctoral fellow, Statistical and Applied Mathematical Sciences Institute (SAMSI) and University of North Carolina at Chapel Hill, Durham/Chapel Hill, North Carolina 27516, USA (e-mail: xli@samsi.info). Minxin Lu is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: mino12@live.unc.edu). Kushal Shah is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: kushshah@live.unc.edu). John Sperger is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: jsperger@live.unc.edu). Matthew C. B. Tsilimigras is a postdoctoral fellow, Department of Epidemiology, Department of Nutrition, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: matthew_tsilimigras@unc.edu). Michael R. Kosorok is the W.R. Kenan, Jr. Distinguished Professor and Chair, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: kosorok@unc.edu).

Statistical Science
2020, Vol. 35, No. 3, 400–403
https://doi.org/10.1214/20-STS794
© Institute of Mathematical Statistics, 2020

Rejoinder: Linear Mixed Models with Endogenous Covariates: Modeling Sequential Treatment Effects with Application to a Mobile Health Study

Tianchen Qian, Predrag Klasnja and Susan A. Murphy

REFERENCES

BORUVKA, A., ALMIRALL, D., WITKIEWITZ, K. and MURPHY, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.* **113** 1112–1121. MR3862343 https://doi.org/10.1080/01621459.2017.1305274

LIU, D., LIN, X. and GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63** 1079–1088, 1311. MR2414585 https://doi.org/10.1111/j.1541-0420.2007.00799.x

MIGLIORETTI, D. L. and HEAGERTY, P. J. (2004). Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* **5** 381–398.

NAHUM-SHANI, I., SMITH, S. N., SPRING, B. J., COLLINS, L. M., WITKIEWITZ, K., TEWARI, A. and MURPHY, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals Behav. Med.* **52** 446–462. https://doi.org/10.1007/s12160-016-9830-8

PEARCE, N. D. and WAND, M. P. (2006). Penalized splines and reproducing kernel methods. *Amer. Statist.* **60** 233–240. MR2246756 https://doi.org/10.1198/000313006X124541

PEARCE, N. D. and WAND, M. P. (2009). Explicit connections between longitudinal data analysis and kernel machines. *Electron. J. Stat.* **3** 797–823. MR2534202 https://doi.org/10.1214/09-EJS428

QIAN, T., YOO, H., KLASNJA, P., ALMIRALL, D. and MURPHY, S. A. (2019). Estimating time-varying causal excursion effect in mobile health with binary outcomes. Preprint. Available at arXiv:1906.00528.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

ROSS, S., PINEAU, J., CHAIB-DRAA, B. and KREITMANN, P. (2011). A Bayesian approach for learning and planning in partially observable Markov decision processes. *J. Mach. Learn. Res.* **12** 1729–1770.

ROY, J., ALDERSON, D., HOGAN, J. W. and TASHIMA, K. T. (2006). Conditional inference methods for incomplete Poisson data with endogenous time-varying covariates: Emergency department use among HIV-infected women. *J. Amer. Statist. Assoc.* **101** 424–434. MR2256164 https://doi.org/10.1198/016214505000001203

SHARDELL, M. and FERRUCCI, L. (2018). Joint mixed-effects models for causal inference with longitudinal data. *Stat. Med.* **37** 829–846. MR3760452 https://doi.org/10.1002/sim.7567

SITLANI, C. M., HEAGERTY, P. J., BLOOD, E. A. and TOSTESON, T. D. (2012). Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Stat. Med.* **31** 1738–1760. MR2947521 https://doi.org/10.1002/sim.4510

SPEED, T. (1991). Comment for "That BLUP is a good thing: The estimation of random effects". *Statist. Sci.* **6** 42–44.

TOMKINS, S., LIAO, P., KLASNJA, P., YEUNG, S. and MURPHY, S. (2020). Rapidly personalizing mobile health treatment policies with limited data. Preprint. Available at arXiv:2002.09971.

VERBYLA, A. P., CULLIS, B. R., KENWARD, M. G. and WELHAM, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **48** 269–311.

WAHBA, G. (1990). *Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442 https://doi.org/10.1137/1.9781611970128

WAND, M. P. (2003). Smoothing and mixed models. *Comput. Statist.* **18** 223–249.

Tianchen Qian is Postdoctoral Fellow, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: qiantianchen@fas.harvard.edu). Predrag Klasnja is Assistant Professor, School of Information, University of Michigan, Ann Arbor, Michigan 48109, USA (e-mail: klasnja@umich.edu). Susan A. Murphy is Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: samurphy@fas.harvard.edu).

Statistical Science
2020, Vol. 35, No. 3, 404–426
https://doi.org/10.1214/19-STS721
© Institute of Mathematical Statistics, 2020

Invariance, Causality and Robustness

2018 Neyman Lecture

Peter Bühlmann

Abstract. We discuss recent work for causal inference and predictive robustness in a unifying way. The key idea relies on a notion of probabilistic invariance or stability: it opens up new insights for formulating causality as a certain risk minimization problem with a corresponding notion of robustness. The invariance itself can be estimated from general heterogeneous or perturbation data which frequently occur with nowadays data collection. The novel methodology is potentially useful in many applications, offering more robustness and better "causal-oriented" interpretation than machine learning or estimation in standard regression or classification frameworks.

Key words and phrases: Anchor regression, causal regularization, distributional robustness, heterogeneous data, instrumental variables regression, interventional data, Random Forests, variable importance.

REFERENCES

ANGRIST, J., IMBENS, G. and RUBIN, D. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

BOTTOU, L., PETERS, J., QUIÑONERO-CANDELA, J., CHARLES, D. X., CHICKERING, D. M., PORTUGALY, E., RAY, D., SIMARD, P. and SNELSON, E. (2013). Counterfactual reasoning and learning systems: The example of computational advertising. *J. Mach. Learn. Res.* **14** 3207–3260. MR3144461

BOWDEN, R. J. and TURKINGTON, D. A. (1990). *Instrumental Variables. Econometric Society Monographs in Quantitative Economics*. Cambridge Univ. Press, Cambridge. MR1113481

BREIMAN, L. (1999). Prediction games and arcing algorithms. *Neural Comput.* **11** 1493–1517.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BRODERSEN, K. H., GALLUSSER, F., KOEHLER, J., REMY, N. and SCOTT, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.* **9** 247–274. MR3341115 https://doi.org/10.1214/14-AOAS788

BÜHLMANN, P. and HOTHORN, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statist. Sci.* **22** 477–505. MR2420454 https://doi.org/10.1214/07-STS242

BÜHLMANN, P., PETERS, J. and ERNEST, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.* **42** 2526–2556. MR3277670 https://doi.org/10.1214/14-AOS1260

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

BÜHLMANN, P. and YU, B. (2003). Boosting with the L_2 loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339. MR1995709 https://doi.org/10.1198/016214503000125

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CHICKERING, D. M. (2002). Optimal structure identification with greedy search *J. Mach. Learn. Res.* **3** 507–554. MR1991085 https://doi.org/10.1162/153244303321897717

CHOW, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **28** 591–605. MR0141193 https://doi.org/10.2307/1910133

DAWID, A. P. (2000). Causal inference without counterfactuals. *J. Amer. Statist. Assoc.* **95** 407–448. MR1803167 https://doi.org/10.2307/2669377

EDITORIAL (2010). Cause and effect. *Nat. Methods* **7** 243.

ERNEST, J. and BÜHLMANN, P. (2015). Marginal integration for nonparametric causal inference. *Electron. J. Stat.* **9** 3155–3194. MR3453973 https://doi.org/10.1214/15-EJS1075

FRIEDBERG, R., TIBSHIRANI, J., ATHEY, S. and WAGER, S. (2018). Local linear forests. Preprint. Available at arXiv:1807.11408.

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

GAO, R., CHEN, X. and KLEYWEGT, A. J. (2017). Wasserstein distributional robustness and regularization in statistical learning. ArXiv preprint. Available at arXiv:1712.06050.

GREENLAND, S., PEARL, J. and ROBINS, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.

GUO, Z., KANG, H., CAI, T. T. and SMALL, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 793–815. MR3849344 https://doi.org/10.1111/rssb.12275

HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. MR0007954 https://doi.org/10.2307/1905714

Peter Bühlmann is Professor, Seminar for Statistics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland (e-mail: buhlmann@stat.math.ethz.ch).

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

HAUSER, A. and BÜHLMANN, P. (2015). Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 291–318. MR3299409 https://doi.org/10.1111/rssb.12071

HEINZE-DEML, C. and MEINSHAUSEN, N. (2017). Conditional variance penalties and domain shift robustness. Preprint. Available at arXiv:1710.11469.

HEINZE-DEML, C. and PETERS, J. (2017). nonlinearICP: Invariant Causal Prediction for Nonlinear Models. R package version 0.1.2.1.

HEINZE-DEML, C., PETERS, J. and MEINSHAUSEN, N. (2018). Invariant causal prediction for nonlinear models. *J. Causal Inf.* **6**.

HERNÁN, M. A. and ROBINS, J. M. (2006). Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* **60** 578–586.

HERNÁN, M. A. and ROBINS, J. M. (2010). *Causal Inference*. CRC, Boca Raton, FL.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HOYER, P., JANZING, D., MOOIJ, J., PETERS, J. and SCHÖLKOPF, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems* 21, 22*nd Annual Conference on Neural Information Processing Systems* (*NIPS* 2008) 689–696.

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

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

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

KANG, H., ZHANG, A., CAI, T. T. and SMALL, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Amer. Statist. Assoc.* **111** 132–144. MR3494648 https://doi.org/10.1080/01621459.2014.994705

KEMMEREN, P., SAMEITH, K., VAN DE PASCH, L., BENSCHOP, J., LENSTRA, T., MARGARITIS, T., O'DUIBHIR, E., APWEILER, E., VAN WAGENINGEN, S. et al. (2014). Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157** 740–752.

KENNEDY, E. H., MA, Z., MCHUGH, M. D. and SMALL, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1229–1245. MR3689316 https://doi.org/10.1111/rssb.12212

KÜNZEL, S., SEKHON, J., BICKEL, P. J. and YU, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. USA* **116** 4156–4165. https://doi.org/10.1073/pnas.1804597116

MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133–3164. MR2549555 https://doi.org/10.1214/09-AOS685

MAATHUIS, M., COLOMBO, D., KALISCH, M. and BÜHLMANN, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7** 247–248.

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

MEINSHAUSEN, N. (2018b). InvariantCausalPrediction: Invariant Causal Prediction. R package version 0.7-2.

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

MEINSHAUSEN, N., HAUSER, A., MOOIJ, J., PETERS, J., VERSTEEG, P. and BÜHLMANN, P. (2016). Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci. USA* **113** 7361–7368.

MURRAY, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *J. Econ. Perspect.* **20** 111–132.

NIE, X. and WAGER, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. Preprint. Available at arXiv:1712.04912.

PAN, S. J. and YANG, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.

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

PETERS, J. and BÜHLMANN, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101** 219–228. MR3180667 https://doi.org/10.1093/biomet/ast043

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. MR3557186 https://doi.org/10.1111/rssb.12167

PFISTER, N., BÜHLMANN, P. and PETERS, J. (2018). Invariant causal prediction for sequential data. *J. Amer. Statist. Assoc.* Published online. https://doi.org/10.1080/01621459.2018.1491403.

PFISTER, N. and PETERS, J. (2017). seqICP: Sequential Invariant Causal Prediction. R package version 1.1.

PRATT, L. Y. (1993). Discriminability-based transfer between neural networks. In *Advances in Neural Information Processing Systems* (*NIPS*) 204–211.

RICHARDSON, T. and SPIRTES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166 https://doi.org/10.1214/aos/1031689015

ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROJAS-CARULLA, M., SCHÖLKOPF, B., TURNER, R. and PETERS, J. (2018). Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **19** 36. MR3862443

ROTHENHÄUSLER, D., BÜHLMANN, P. and MEINSHAUSEN, N. (2019). Causal Dantzig: Fast inference in linear structural equation models with hidden variables under additive interventions. *Ann. Statist.* **47** 1688–1722. MR3911127 https://doi.org/10.1214/18-AOS1732

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2018). Anchor regression: heterogeneous data meets causality. Preprint. Available at arXiv:1801.06229.

SHIMIZU, S., HOYER, P. O., HYVÄRINEN, A. and KERMINEN, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** 2003–2030. MR2274431

SINHA, A., NAMKOONG, H. and DUCHI, J. (2017). Certifiable distributional robustness with principled adversarial training. Preprint. Available at arXiv:1710.10571. Presented at Sixth International Conference on Learning Representations (ICLR 2018).

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. MR1092986

STEKHOVEN, D., MORAES, I., SVEINBJÖRNSSON, G., HENNIG, L., MAATHUIS, M. and BÜHLMANN, P. (2012). Causal stability ranking. *Bioinformatics* **28** 2819–2823.

STOCK, J. H., WRIGHT, J. H. and YOGO, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econom. Statist.* **20** 518–529. MR1973801 https://doi.org/10.1198/073500102288618658

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TUKEY, J. (1954). Causation, regression, and path analysis. In *Statistics and Mathematics in Biology* (O. Kempthorne, ed.) 35–66. Iowa State College Press, Ames, IA.

VANDERWEELE, T. (2015). *Explanation in Causal Inference*: *Methods for Mediation and Interaction*. Oxford Univ. Press.

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

Statistical Science
2020, Vol. 35, No. 3, 427–429
https://doi.org/10.1214/20-STS768
© Institute of Mathematical Statistics, 2020

Comment: Invariance, Causality and Robustness

Vanessa Didelez

REFERENCES

DAWID, A. P. and DIDELEZ, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Stat. Surv.* **4** 184–231. MR2740837 https://doi.org/10.1214/10-SS081

DING, P., VANDERWEELE, T. J. and ROBINS, J. M. (2017). Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika* **104** 291–302. MR3698254 https://doi.org/10.1093/biomet/asx009

GREENLAND, S. and PEARL, J. (2011). Adjustments and their consequences—collapsibility analysis using graphical models. *Int. Stat. Rev.* **79** 401–426.

HERNAN, M. J. (2016). Does water kill? A call for less casual causal inferences. *Ann. Epidemiol.* **26** 674–680.

MIDDLETON, J. A., SCOTT, M. A., DIAKOW, R. and HILL, J. L. (2015). Bias amplification and bias unmasking. *Polit. Anal.* **24** 307–323.

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

PEARL, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence* (*UAI*-10) 417–424. Morgan Kaufmann, San Mateo, CA.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

Vanessa Didelez is Professor, Leibniz Institute for Prevention Research and Epidemiology—BIPS, and Faculty of Mathematics and Computer Science, University of Bremen, Germany (e-mail: didelez@leibniz-bips.de).

Statistical Science
2020, Vol. 35, No. 3, 430–433
https://doi.org/10.1214/20-STS772
© Institute of Mathematical Statistics, 2020

Comment: Invariance and Causal Inference

Stefan Wager

REFERENCES

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

ATHEY, S. and WAGER, S. (2017). Efficient policy learning. Preprint. Available at arXiv:1702.02896.

BAREINBOIM, E. and PEARL, J. (2016). Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. USA* **113** 7345–7352.

BASSE, G. W., FELLER, A. and TOULIS, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106** 487–494. MR3949317 https://doi.org/10.1093/biomet/asy072

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.* **90** 389–405.

DING, P., FELLER, A. and MIRATRIX, L. (2019). Decomposing treatment effect variation. *J. Amer. Statist. Assoc.* **114** 304–317. MR3941256 https://doi.org/10.1080/01621459.2017.1407322

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242 https://doi.org/10.2307/2998560

HAHN, J., TODD, P. and VAN DER KLAAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.

HERNÁN, M. A. and VANDERWEELE, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology* **22** 368.

HIRSHBERG, D. A., MALEKI, A. and ZUBIZARRETA, J. (2019). Minimax linear estimation of the retargeted mean. Preprint. Available at arXiv:1901.10296.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HOTZ, V. J., IMBENS, G. W. and KLERMAN, J. A. (2006). Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program. *J. Labor Econ.* **24** 521–566.

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

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

IMBENS, G. and WAGER, S. (2019). Optimized regression discontinuity designs. *Rev. Econ. Stat.* **101** 264–278.

KITAGAWA, T. and TETENOV, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* **86** 591–616. MR3783340 https://doi.org/10.3982/ECTA13288

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420 https://doi.org/10.1214/12-AOAS583

MEINSHAUSEN, N., HAUSER, A., MOOIJ, J. M., PETERS, J., VERSTEEG, P. and BÜHLMANN, P. (2016). Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci. USA* **113** 7361–7368.

NEYMAN, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Rocz. Nauk Rol.* **10** 1–51.

NIE, X. and WAGER, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. Preprint. Available at arXiv:1712.04912.

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. MR3557186 https://doi.org/10.1111/rssb.12167

ROBINS, J. M. and RICHARDSON, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures* 103–158. Oxford Univ. Press, Oxford.

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762 https://doi.org/10.2307/1912705

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2018). Anchor regression: Heterogeneous data meets causality. Preprint. Available at arXiv:1801.06229.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071 https://doi.org/10.1198/016214504000001880

VAN DER LAAN, M. J. and DUDOIT, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Paper 130, U.C. Berkeley Division of Biostatistics Working Paper Series.

Stefan Wager is Assistant Professor of Operations, Information and Technology, and Assistant Professor of Statistics (by courtesy), Graduate School of Business, Stanford University, Stanford, California 94305, USA (e-mail: swager@stanford.edu).

Statistical Science
2020, Vol. 35, No. 3, 434–436
https://doi.org/10.1214/20-STS797
© Institute of Mathematical Statistics, 2020

Rejoinder: Invariance, Causality and Robustness

Peter Bühlmann

Abstract. We sincerely thank Vanessa Didelez and Stefan Wager for their insightful and inspiring comments. Their views and thoughts on the topic of my article are of great value and truly contribute to put it into greater perspective.

Key words and phrases: Anchor regression, causal regularization, distributional robustness, heterogeneous treatment effects, instrumental variables regression, random forests.

REFERENCES

ĆEVID, D., BÜHLMANN, P. and MEINSHAUSEN, N. (2018). Spectral deconfounding and perturbed sparse linear models. Preprint. Available at arXiv:1811.05352.

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067 https://doi.org/10.1214/11-AOS949

DAWID, A. P. (2020). Decision-theoretic foundations for statistical causality. Preprint. Available at arXiv:2004.12493.

DAWID, A. P. and DIDELEZ, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Stat. Surv.* **4** 184–231. MR2740837 https://doi.org/10.1214/10-SS081

GUO, Z., ĆEVID, D. and BÜHLMANN, P. (2020). Doubly debiased lasso: High-dimensional inference under hidden confounding and measurement errors. Preprint. Available at arXiv:2004.03758.

JAKOBSEN, M. and PETERS, J. (2020). Distributional robustness of K-class estimators and the PULSE. Preprint. Available at arXiv:2005.03353.

NOVEMBRE, J., JOHNSON, T., BRYC, K., KUTALIK, Z., BOYKO, A. R., AUTON, A., INDAP, A., KING, K. S., BERGMANN, S. et al. (2008). Genes mirror geography within Europe. *Nature* **456** 98–101.

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. MR3557186 https://doi.org/10.1111/rssb.12167

PFISTER, N., WILLIAM, E., PETERS, J., AEBERSOLD, R. and BÜHLMANN, P. (2019). Stabilizing variable selection and regression. Preprint. Available at arXiv:1911.01850.

RICHARDSON, T. and ROBINS, J. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. Preprint. Available at http://www.csss.washington.edu/Papers/wp128.pdf.

ROJAS-CARULLA, M., SCHÖLKOPF, B., TURNER, R. and PETERS, J. (2018). Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **19** 36. MR3862443

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2018). Anchor regression: Heterogeneous data meets causality. Preprint. Available at arXiv:1801.06229.

SHAH, R. D., FROT, B., THANEI, G.-A. and MEINSHAUSEN, N. (2018). RSVP-graphs: Fast high-dimensional covariance matrix estimation under latent confounding. *J. Roy. Statist. Soc. Ser. B.* Preprint. Available at arXiv:1811.01076.

THEIL, H. (1953). Repeated least squares applied to complete equation systems. The Hague: Central Planning Bureau.

THEIL, H. (1958). *Economic Forecasts and Policy*. North-Holland, Amsterdam, Netherlands.

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

Peter Bühlmann is Professor of Statistics, Seminar for Statistics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland (e-mail: buhlmann@stat.math.ethz.ch).

Statistical Science
2020, Vol. 35, No. 3, 437–466
https://doi.org/10.1214/19-STS728
© Institute of Mathematical Statistics, 2020

Outcome-Wide Longitudinal Designs for Causal Inference: A New Template for Empirical Studies

Tyler J. VanderWeele, Maya B. Mathur and Ying Chen

Abstract. In this paper, we propose a new template for empirical studies intended to assess causal effects: the outcome-wide longitudinal design. The approach is an extension of what is often done to assess the causal effects of a treatment or exposure using confounding control, but now, over numerous outcomes. We discuss the temporal and confounding control principles for such outcome-wide studies, metrics to evaluate robustness or sensitivity to potential unmeasured confounding for each outcome and approaches to handle multiple testing. We argue that the outcome-wide longitudinal design has numerous advantages over more traditional studies of single exposure-outcome relationships including results that are less subject to investigator bias, greater potential to report null effects, greater capacity to compare effect sizes, a tremendous gain in the efficiency for the research community, a greater policy relevance and a more rapid advancement of knowledge. We discuss both the practical and theoretical justification for the outcome-wide longitudinal design and also the pragmatic details of its implementation, providing publicly available R code.

Key words and phrases: Causal inference, confounding, multiple testing, sensitivity analysis, bias, longitudinal data.

REFERENCES

ABADIE, A. (2018). Statistical non-significance in empirical economics. Working Paper. Available at: https://economics.mit.edu/files/14851.

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Amer. Statist. Assoc.* **91** 444–472.

ANGRIST, J. D. and PISCHKE, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Univ. Press, Princeton.

BARNOW, B. S., CAIN, G. G. and GOLDBERGER, A. S. (1980). Issues in the analysis of selectivity bias. In *Evaluation Studies* **65** (E. E. Stromsdorfer and G. G. Farkas, eds.). Sage, San Francisco.

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

BENJAMIN, D. J., BERGER, J. O., JOHANNESSON, M., NOSEK, B. A., WAGENMAKERS, E.-J., BERK, R., BOLLEN, K. A., BREMBS, B., BROWN, L. et al. (2018). Redefine statistical significance. *Nat. Hum. Behav.* **2** 6–10. https://doi.org/10.1038/s41562-017-0189-z

BERNAL, J. L., CUMMINS, S. and GASPARRINI, A. (2017). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *Int. J. Epidemiol.* **46** 348–355. https://doi.org/10.1093/ije/dyw098

BETANCOURT, T., GILMAN, S., BRENNAN, R., ZAHN, I. and VANDERWEELE, T. J. (2015). Identifying priorities for mental health interventions in war-affected youth: A longitudinal study. *Pediatrics* **136** e344–350.

BLACKWELL, M., HONAKER, J. and KING, G. (2017). A unified approach to measurement error and missing data: Overview and applications. *Sociol. Methods Res.* **46** 303–341. MR3671517 https://doi.org/10.1177/0049124115585360

BOR, J., MOSCOE, E., MUTEVEDZI, P., NEWELL, M. L. and BAERNIGHAUSEN, T. (2014). Regression discontinuity designs in epidemiology: Causal inference without randomized trials. *Epidemiology* **25** 729–737.

BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. T. and ROTHSTEIN, H. R. (2009). *Introduction to Meta-Analysis*. Wiley, New York.

BROSS, I. (1954). Misclassification in 2×2 tables. *Biometrics* **10** 478–486. MR0068796 https://doi.org/10.2307/3001619

Tyler J. VanderWeele is John L. Loeb and Frances Lehman Loeb Professor of Epidemiology, Department of Epidemiology, Harvard University, Boston, Massachusetts, USA (e-mail: tvanderw@hsph.harvard.edu). Maya B. Mathur is Postdoctoral Research Fellow, Department of Epidemiology, Harvard University, Boston, Massachusetts, USA (e-mail: mmathur@stanford.edu). Ying Chen is Research Scientist, Institute for Quantitative Social Science, Harvard University, Boston, Massachusetts, USA (e-mail: yic867@mail.harvard.edu).

CAMERER, C. F., DREBER, A., FORSELL, E., HO, T.-H., HUBER, J., JOHANNESSON, M. et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science* **351** 1433–1436.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

CEPEDA, M. S., BOSTON, R., FARRAR, J. T. and STROM, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am. J. Epidemiol.* **158** 280–287. https://doi.org/10.1093/aje/kwg115

CHEN, Y., HARRIS, S. K., WORTHINGTON, E. L. and VANDERWEELE, T. J. (2018). Religiously or spiritually-motivated forgiveness and subsequent health and well-being among young adults: An outcome-wide analysis. *J. Posit. Psych.* **187** 2355–2364.

CHEN, Y., KUBZANSKY, L. D. and VANDERWEELE, T. J. (2019). Parental warmth and flourishing in mid-life. *Soc. Sci. Med.* **220** 65–72. https://doi.org/10.1016/j.socscimed.2018.10.026

CHEN, Y. and VANDERWEELE, T. J. (2018). Associations of religious upbringing with subsequent health and well-being from adolescence to young adulthood: An outcome-wide analysis. *Am. J. Epidemiol.* **187** 2355–2364. https://doi.org/10.1093/aje/kwy142

COLE, S. R., PLATT, R. W., SCHISTERMAN, E. F., CHU, H., WESTREICH, D., RICHARDSON, D. and POOLE, C. (2010). Illustrating bias due to conditioning on a collider. *Int. J. Epidemiol.* **39** 417–420.

COOK, R. J. and FAREWELL, V. T. (1996). Multiplicity considerations in the design and analysis of clinical trials. *J. Roy. Statist. Soc. Ser. A* **159** 93–110.

CUNADO, J. and DE GRACIA, F. P. (2012). Does education affect happiness? Evidence for Spain. *Soc. Indic. Res.* **108** 185–196.

DANAEI, G., PAN, A., HU, F. B. and HERNÁN, M. A. (2013). Hypothetical lifestyle interventions in middle-aged women and risk of type 2 diabetes: A 24-year prospective study. *Epidemiology* **24** 122–128.

DANAEI, G., TAVAKKOLI, M. and HERNÁN, M. A. (2012). Bias in observational studies of prevalent users: Lessons for comparative effectiveness research from a meta-analysis of statins. *Am. J. Epidemiol.* **175** 250–262.

DING, P. and MIRATRIX, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias (with comments). *J. Causal Infer.* **3** 41–57.

DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368–377.

DING, P., VANDERWEELE, T. J. and ROBINS, J. M. (2017). Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika* **104** 291–302. MR3698254 https://doi.org/10.1093/biomet/asx009

FRANE, A. V. (2015). Are per-family type I error rates relevant in social and behavioral science? *J. Mod. Appl. Stat. Methods* **14** 5.

GARCIA-AYMERICH, J., VARRASO, R., DANAEI, G., CAMARGO, C. A. and HERNÁN, M. A. (2014). Incidence of adult-onset asthma after hypothetical interventions on body mass index and physical activity. An application of the parametric g-formula. *Am. J. Epidemiol.* **179** 20–26.

GELMAN, A. and LOKEN, E. (2014). The statistical crisis in science. *Am. Sci.* **102** 460–465.

GLYMOUR, M. M., WEUVE, J. and CHEN, J. T. (2008). Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: Measurement, selection, and bias. *Neuropsychol. Rev.* **18** 194–213.

GORDON, A., GLAZKO, G., QIU, X. and YAKOVLEV, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Ann. Appl. Stat.* **1** 179–190. MR2393846 https://doi.org/10.1214/07-AOAS102

GOSLING, S. D., RENTFROW, P. J. and SWANN, W. B. JR. (2003). A very brief measure of the big-five personality domains. *J. Res. Pers.* **37** 504–528.

GREENLAND, S. and ROBINS, J. M. (1986). Identifiability, exchangeability, and epidemiologic confounding. *Int. J. Epidemiol.* **15** 413–419.

GREENLAND, S., SCHLESSELMAN, J. J. and CRIQUI, M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am. J. Epidemiol.* **123** 203–208.

GREENLAND, S., SENN, S. J., ROTHMAN, K. J., CARLIN, J. B., POOLE, C., GOODMAN, S. N. and ALTMAN, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology* **31** 337–350.

HASSELBLAD, V. and HEDGES, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychol. Bull.* **117** 167–178.

HAYDEN, D., PAULER, D. K. and SCHOENFELD, D. (2005). An estimator for treatment comparisons among survivors in randomized trials. *Biometrics* **61** 305–310. MR2135873 https://doi.org/10.1111/j.0006-341X.2005.030227.x

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.

HERNÁN, M. A. (2015). Epidemiology to guide decision-making: Moving away from practice-free research. *Am. J. Epidemiol.* **182** 834–839.

HERNÁN, M. A. and ROBINS, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183** 758–764. https://doi.org/10.1093/aje/kwv254

HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference*. Princeton University Press, Chapman & Hall/CRC.

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

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

HUNTER, D. J. (2012). Lessons from genome-wide association studies for epidemiology. *Epidemiology* **23** 363–367.

IACUS, S. M., KING, G. and PORRO, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Polit. Anal.* **20** 1–24.

IMAI, K., KEELE, L. and TINGLEY, D. (2010). A general approach to causal mediation analysis. *Psychol. Methods* **15** 309–334.

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93** 126–132.

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

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

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

IOANNIDIS, J. P. A. (2016). Exposure-wide epidemiology: Revisiting Bradford Hill. *Stat. Med.* **35** 1749–1762. MR3513482 https://doi.org/10.1002/sim.6825

JOHNSON, R. A. and WICHERN, D. (2002). *Multivariate Analysis*. Wiley, New York.

KENNEDY, E. H., KANGOVI, S. and MITRA, N. (2019). Estimating scaled treatment effects with multiple outcomes. *Stat. Methods Med. Res.* **28** 1094–1104. MR3934637 https://doi.org/10.1177/0962280217747130

KIM, E. S. and VANDERWEELE, T. J. (2019). Mediators of the association between religious service attendance and mortality. *Am. J. Epidemiol.*. **188** 96–101.

KING, G., KEOHANE, R. O. and VERBA, S. (1994). *Designing Social Inquiry*: *Scientific Inference in Qualitative Research*. Princeton Univ. Press, Princeton.

KING, G. and NIELSEN, R. (2019). Why propensity scores should not be used for matching. *Polit. Anal.* **27** 4. https://doi.org/10.1017/pan.2019.11

KING, G., TOMZ, M. and WITTENBERG, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *Amer. J. Polit. Sci.* **44** 341–355.

KNOL, M. J., LE CESSIE, S., ALGRA, A., VANDENBROUCKE, J. P. and GROENWOLD, R. H. H. (2012). Overestimation of risk ratios by odds ratios in trials and cohort studies: Alternatives to logistic regression. *CMAJ*, *Can. Med. Assoc. J.* **184** 895–899.

KOENIG, H. G., KING, D. E. and CARSON, V. B. (2012). *Handbook of Religion and Health*, 2nd ed. Oxford Univ. Press, Oxford, New York.

KUTOB, R. M., YUAN, N. P., WERTHEIM, B. C., SBARRA, D. A., LOUCKS, E. B., NASSIR, R., BAREH, G., KIM, M. M., SNETSELAAR, L. G. et al. (2017). Relationship between marital transitions, health behaviors, and health indicators of postmenopausal women: Results from the women's health initiative. *J. Women's Health* (*Larchmt.*) **26** 313–320.

LASH, T. L., FOX, M. P. and FINK, A. K. (2009). *In Applying Quantitative Bias Analysis to Epidemiologic Data*. Spring, New York.

LEE, D. S. and LEMIEUX, T. (2010). Regression discontinuity designs in economics. *J. Econ. Lit.* **2010** 281–355.

LIN, D. Y., PSATY, B. M. and KRONRNAL, R. A. (1998). Assessing the sensitivity of regression results tounmeasured confounders in observational studies. *Biometrics* **54** 948–963.

LINDEN, A., MATHUR, M. B. and VANDERWEELE, T. J. (2019). EVALUE: Stata module for conducting sensitivity analyses for unmeasured confounding in observational studies. Statistical Software Components S458592. Boston College Department of Economics. Revised 16 Feb 2019.

LIPSITCH, M., TCHETGEN TCHETGEN, E. and COHEN, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21** 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb

LITTLE, R. J. A. and RUBIN, D. B. (2014). *Statistical Analysis with Missing Data*, Wiley, Hoboken, NJ.

MARKS, N. F. and LAMBERT, J. D. (1998). Marital status continuity and change among young and midlife adults longitudinal effects on psychological well-being. *J. Fam. Issues* **19** 652–686.

MATHUR, M. B., DING, P., RIDDELL, C. A. and VANDERWEELE, T. J. (2018). Web site and R package for computing E-values. *Epidemiology* **29** e45–e47.

MATHUR, M. B. and VANDERWEELE, T. J. (2018). New metrics for multiple testing with correlated outcomes. Preprint. https://doi.org/10.31219/osf.io/k9g3b.

MATHUR, M. B. and VANDERWEELE, T. J., (2019). Sensitivity analysis for unmeasured confounding in meta-analyses. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2018.1529598

MORGAN, S. L. and WINSHIP, C. (2015). *Counterfactuals and Causal Inference*, 2nd ed. Cambridge Univ. Press, Cambridge, UK.

MYERS, J. A., RASSEN, J. A., GAGNE, J. G., HUYBRECHTS, K. F., SCHNEEWEISS, S., ROTHMAN, K. J., JOFFE, M. M. and GLYNN, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am. J. Epidemiol.* **174** 1213–1222.

NISBETT, R. E., ARONSON, J., BLAIR, C., DICKENS, W., FLYNN, J., HALPERN, D. F. and TURKHEIMER, E. (2012). Intelligence: New findings and theoretical developments. *Am. Psychol.* **67** 130–159.

OGBURN, E. L. and VANDERWEELE, T. J. (2013). Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders. *Biometrika* **100** 241–248. MR3034338 https://doi.org/10.1093/biomet/ass054

OLIVEIRA, R. and TEIXEIRA-PINTO, A. (2015). Analyzing multiple outcomes: Is it really worth the use of multivariate linear regression? *J. Biometr. Biostat.* **6**.

OPEN SCIENCE COLLABORATION (2015). Estimating the reproducibility of psychological science. *Science* **349** aac4716. https://doi.org/10.1126/science.aac4716

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

PEARL, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In *Proc. 26th Conf. Uncert. Artif. Intel.* (*UAI* 2010) (P. Grunwald and P. Spirtes, eds.), 425–432, Association for Uncetainty in Artificial Intelligence, Corvallis, OR.

PIMENTEL, S. D., SMALL, D. S. and ROSENBAUM, P. R. (2016). Constructed second control groups and attenuation of unmeasured biases. *J. Amer. Statist. Assoc.* **111** 1157–1167. MR3561939 https://doi.org/10.1080/01621459.2015.1076342

POWDTHAVEE, N., LEKFUANGFUB, W. N. and WOODEN, M. (2015). What's the good of education on our overall quality of life? A simultaneous equation model of education and life satisfaction for Australia. *J. Behav. Exp. Econ.* **54** 10–21.

ROBINS, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79** 321–334. MR1185134 https://doi.org/10.1093/biomet/79.2.321

ROBINS, J. M. and HERNÁN, M. A. (2009). Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 553–599. CRC Press, Boca Raton, FL. MR1500133

ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROMANO, J. P. and WOLF, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.* **35** 1378–1408. MR2351090 https://doi.org/10.1214/009053606000001622

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROSENTHAL, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull.* **86** 638–641.

ROTHMAN, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology* **1** 43–46.

ROTHMAN, K. J., GREENLAND, S. and LASH, T. L. (2008). *Modern Epidemiology*, 3rd ed. Lippincott.

RUBIN, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statist. Sci.* **21** 299–309. MR2339125 https://doi.org/10.1214/088342306000000114

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795 https://doi.org/10.1214/08-AOAS187

RUBIN, D. B. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?. *Stat. Med.* **28** 1420–1423. MR2724703 https://doi.org/10.1002/sim.3565

SCHULER, M. and ROSE, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *Am. J. Epidemiol.* **185** 65–73.

SHOR, E., ROELFS, D. J., BUGYI, P. and SCHWARTZ, J. E. (2012). Meta-analysis of marital dissolution and mortality: Reevaluating the intersection of gender and age. *Soc. Sci. Med.* **75** 46–59. https://doi.org/10.1016/j.socscimed.2012.03.010

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.

SJØLANDER, A. (2009). Letter to the editor. *Stat. Med.* **28** 1416–1420.

SMITH, L. and VANDERWEELE, T. J. (2019). Bounding bias due to selection. *Epidemiology* **30** 509–516.

SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573 https://doi.org/10.1198/016214506000000636

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

STUTZER, A. and FREY, B. S. (2006). Does marriage make people happy, or do happy people get married? *J. Socio-Econ.* **35** 326–347.

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779

TUKEY, J. W. (1980). We need both exploratory and confirmatory. *Amer. Statist.* **34** 23–25.

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics*. Springer, New York. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1

VAN DER LAAN, M. J. and ROSE, S. (2018). *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies. Springer Series in Statistics*. Springer, Cham. MR3791826 https://doi.org/10.1007/978-3-319-65304-4

VANDERWEELE, T. J. (2009). On the distinction between interaction and effect modification. *Epidemiology* **20** 863–871.

VANDERWEELE, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford Univ. Press, New York.

VANDERWEELE, T. J. (2017a). Outcome-wide epidemiology. *Epidemiology* **28** 399–402.

VANDERWEELE, T. J. (2017b). On the promotion of human flourishing. *Proc. Natl. Acad. Sci. USA* **31** 8148–8156.

VANDERWEELE, T. J. (2017c). Religious communities and human flourishing. *Curr. Dir. Psychol. Sci.* **26** 476–481.

VANDERWEELE, T. J. (2017d). On a square-root transformation of the odds ratio for a common outcome. *Epidemiology* **28** e58–e60.

VANDERWEELE, T. J. (2019). Principles of confounder selection. *Eur. J. Epidemiol.* **34** 211–219.

VANDERWEELE, T. J. and ARAH, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* **22** 42–52. https://doi.org/10.1097/EDE.0b013e3181f74493

VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Ann. Intern. Med.* **167** 268–274. https://doi.org/10.7326/M16-2607

VANDERWEELE, T. J., DING, P. and MATHUR, M. (2019). Technical considerations in the use of the E-value. *J. Causal Inference*. **7** 1–11. https://doi.org/10.1515/jci-2018-0007

VANDERWEELE, T. J. and HERNÁN, M. A. (2012). Results on differential and dependent measurement error of the exposure and the outcome using signed DAGs. *Am. J. Epidemiol.* **175** 1303–1310.

VANDERWEELE, T. J., JACKSON, J. W. and LI, S. (2016). Causal inference and longitudinal data: A case study of religion and mental health. *Soc. Psychiatry Psychiatr. Epidemiol.* **51** 1457–1466.

VANDERWEELE, T. J. and LI, Y. (2019). Simple sensitivity analysis for differential measurement error. *Am. J. Epidemiol.* **188** 1823–1829.

VANDERWEELE, T. J. and MATHUR, M. B. (2019). Some desirable properties of the Bonferroni correction: Is the Bonferroni correction really so bad? *Am. J. Epidemiol.* **188** 617–618.

VANDERWEELE, T. J, MATHUR, M. B and CHEN, Y. (2020). Supplement to "Outcome-Wide Longitudinal Designs for Causal Inference: A New Template for Empirical Studies." https://doi.org/10.1214/19-STS728SUPP.

VANDERWEELE, T. J. and SHPITSER, I. (2011). A new criterion for confounder selection. *Biometrics* **67** 1406–1413. MR2872391 https://doi.org/10.1111/j.1541-0420.2011.01619.x

VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2014). Attributing effects to interactions. *Epidemiology* **25** 711–722.

VANDERWEELE, T. J. and VANSTEELANDT, S. (2013). Mediation analysis with multiple mediators. *Epidemiol. Methods* **2** 95–115.

WAITE, L. J. and GALLAGHER, M. (2000). *The Case for Marriage*. Doubleday, New York.

WEINBERG, C. R. (1993). Toward a clearer definition of confounding. *Am. J. Epidemiol.* **137** 1–8.

WEINBERG, C. A., UMBACH, D. M. and GREENLAND, S. (1994). When will nondifferential misclassification of an exposure preserve the direction of a trend?. *Am. J. Epidemiol.* **140** 565–571.

WELTER, D., MACARTHUR, J., MORALES, J., BURDETT, T., HALL, P., JUNKINS, H., KLEMM, A., FLICEK, P., MANOLIO, T. et al. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42** (Database issue), D1001-D1006.

WILCOX, W. B. (2011). *Why Marriage Matters: 30 Conclusions from the Social Sciences*, 3rd ed. Institute for American Values/National Marriage Project, New York.

WOOLDRIDGE, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.

YELLAND, L. N., SALTER, A. B. and RYAN, P. (2011). Relative risk estimation in randomized controlled trials: A comparison of methods for independent observations. *Int. J. Biostat.* **7** 5. MR2753573 https://doi.org/10.2202/1557-4679.1278

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348–368. MR0139235

ZILIAK, S. T. and MCCLOSKEY, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. Economics, Cognition, and Society*. Univ. Michigan Press, Ann Arbor, MI. MR2730043 https://doi.org/10.3998/mpub.186351

Statistical Science
2020, Vol. 35, No. 3, 467–471
https://doi.org/10.1214/20-STS769
© Institute of Mathematical Statistics, 2020

Comment: On the Potential for Misuse of Outcome-Wide Study Designs, and Ways to Prevent It

Stijn Vansteelandt and Oliver Dukes

REFERENCES

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

BELLONI, A., CHERNOZHUKOV, V. and WEI, Y. (2016). Post-selection inference for generalized linear models with many controls. *J. Bus. Econom. Statist.* **34** 606–619. MR3547999 https://doi.org/10.1080/07350015.2016.1166116

BRANSON, Z. and BIND, M.-A. (2019). Randomization-based inference for Bernoulli trial experiments and implications for observational studies. *Stat. Methods Med. Res.* **28** 1378–1398. MR3941082 https://doi.org/10.1177/0962280218756689

CHAUSSÉ, P. (2010). Computing generalized method of moments and generalized empirical likelihood with R. *J. Stat. Softw.* **34**.

DUKES, O., AVAGYAN, V. and VANSTEELANDT, S. (2020). Doubly robust tests of exposure effects under high-dimensional confounding. *Biometrics*. https://doi.org/10.1111/biom.13231

DUKES, O. and VANSTEELANDT, S. (2020). How to obtain valid tests and confidence intervals after propensity score variable selection? *Stat. Methods Med. Res.* **29** 677–694. MR4078242 https://doi.org/10.1177/0962280219862005

FARRELL, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Econometrics* **189** 1–23. MR3397349 https://doi.org/10.1016/j.jeconom.2015.06.017

LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856 https://doi.org/10.1017/S0266466605050036

VANDERWEELE, T. J., MATHUR, M. B. and CHEN, Y. (2020). Outcome-wide longitudinal designs for causal inference: A new template for empirical studies. *Statist. Sci.* **35** 437–466. https://doi.org/10.1214/19-STS728

Stijn Vansteelandt is Professor, Department of Applied Mathematics, Computer Sciences and Statistics, Ghent University, Krijgslaan 281, S9, 9000 Gent, Belgium (e-mail: stijn.vansteelandt@ugent.be). Stijn Vansteelandt is also affiliated as a part-time professor in the Department of Medical Statistics of the London School of Hygiene and Tropical Medicine, London, United Kingdom. Oliver Dukes is Postdoctoral Researcher, Department of Applied Mathematics, Computer Sciences and Statistics, Ghent University, Krijgslaan 281, S9, 9000 Gent, Belgium (e-mail: oliver.dukes@ugent.be).

Statistical Science
2020, Vol. 35, No. 3, 472–475
https://doi.org/10.1214/20-STS771
© Institute of Mathematical Statistics, 2020

Comment: Outcome-Wide Individualized Treatment Strategies

Ashkan Ertefaie and Brent A. Johnson

REFERENCES

BENKESER, D., CARONE, M., VAN DER LAAN, M. J. and GILBERT, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104** 863–880. MR3737309 https://doi.org/10.1093/biomet/asx053

ERTEFAIE, A. and STRAWDERMAN, R. L. (2019). Robust Q-learning Technical report. Available at arXiv:2003.12427.

ERTEFAIE, A., WU, T., LYNCH, K. G. and NAHUM-SHANI, I. (2016). Identifying a set that contains the best dynamic treatment regimes. *Biostatistics* **17** 135–148. MR3449856 https://doi.org/10.1093/biostatistics/kxv025

FARD, M. M. and PINEAU, J. (2009). MDPs with non-deterministic policies. In *Advances in Neural Information Processing Systems* 1065–1072.

HSU, J. C. (1981). Simultaneous confidence intervals for all distances from the "best". *Ann. Statist.* **9** 1026–1034. MR0628758

HSU, J. C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best. *Ann. Statist.* **12** 1136–1144. MR0751303 https://doi.org/10.1214/aos/1176346732

LABER, E. B., LIZOTTE, D. J. and FERGUSON, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics* **70** 53–61. MR3251666 https://doi.org/10.1111/biom.12132

LIAW, A., WIENER, M. et al. (2002). Classification and regression by randomForest. *R News* **2** 18–22.

LIZOTTE, D. J. and LABER, E. B. (2016). Multi-objective Markov decision processes for data-driven decision support. *J. Mach. Learn. Res.* **17** Art. ID 211. MR3595145

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

ROSENBAUM, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika* **75** 577–581. MR0967598 https://doi.org/10.1093/biomet/75.3.577

ROSENBAUM, P. R. (1989). On permutation tests for hidden biases in observational studies: An application of Holley's inequality to the Savage lattice. *Ann. Statist.* **17** 643–653. MR0994256 https://doi.org/10.1214/aos/1176347131

ROSENBAUM, P. R. (1991). Sensitivity analysis for matched case-control studies. *Biometrics* **47** 87–100. MR1108691 https://doi.org/10.2307/2532498

ROSENBAUM, P. R. (2002). Overt bias in observational studies. In *Observational Studies* 71–104. Springer, Berlin.

VAN DER LAAN, M. J. and DUDOIT, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. U.C. Berkeley Division of Biostatistics Working Paper Series 130.

VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Ann. Intern. Med.* **167** 268–274. https://doi.org/10.7326/M16-2607

VAN DER LAAN, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int. J. Biostat.* **10** 29–57. MR3208072 https://doi.org/10.1515/ijb-2012-0038

VAN DER LAAN, M. J., BENKESER, D. and CAI, W. (2019). Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. Preprint. Available at arXiv:1908.05607.

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. ID 25. MR2349918 https://doi.org/10.2202/1544-6115.1309

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, Berlin.

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. MR3040007 https://doi.org/10.1111/j.1541-0420.2012.01763.x

ZHAO, Q., SMALL, D. S. and ERTEFAIE, A. (2017). Selective inference for effect modification via the lasso. Preprint. Available at arXiv:1705.08020.

Ashkan Ertefaie is an Assistant Professor, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York 14642, USA (e-mail: ashkan_ertefaie@urmc.rochester.edu). Brent A. Johnson is a Professor, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York 14642, USA (e-mail: brent_johnson@urmc.rochester.edu).

Statistical Science
2020, Vol. 35, No. 3, 476–478
https://doi.org/10.1214/20-STS776
© Institute of Mathematical Statistics, 2020

A New Template for Empirical Studies: From positivity to Positivity

Rhian Daniel

REFERENCES

[1] PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537 https://doi.org/10.1177/0962280210386207

[2] STERNE, J. A. C., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. and CARPENTER, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *Br. Med. J.* **338** b2393.

[3] VANDERWEELE, T. J., MATHUR, M. B. and CHEN, Y. (2020). Outcome-wide longitudinal designs for causal inference: A new template for empirical studies. *Statist. Sci.* **35** 437–466.

Rhian Daniel is a Reader in Medical Statistics, Division of Population Medicine, Cardiff University, Cardiff, United Kingdom (e-mail: danielr8@cardiff.ac.uk).

Statistical Science
2020, Vol. 35, No. 3, 479–483
https://doi.org/10.1214/20-STS791
© Institute of Mathematical Statistics, 2020

Rejoinder: The Future of Outcome-Wide Studies

Tyler J. VanderWeele, Maya B. Mathur and Ying Chen

REFERENCES

AMRHEIN, V., GREENLAND, S. and MCSHANE, B. (2019). Scientists rise up against statistical significance. *Nature* **567** 305–307. https://doi.org/10.1038/d41586-019-00857-9

CHEN, Y., KUBZANSKY, L. D. and VANDERWEELE, T. J. (2019). Parental warmth and flourishing in mid-life. *Soc. Sci. Med.* **220** 65–72.

CHEN, Y. and VANDERWEELE, T. J. (2018). Associations of religious upbringing with subsequent health and well-being from adolescence to young adulthood: An outcome-wide analysis. *Am. J. Epidemiol.* **187** 2355–2364. https://doi.org/10.1093/aje/kwy142

CHEN, Y., KIM, E. S., KOH, H. K., FRAZIER, A. L. and VANDERWEELE, T. J. (2019a). Sense of mission and subsequent health and well-being among young adults: An outcome-wide analysis. *Am. J. Epidemiol.* **188** 664–673.

CHEN, Y., HARRIS, S. K., WORTHINGTON, E. L. JR. and VANDERWEELE, T. J. (2019b). Religiously or spiritually-motivated forgiveness and subsequent health and well-being among young adults: An outcome-wide analysis. *J. Posit. Psychol.* **14** 649–658. https://doi.org/10.1080/17439760.2018.1519591

CHEN, Y., HAINES, J., CHARLTON, B. M. and VANDERWEELE, T. J. (2019c). Positive parenting improves multiple aspects of health and well-being in young adulthood. *Nat. Hum. Behav.* **3** 684–691.

DANIEL, R. (2020). A new template for empirical studies: From positivity to Positivity. *Statist. Sci.* **35** 476–478.

ERTEFAIE, A. and JOHNSON, B. A. (2020). Comment: Outcome-wide individualized treatment strategies. *Statist. Sci.* **35** 472–475.

GELMAN, A. and LOKEN, E. (2014). The statistical crisis in science. *Am. Sci.* **102** 460–465.

GREENLAND, S. (2017). Invited commentary: The need for cognitive science in methodology. *Am. J. Epidemiol.* **186** 639–645. https://doi.org/10.1093/aje/kwx259

KIM, E. S., WHILLANS, A. V., LEE, M. T., CHEN, Y. and VANDERWEELE, T. J. (2020). Volunteering and subsequent health and well-being in older adults: An outcome-wide longitudinal approach. *Am. J. Prev. Med.* **59** 176–186.

LASH, T. L., FOX, M. P. and FINK, A. K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer, New York.

LONG, K. N., KIM, E. S., CHEN, Y., WILSON, M. F., WORTHINGTON, E. L. JR and VANDERWEELE, T. J. (2020). The role of hope in subsequent health and well-being for older adults: An outcome-wide longitudinal approach. *Global Epidemiol.* **2** 100018.

LUEDTKE, A., SADIKOVA, E. and KESSLER, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clin. Psychol. Sci.* **7** 445–461.

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. *J. Causal Inference* **3** 61–95.

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

MATHUR, M. B. and VANDERWEELE, T. J. (2018). New metrics for multiple testing with correlated outcomes. Preprint. https://doi.org/10.31219/osf.io/k9g3b

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2

ROTHMAN, K. J., GREENLAND, S. and LASH, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins, Philadelphia, PA. 345–380.

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.

VANDERWEELE, T. J. (2015). *Explanation in Causal Inference*: *Methods for Mediation and Interaction*. Oxford Univ. Press, New York.

VANDERWEELE, T. J. (2017). On the promotion of human flourishing. *Proc. Natl. Acad. Sci. USA* **31** 8148–8156.

VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Ann. Intern. Med.* **167** 268–274. https://doi.org/10.7326/M16-2607

VANDERWEELE, T. J., DING, P. and MATHUR, M. B. (2019). Technical considerations in the use of the E-value. *J. Causal Inference* **7** 1–11. https://doi.org/doi.org/10.1515/jci-2018-0007

VANDERWEELE, T. J. and MATHUR, M. B. (2020). Commentary: Developing best practice guidelines for the reporting of E-values. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyaa094

VANDERWEELE, T. J., MATHUR, M. B. and CHEN, Y. (2020). Outcome-wide longitudinal designs for causal inference: A new template for empirical studies. *Statist. Sci.* **35** 437–466.

VANDERWEELE, T. J., MATHUR, M. B. and DING, P. (2019). Correcting misinterpretations of the E-value. *Ann. Intern. Med.* **170** 131–132.

VANDERWEELE, T. J., MCNEELY, E. and KOH, H. K. (2019). Reimagining health: Flourishing. *J. Am. Med. Assoc.* **321** 1667–1668.

Tyler J. VanderWeele is John L. Loeb and Frances Lehman Loeb Professor of Epidemiology, Harvard T.H. Chan School of Public Health, and Institute for Quantitative Social Science, Harvard University, Boston, Massachusetts 02115, United States (e-mail: tvanderw@hsph.harvard.edu). Maya B. Mathur is Assistant Professor, Quantitative Sciences Unit, Stanford University, Palo Alto, California 94304, United States. Ying Chen is Research Associate, Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts 02138, United States.

VANDERWEELE, T. J., LUEDTKE, A. R., VAN DER LAAN, M. J. and KESSLER, R. C. (2019). Selecting optimal subgroups for treatment using many covariates. *Epidemiology* **30** 334–341. https://doi.org/10.1097/EDE.0000000000000991

VANSTEELANDT, S. and DUKES, O. (2020). On the potential for misuse of outcome-wide study designs, and ways to prevent it. *Statist. Sci*. To appear.

Statistical Science
2020, Vol. 35, No. 3, 484–495
https://doi.org/10.1214/19-STS735
© Institute of Mathematical Statistics, 2020

A Nonparametric Super-Efficient Estimator of the Average Treatment Effect

David Benkeser, Weixin Cai and Mark J. van der Laan

Abstract. Doubly robust estimators are a popular means of estimating causal effects. Such estimators combine an estimate of the conditional mean of the outcome given treatment and confounders (the so-called outcome regression) with an estimate of the conditional probability of treatment given confounders (the propensity score) to generate an estimate of the effect of interest. In addition to enjoying the double-robustness property, these estimators have additional benefits. First, flexible regression tools, such as those developed in the field of machine learning, can be utilized to estimate the relevant regressions, while the estimators of the treatment effects retain desirable statistical properties. Furthermore, these estimators are often statistically efficient, achieving the lower bound on the variance of regular, asymptotically linear estimators. However, in spite of their asymptotic optimality, in problems where causal estimands are weakly identifiable, these estimators may behave erratically. We propose new estimation techniques for use in these challenging settings. Our estimators build on two existing frameworks for efficient estimation: targeted minimum loss estimation and one-step estimation. However, rather than using an estimate of the propensity score in their construction, we instead opt for an alternative regression quantity when building our estimators: the conditional probability of treatment given the conditional mean outcome. We discuss the theoretical implications and demonstrate the estimators' performance in simulated and real data.

Key words and phrases: Causal inference, average treatment effect, asymptotic linearity, efficient influence function, collaborative targeted minimum loss estimation, super efficiency.

REFERENCES

BENKESER, D., CAI, W. and VAN DER LAAN, M. J. (2020). Supplement to "A Nonparametric Super-Efficient Estimator of the Average Treatment Effect." https://doi.org/10.1214/19-STS735SUPPA, https://doi.org/10.1214/19-STS735SUPPB.

BENKESER, D. and VAN DER LAAN, M. J. (2016). The highly adaptive lasso estimator. In *Proceedings of the International Conference on Data Science and Advanced Analytics* 2016 689–696. https://doi.org/10.1109/DSAA.2016.93.

BENKESER, D., CARONE, M., VAN DER LAAN, M. J. and GILBERT, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104** 863–880. MR3737309 https://doi.org/10.1093/biomet/asx053

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. MR1623559

GILBERT, P. B., JURASKA, M., DECAMP, A. C. et al. (2017). Basis and statistical design of the passive HIV-1 antibody mediated prevention (AMP) test-of-concept efficacy trials. *Stat. Commun. Infec. Dis.* **9** 20160001. MR3743441 https://doi.org/10.1515/scid-2016-0001

GRUBER, S. and VAN DER LAAN, M. J. (2010). An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int. J. Biostat.* **6** Art. 18, 31. MR2653847 https://doi.org/10.2202/1557-4679.1182

IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation: Asymptotic Theory. Applications of Mathematics* **16**. Springer, New York–Berlin. Translated from the Russian by Samuel Kotz. MR0620321

JU, C., SCHWAB, J. and VAN DER LAAN, M. J. (2019). On adaptive propensity score truncation in causal inference. *Stat. Methods Med. Res.* **28** 1741–1760. MR3961963 https://doi.org/10.1177/0962280218774817

David Benkeser is an Assistant Professor at Emory University in the Rollins School of Public Health, Department of Biostatistics and Bioinformatics, Atlanta, Georgia 30322, USA (e-mail: benkeser@emory.edu). Weixin Cai is a quantitative researcher at Citadel LLC, Seattle, Washington 98122, USA. Mark J. van der Laan is the Jiann-Ping Hsu/Karl E. Peace Professor of Biostatistics and Statistics at the University of California, Berkeley, Berkeley, California 94720, USA.

JU, C., GRUBER, S., LENDLE, S. D., CHAMBAZ, A., FRANKLIN, J. M., WYSS, R., SCHNEEWEISS, S. and VAN DER LAAN, M. J. (2019a). Scalable collaborative targeted learning for high-dimensional data. *Stat. Methods Med. Res.* **28** 532–554. MR3903757 https://doi.org/10.1177/0962280217729845

JU, C., WYSS, R., FRANKLIN, J. M., SCHNEEWEISS, S., HÄGGSTRÖM, J. and VAN DER LAAN, M. J. (2019b). Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data. *Stat. Methods Med. Res.* **28** 1044–1063. MR3934634 https://doi.org/10.1177/0962280217744588

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

LUO, W., ZHU, Y. and GHOSH, D. (2017). On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika* **104** 51–65. MR3626479 https://doi.org/10.1093/biomet/asw068

MAGARET, C. A., BENKESER, D. C., WILLIAMSON, B. D., BORATE, B. R., CARPP, L. N., GEORGIEV, I. S., SETLIFF, I., DINGENS, A. S., SIMON, N. et al. (2019). Prediction of VRC01 neutralization sensitivity by HIV-1 gp160 sequence features. *PLoS Comput. Biol.* **15** e1006952.

PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537 https://doi.org/10.1177/0962280210386207

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. *Lecture Notes in Statistics* **13**. Springer, New York-Berlin. With the assistance of W. Wefelmeyer. MR0675954

SHORTREED, S. M. and ERTEFAIE, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* **73** 1111–1122. MR3744525 https://doi.org/10.1111/biom.12679

SILVERMAN, B. W. (1986). *Density Estimation*. CRC Press, London, UK.

STITELMAN, O. M. and VAN DER LAAN, M. J. (2010). Collaborative targeted maximum likelihood for time to event data. *Int. J. Biostat.* **6** Art. 21, 46. MR2659056 https://doi.org/10.2202/1557-4679.1249

VAN DER LAAN, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int. J. Biostat.* **10** 29–57. MR3208072 https://doi.org/10.1515/ijb-2012-0038

VAN DER LAAN, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive Lasso. *Int. J. Biostat.* **13** 20150097, 35. MR3724476 https://doi.org/10.1515/ijb-2015-0097

VAN DER LAAN, M. J. and GRUBER, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *Int. J. Biostat.* **6** Art. 17, 70. MR2653848 https://doi.org/10.2202/1557-4679.1181

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning*: *Causal Inference for Observational and Experimental Data*. *Springer Series in Statistics*. Springer, New York. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11, 40. MR2306500 https://doi.org/10.2202/1557-4679.1043

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

WANG, H., ROSE, S. and VAN DER LAAN, M. J. (2011). Finding quantitative trait loci genes with collaborative targeted maximum likelihood learning. *Statist. Probab. Lett.* **81** 792–796. MR2793745 https://doi.org/10.1016/j.spl.2010.11.001

YOON, H., MACKE, J., WEST JR, A. P., FOLEY, B., BJORKMAN, P. J., KORBER, B. and YUSIM, K. (2015). CATNAP: A tool to compile, analyze and tally neutralizing antibody panels. *Nucleic Acids Res.* **43** W213–W219.

Statistical Science
2020, Vol. 35, No. 3, 496–498
https://doi.org/10.1214/20-STS770
© Institute of Mathematical Statistics, 2020

Comment: Increasing Real World Usage of Targeted Minimum Loss-Based Estimators

Mireille E. Schnitzer

REFERENCES

[1] BAHAMYIROU, A., BLAIS, L., FORGET, A. and SCHNITZER, M. E. (2019). Understanding and diagnosing the potential for bias when using machine learning methods with doubly robust causal estimators. *Stat. Methods Med. Res.* **28** 1637–1650. MR3961955 https://doi.org/10.1177/0962280218772065

[2] BENKESER, D., CARONE, M., VAN DER LAAN, M. J. and GILBERT, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104** 863–880. MR3737309 https://doi.org/10.1093/biomet/asx053

[3] BENKESER, D. and VAN DER LAAN, M. (2016). The highly adaptive lasso estimator. In *Proceedings of the International Conference on Data Science and Advanced Analytics*, *IEEE International Conference on Data Science and Advanced Analytics*, 689–696.

[4] COYLE, J., HEJAZI, N., MALENICA, I., PHILLIPS, R., HUBBARD, A. and VAN DER LAAN, M. J. The hitchhiker's guide to the tlverse or a targeted learning practitioner's handbook.

[5] GRUBER, S. and VAN DER LAAN, M. J. (2010). An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int. J. Biostat.* **6** Art. ID 18. MR2653847 https://doi.org/10.2202/1557-4679.1182

[6] JU, C., GRUBER, S., LENDLE, S. D., CHAMBAZ, A., FRANKLIN, J. M., WYSS, R., SCHNEEWEISS, S. and VAN DER LAAN, M. J. (2019). Scalable collaborative targeted learning for high-dimensional data. *Stat. Methods Med. Res.* **28** 532–554. MR3903757 https://doi.org/10.1177/0962280217729845

[7] JU, C., SCHWAB, J. and VAN DER LAAN, M. J. (2019). On adaptive propensity score truncation in causal inference. *Stat. Methods Med. Res.* **28** 1741–1760. MR3961963 https://doi.org/10.1177/0962280218774817

[8] JU, C., WYSS, R., FRANKLIN, J. M., SCHNEEWEISS, S., HÄGGSTRÖM, J. and VAN DER LAAN, M. J. (2019). Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data. *Stat. Methods Med. Res.* **28** 1044–1063. MR3934634 https://doi.org/10.1177/0962280217744588

[9] PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537 https://doi.org/10.1177/0962280210386207

[10] SCHNITZER, M. E. and CEFALU, M. (2018). Collaborative targeted learning using regression shrinkage. *Stat. Med.* **37** 530–543. MR3748759 https://doi.org/10.1002/sim.7527

[11] SCHNITZER, M. E., LOK, J. J. and GRUBER, S. (2016). Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *Int. J. Biostat.* **12** 97–115. MR3505689 https://doi.org/10.1515/ijb-2015-0017

[12] SCHNITZER, M. E., SANGO, J., FERREIRA GUERRA, S. and VAN DER LAAN, M. J. (2020). Data-adaptive longitudinal model selection in causal inference with collaborative targeted minimum loss-based estimation. *Biometrics*. To appear. Available at https://onlinelibrary.wiley.com/doi/full/10.1111/biom.13135.

[13] VAN DER LAAN, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int. J. Biostat.* **10** 29–57. MR3208072 https://doi.org/10.1515/ijb-2012-0038

[14] VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning*: *Causal Inference for Observational and Experimental Data*. *Springer Series in Statistics*. Springer, New York. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1

[15] ZHENG, W. and VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*: *Causal Inference for Observational and Experimental Data* (M. J. van der Laan and S. Rose, eds.). *Springer Series in Statistics* 459–474. Springer, New York. MR2867139 https://doi.org/10.1007/978-1-4419-9782-1_27

Mireille E. Schnitzer is Associate Professor of Biostatistics, Faculty of Pharmacy and School of Public Health, Université de Montréal, 2940, chemin de Polytechnique, Montréal, Québec, Canada, H3T 1J4 (e-mail: mireille.schnitzer@umontreal.ca).

Statistical Science
2020, Vol. 35, No. 3, 499–502
https://doi.org/10.1214/20-STS773
© Institute of Mathematical Statistics, 2020

Comment: Automated Analyses: Because We Can, Does It Mean We Should?

Susan M. Shortreed and Erica E. M. Moodie

REFERENCES

ALAM, S., MOODIE, E. E. M. and STEPHENS, D. A. (2019). Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. *Stat. Med.* **38** 1690–1702. MR3934814 https://doi.org/10.1002/sim.8075

BAKER, M. (2016). Is there a reproducibility crisis? A nature survey lifts the lid on how researchers view the 'crisis rocking science and what they think will help. *Nature* **533** 452–454.

BENKESER, D., CAI, W. and VAN DER LAAN, M. (2020). A nonparametric super-efficient estimator of the average treatment effect. *Statist. Sci.* **35** 484–495.

EFRON, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331. MR0711106

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 https://doi.org/10.1007/978-1-4899-4541-9

GREENLAND, S. (2008). Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *Am. J. Epidemiol.* **167** 523–9; discussion 530–1. https://doi.org/kwm355[pii]10.1093/aje/kwm355

HERNÁN, M. A. and ROBINS, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183** 758–764. https://doi.org/10.1093/aje/kwv254

JU, C., BENKESER, D. and VAN DER LAAN, M. J. (2019). Robust inference on the average treatment effect using the outcome highly adaptive lasso. *Biometrics*. https://doi.org/10.1111/biom.13121

LANCET EDITORS (2010). Should protocols for observational studies be registered? *Lancet* **375** 348.

LAZER, D., KENNEDY, R., KING, G. and VESPIGNANI, A. (2014). Big data. The parable of Google flu: Traps in big data analysis. *Science* **343** 1203–1205. https://doi.org/10.1126/science.1248506

LITTLE, R. J. and RUBIN, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annu. Rev. Public Health* **21** 121–145. https://doi.org/10.1146/annurev.publhealth.21.1.121

LODER, E., GROVES, T. and MACAULEY, D. (2010). Registration of observational studies: The next step toward research transparency. *BMJ* **340** c950. https://doi.org/10.1136/bmj.c950

LUM, K. and ISAAC, W. (2016). To predict amd serve? *Significance*.

O'NEILL, I. Apollo Landing Terrifying Moments. https://urldefense.com/v3/__https://www.history.com/news/apollo-11-moon-landing-terrifying-moments__;!7TrXCGkIugIq!4-s2OGm68zL4bh2pdYQHkMN56a89SJUDHoYeoBUWwSbZCd8WnK_Icg4Orhxehkk87oE.

OBERMEYER, Z., POWERS, B., VOGELI, C. and MULLAINATHAN, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366** 447–53.

PENG, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance* **12** 30–32.

PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537 https://doi.org/10.1177/0962280210386207

PIRRACCHIO, R. and CARONE, M. (2018). The Balance Super Learner: A robust adaptation of the Super Learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Stat. Methods Med. Res.* **27** 2504–2518. MR3825922 https://doi.org/10.1177/0962280216682055

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROTNITZKY, A., LI, L. and LI, X. (2010). A note on overadjustment in inverse probability weighted estimation. *Biometrika* **97** 997–1001. MR2746169 https://doi.org/10.1093/biomet/asq049

RUBIN, R. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2** 169–88.

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795 https://doi.org/10.1214/08-AOAS187

SCHUEMIE, M. J., RYAN, P. B., HRIPCSAK, G., MADIGAN, D. and SUCHARD, M. A. (2018). Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **376** 1–17.

SHORTREED, S. M. and ERTEFAIE, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* **73** 1111–1122. MR3744525 https://doi.org/10.1111/biom.12679

SMITH, G. C., SEAMAN, S. R., WOOD, A. M., ROYSTON, P. and WHITE, I. R. (2014). Correcting for optimistic prediction in small data sets. *Am. J. Epidemiol.* **180** 318–24. https://doi.org/10.1093/aje/kwu140

VELDKAMP, C. (2017). The human fallibility of scientists: Dealing with error and bias in academic research Ph.D. thesis.

WILLIAMS, R. J., TSE, T., HARLAN, W. R. and ZARIN, D. A. (2010). Registration of observational studies: Is it time? *CMAJ, Can. Med. Assoc. J.* **182** 1638–1642. https://doi.org/10.1503/cmaj.092252

Susan M. Shortreed is Senior Investigator Kaiser Permanente Washignton Heath Research Institute, 1730 Minor Ave, Suite 1300, Seattle, Washington 98101, USA (e-mail: susan.m.shortreed@kp.org). Erica E. M. Moodie is William Dawson Scholar and Associate Professor of Biostatistics, McGill University, 1020 Pine Ave W., Montreal, Quebec, Canada H3A 1A2 (e-mail: erica.moodie@mcgill.ca).

Statistical Science
2020, Vol. 35, No. 3, 503–510
https://doi.org/10.1214/20-STS774
© Institute of Mathematical Statistics, 2020

Comment: Stabilizing the Doubly-Robust Estimators of the Average Treatment Effect under Positivity Violations

Fan Li

Abstract. Doubly-robust estimators within the one-step and TMLE frameworks could exhibit finite-sample bias and excess variability under positivity violations. We comment on how the application of a stabilization factor may improve the efficiency property of one-step estimator and TMLE, and the comparisons with their collaborative counterparts using the adaptive propensity scores.

Key words and phrases: Efficient influence function, overlap weighting, trimming, one-step estimation, targeted maximum likelihood estimation.

REFERENCES

BEMBOM, O. and VAN DER LAAN, M. (2008). Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. Technical Report 230, Division of Biostatstics, Univ. California, Berkeley.

BENKESER, D. and VAN DER LAAN, M. (2016). The highly adaptive lasso estimator. In *Proceedings—3rd IEEE International Conference on Data Science and Advanced Analytics*, *DSAA* 2016 689–696.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144 https://doi.org/10.1093/biomet/asn055

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710. MR1789821 https://doi.org/10.1093/biomet/87.3.706

JU, C., SCHWAB, J. and VAN DER LAAN, M. J. (2019). On adaptive propensity score truncation in causal inference. *Stat. Methods Med. Res.* **28** 1741–1760. MR3961963 https://doi.org/10.1177/0962280218774817

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

LI, F. and LI, F. (2019). Propensity score weighting for causal inference with multiple treatments. *Ann. Appl. Stat.* **13** 2389–2415. MR4037435 https://doi.org/10.1214/19-aoas1282

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LI, F., THOMAS, L. E. and LI, F. (2019). Addressing extreme propensity scores via the overlap weights. *Am. J. Epidemiol.* **1** 250–257.

PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537 https://doi.org/10.1177/0962280210386207

VAN DER LAAN, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive Lasso. *Int. J. Biostat.* **13** Art. ID 20150097. MR3724476 https://doi.org/10.1515/ijb-2015-0097

Fan Li is Assistant Professor, Department of Biostatistics, Yale School of Public Health, 135 College Street, Suite 200, New Haven, Connecticut 06510, USA (e-mail: fan.f.li@yale.edu).

Statistical Science
2020, Vol. 35, No. 3, 511–517
https://doi.org/10.1214/20-STS789
© Institute of Mathematical Statistics, 2020

Rejoinder: A Nonparametric Superefficient Estimator of the Average Treatment Effect

David Benkeser, Weixin Cai and Mark J. van der Laan

REFERENCES

ALAM, S., MOODIE, E. E. M. and STEPHENS, D. A. (2019). Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. *Stat. Med.* **38** 1690–1702. MR3934814 https://doi.org/10.1002/sim.8075

BEMBOM, O. and VAN DER LAAN, M. J. (2008). Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 230.

BENKESER, D. (2020). drtmle: Doubly-robust nonparametric estimation and inference. R package version 1.0.5. https://doi.org/10.5281/zenodo.844836

BENKESER, D. and VAN DER LAAN, M. J. (2016). The highly adaptive lasso estimator. In *Proceedings of the International Conference on Data Science and Advanced Analytics* 2016 689–696. https://doi.org/10.1109/DSAA.2016.93

BENKESER, D., CARONE, M., VAN DER LAAN, M. J. and GILBERT, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104** 863–880. MR3737309 https://doi.org/10.1093/biomet/asx053

BREIMAN, L. (1996). Stacked regressions. *Mach. Learn.* **24** 49–64. https://doi.org/10.1007/bf00117832

CORBETT-DAVIES, S., PIERSON, E., FELLER, A., GOEL, S. and HUQ, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD* '17 797–806. Association for Computing Machinery, New York. https://doi.org/10.1145/3097983.3098095

HARTNETT, K. (2018). To build truly intelligent machines, teach them cause and effect [online; accessed 13-January-2020].

HUBBARD, A. E., KENNEDY, C. J. and VAN DER LAAN, M. J. (2018). Data-adaptive target parameters. In *Targeted Learning in Data Science. Springer Ser. Statist.* 125–142. Springer, Cham. MR3820724

HUBBARD, A. E., KHERAD-PAJOUH, S. and VAN DER LAAN, M. J. (2016). Statistical inference for data adaptive target parameters. *Int. J. Biostat.* **12** 3–19. MR3505683 https://doi.org/10.1515/ijb-2015-0013

KUSNER, M. J., LOFTUS, J., RUSSELL, C. and SILVA, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 4066–4076. Curran Associates, Red Hook, NY.

LAZER, D., KENNEDY, R., KING, G. and VESPIGNANI, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science* **343** 1203–1205.

PETERSEN, M. L. and VAN DER LAAN, M. J. (2014). Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology* **25** 418–426. https://doi.org/10.1097/EDE.0000000000000078

PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537 https://doi.org/10.1177/0962280210386207

PIRRACCHIO, R. and CARONE, M. (2018). The Balance Super Learner: A robust adaptation of the Super Learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Stat. Methods Med. Res.* **27** 2504–2518. MR3825922 https://doi.org/10.1177/0962280216682055

SHPITSER, I. and PEARL, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence* **2** 1219. AAAI Press, Menlo Park, CA; MIT Press, Cambridge, MA.

TIAN, J. and PEARL, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence* 567–573.

VAN DER LAAN, M. J., BENKESER, D. and CAI, W. (2019). Causal inference based on undersmoothing the highly adaptive lasso. In *Proceedings of the* 2019 *AAAI Spring Symposium*. Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

VAN DER LAAN, M. J. and LUEDTKE, A. R. (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. *J. Causal Inference* **3** 61–95.

VAN DER LAAN, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int. J. Biostat.* **10** 29–57. MR3208072 https://doi.org/10.1515/ijb-2012-0038

VAN DER LAAN, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive Lasso. *Int. J. Biostat.* **13** Art. ID 20150097. MR3724476 https://doi.org/10.1515/ijb-2015-0097

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. ID 25. MR2349918 https://doi.org/10.2202/1544-6115.1309

WOLPERT, D. H. (1992). Stacked generalization. *Neural Netw.* **5** 241–259. https://doi.org/10.1016/s0893-6080(05)80023-1

XIAO, Y., MOODIE, E. E. and ABRAHAMOWICZ, M. (2013). Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiol. Methods* **2** 1–20.

David Benkeser is an Assistant Professor of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia 30322, USA (e-mail: benkeser@emory.edu). Weixin Cai is a researcher at Citadel, Seattle, Washington USA. Mark J. van der Laan is Professor of Biostatistics and Statistics, University of California, Berkeley, California 94720, USA.

Statistical Science
2020, Vol. 35, No. 3, 518–539
https://doi.org/10.1214/20-STS786
© Institute of Mathematical Statistics, 2020

On Nearly Assumption-Free Tests of Nominal Confidence Interval Coverage for Causal Parameters Estimated by Machine Learning

Lin Liu, Rajarshi Mukherjee and James M. Robins

Abstract. For many causal effect parameters of interest, doubly robust machine learning (DRML) estimators $\hat{\psi}_1$ are the state-of-the-art, incorporating the good prediction performance of machine learning; the decreased bias of doubly robust estimators; and the analytic tractability and bias reduction of sample splitting with cross-fitting. Nonetheless, even in the absence of confounding by unmeasured factors, the nominal $(1-\alpha)$ Wald confidence interval $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$ may still undercover even in large samples, because the bias of $\hat{\psi}_1$ may be of the same or even larger order than its standard error of order $n^{-1/2}$.

In this paper, we introduce essentially assumption-free tests that (i) can falsify the null hypothesis that the bias of $\hat{\psi}_1$ is of smaller order than its standard error, (ii) can provide a upper confidence bound on the true coverage of the Wald interval, and (iii) are valid under the null under no smoothness/sparsity assumptions on the nuisance parameters. The tests, which we refer to as Assumption Free Empirical Coverage Tests (AFECTs), are based on a U-statistic that estimates part of the bias of $\hat{\psi}_1$.

Our claims need to be tempered in several important ways. First no test, including ours, of the null hypothesis that the ratio of the bias to its standard error is smaller than some threshold δ can be consistent [without additional assumptions (e.g., smoothness or sparsity) that may be incorrect]. Second, the above claims only apply to certain parameters in a particular class. For most of the others, our results are unavoidably less sharp. In particular, for these parameters, we cannot directly test whether the nominal Wald interval $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$ undercovers. However, we can often test the validity of the smoothness and/or sparsity assumptions used by an analyst to justify a claim that the reported Wald interval's actual coverage is no less than nominal. Third, in the main text, with the exception of the simulation study in Section 1, we assume we are in the semisupervised data setting (wherein there is a much larger dataset with information only on the covariates), allowing us to regard the covariance matrix of the covariates as known. In the simulation in Section 1, we consider the setting in which estimation of the covariance matrix is required. In the simulation, we used a data adaptive estimator which performs very well in our simulations, but the estimator's theoretical sampling behavior remains unknown.

Key words and phrases: Causal inference, assumption-free, valid inference, U-statistics, higher-order influence functions.

Lin Liu is Assistant Professor, Institute of Natural Sciences, School of Mathematical Sciences and SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai 200420, China (e-mail: linliu@alumni.tongji.edu.cn). Rajarshi Mukherjee is Assistant Professor, Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: ram521@mail.harvard.edu). James M. Robins is Mitchell L. and Robin LaFoley Dong Professor, Department of

REFERENCES

AYYAGARI, R. (2010). Applications of influence functions to semiparametric regression models. Ph.D. thesis, Harvard Univ. MR2813909

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186** 345–366. MR3343791 https://doi.org/10.1016/j.jeconom.2015.02.014

BHATTACHARYA, R. N. and GHOSH, J. K. (1992). A class of U-statistics and asymptotic normality of the number of k-clusters. *J. Multivariate Anal.* **43** 300–330. MR1193616 https://doi.org/10.1016/0047-259X(92)90038-H

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

CHAKRABORTTY, A. and CAI, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.* **46** 1541–1572. MR3819109 https://doi.org/10.1214/17-AOS1594

CHAPELLE, O., SCHÖLKOPF, B. and ZIEN, A. (2010). *Semi-Supervised Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139. MR1473055 https://doi.org/10.1006/jcss.1997.1504

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105.

KUCHIBHOTLA, A. K. and CHAKRABORTTY, A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. Preprint. Available at arXiv:1804.02605.

LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942 https://doi.org/10.1214/12-AOS989

LIU, L., MUKHERJEE, R. and ROBINS, J. (2020). Supplement to "On nearly assumption-free tests of confidence interval coverage of parameters estimated by machine learning." https://doi.org/10.1214/20-STS786SUPP

MUKHERJEE, R., NEWEY, W. K. and ROBINS, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. Preprint. Available at arXiv:1705.07577.

NEWEY, W. K. and ROBINS, J. M. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. Preprint. Available at arXiv:1801.09138.

RITOV, Y., BICKEL, P. J., GAMST, A. C. and KLEIJN, B. J. K. (2014). The Bayesian analysis of complex, high-dimensional models: Can it be CODA? *Statist. Sci.* **29** 619–639. MR3300362 https://doi.org/10.1214/14-STS483

ROBINS, J. M. and ROTNITZKY, A. (2001). Comments on "Inference for semiparametric models: Some questions and an answer." *Statist. Sinica* **11** 920–936.

ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics*: *Essays in Honor of David A. Freedman* 335–421. IMS, Beachwood, OH.

ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

ROBINS, J. M., ZHANG, P., AYYAGARI, R., LOGAN, R., TCHETGEN TCHETGEN, E., LI, L., LUMLEY, T. and VAN DER VAART, A. (2013). New statistical approaches to semiparametric regression with application to air pollution research. Research Report 175, Health Effects Institute.

ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN TCHETGEN, E. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515

ROSENBAUM, P. R. (2008). Testing hypotheses in order. *Biometrika* **95** 248–252. MR2409727 https://doi.org/10.1093/biomet/asm085

RUBIN, H. and VITALE, R. A. (1980). Asymptotic distribution of symmetric statistics. *Ann. Statist.* **8** 165–170. MR0557561

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999a). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478 https://doi.org/10.2307/2669923

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999b). Rejoinder. *J. Amer. Statist. Assoc.* **94** 1135–1146.

SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. MR0856811 https://doi.org/10.1214/aos/1176350055

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

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

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

ZHENG, W. and VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*. *Springer Series in Statistics* 459–474. Springer, New York. MR2867139 https://doi.org/10.1007/978-1-4419-9782-1_27

Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: robins@hsph.harvard.edu).

Statistical Science
2020, Vol. 35, No. 3, 540–544
https://doi.org/10.1214/20-STS796
© Institute of Mathematical Statistics, 2020

Discussion of "On Nearly Assumption-Free Tests of Nominal Confidence Interval Coverage for Causal Parameters Estimated by Machine Learning"

Edward H. Kennedy, Sivaraman Balakrishnan and Larry Wasserman

REFERENCES

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393. MR1065550

BIRGÉ, L. and MASSART, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29. MR1331653 https://doi.org/10.1214/aos/1176324452

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369. MR1323344

LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24** 659–681. MR1394981 https://doi.org/10.1214/aos/1032894458

LIU, L., MUKHERJEE, R. and ROBINS, J. M. (2020). On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statist. Sci.*. **35** 518–539.

NEWEY, W. K. and ROBINS, J. M. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. Preprint. Available at arXiv:1801.09138.

ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman. Inst. Math. Stat. (IMS) Collect.* **2** 335–421. IMS, Beachwood, OH. MR2459958 https://doi.org/10.1214/193940307000000527

ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515

TSYBAKOV, A. B. (2003). Optimal rates of aggregation. *Learning Theory and Kernel Machines* 303–313.

WANG, L., BROWN, L. D., CAI, T. T. and LEVINE, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *Ann. Statist.* **36** 646–664. MR2396810 https://doi.org/10.1214/009053607000000901

WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference. *Proc. Natl. Acad. Sci. USA.* **117** 16880–16890.

Edward H. Kennedy is Assistant Professor, Department of Statistics & Data Science, Carnegie Mellon University, USA (e-mail: edward@stat.cmu.edu). Sivaraman Balakrishnan is Assistant Professor, Department of Statistics & Data Science, Carnegie Mellon University, USA (e-mail: siva@stat.cmu.edu). Larry Wasserman is University Professor, Department of Statistics & Data Science, Carnegie Mellon University, USA (e-mail: larry@stat.cmu.edu).

Statistical Science
2020, Vol. 35, No. 3, 545–554
https://doi.org/10.1214/20-STS804
© Institute of Mathematical Statistics, 2020

Rejoinder: On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning

Lin Liu, Rajarshi Mukherjee and James M. Robins

REFERENCES

ADAMCZAK, R. (2006). Moment inequalities for U-statistics. *Ann. Probab.* **34** 2288–2314. MR2294982 https://doi.org/10.1214/009117906000000476

ALI, S. M. and SILVEY, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–142. MR0196777

BARRON, A. R. and KLUSOWSKI, J. M. (2018). Approximation and estimation for high-dimensional deep learning networks. Preprint. Available at arXiv:1809.03090.

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186** 345–366. MR3343791 https://doi.org/10.1016/j.jeconom.2015.02.014

CAI, T. T., LEVINE, M. and WANG, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivariate Anal.* **100** 126–136. MR2460482 https://doi.org/10.1016/j.jmva.2008.03.007

CHERNOZHUKOV, V., NEWEY, W. K. and SINGH, R. (2018). Learning L2 Continuous Regression Functionals via Regularized Riesz Representers. Preprint. Available at arXiv:1809.05224.

CSISZÁR, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **8** 85–108. MR0164374

CUI, Y. and TCHETGEN TCHETGEN, E. (2019). Selective machine learning of doubly robust functionals. Preprint. Available at arXiv:1911.02029.

GINÉ, E., LATAŁA, R. and ZINN, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability, II* (*Seattle, WA*, 1999). *Progress in Probability* **47** 13–38. Birkhäuser, Boston, MA. MR1857312

HAYAKAWA, S. and SUZUKI, T. (2020). On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Netw.* **123** 343–361. https://doi.org/10.1016/j.neunet.2019.12.014

KENNEDY, E. H., BALAKRISHNAN, S. and WASSERMAN, L. (2020). Discussion of "On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning" by L. Liu, R. Mukherjee, and J. Robins. *Statist. Sci.*. **35** 540-544.

LIU, L., MUKHERJEE, R. and ROBINS, J. M. (2020a). An assumption-lean skepticism test of inference validity for doubly robust functionals. Technical report. Available upon request.

LIU, L., MUKHERJEE, R. and ROBINS, J. M. (2020). Supplement to "Rejoinder: On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning." https://doi.org/10.1214/20-STS804SUPP

LIU, L., MUKHERJEE, R. and ROBINS, J. M. (2020). On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statist. Sci.*. **35** 518-539.

LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412 https://doi.org/10.1214/aos/1030741084

MUKHERJEE, R., NEWEY, W. K. and ROBINS, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. Preprint. Available at arXiv:1705.07577.

MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168 https://doi.org/10.2307/2669386

NEWEY, W. K. and ROBINS, J. M. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. Preprint. Available at arXiv:1801.09138.

RINALDO, A., WASSERMAN, L. and G'SELL, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.* **47** 3438–3469. MR4025748 https://doi.org/10.1214/18-AOS1784

ROBINS, J. M. (1988). Confidence intervals for causal parameters. *Stat. Med.* **7** 773–785.

ROBINS, J., LI, L., TCHETGEN TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman* 335–421. IMS, Beachwood, OH, USA.

ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

ROBINS, J., LI, L., TCHETGEN TCHETGEN, E. and VAN DER VAART, A. (2016). Technical Report: Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals. Preprint. Available at arXiv:1601.05820.

Lin Liu is Assistant Professor, Institute of Natural Sciences, School of Mathematical Sciences and SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: linliu@sjtu.edu.cn). Rajarshi Mukherjee is Assistant Professor, Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: ram521@mail.harvard.edu). James M. Robins is Professor, Department of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: robins@hsph.harvard.edu).

ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515

ROTNITZKY, A., SMUCLER, E. and ROBINS, J. M. (2019). Characterization of parameters with a mixed bias property. Preprint. Available at arXiv:1904.03725.

SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.*. To appear.

SHEN, Y., GAO, C., WITTEN, D. and HAN, F. (2019). Optimal estimation of variance in nonparametric regression with random design. Preprint. Available at arXiv:1902.10822.

WANG, L., BROWN, L. D., CAI, T. T. and LEVINE, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *Ann. Statist.* **36** 646–664. MR2396810 https://doi.org/10.1214/009053607000000901

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

EDITORIAL POLICY

The aim of *Statistical Science* is to present the full range of contemporary statistical thought at a technical level accessible to the broad community of practitioners, teachers, researchers and students of statistics and probability.

The journal will publish discussions of methodological and theoretical topics of current interest and importance, surveys of substantive research areas with promising statistical applications, comprehensive book reviews, discussions of classic articles from the statistical literature and interviews with distinguished statisticians and probabilists.

The opinions expressed are those of the authors and do not necessarily reflect those of the editors or of the IMS.

GENERAL INFORMATION

Journal Homepage: http://www.imstat.org/journals-and-publications/statistical-science/

Submissions: Manuscripts for *Statistical Science* should be submitted online. Authors may access the Electronic Journals Management System (EJMS) at http://www.e-publications.org/ims/submission/

Preparation of Manuscripts: www.imstat.org/journals-and-publications/statistical-science/statistical-science-manuscript-preparation/

Permissions Policy. Authorization to photocopy items for internal or personal use is granted by the Institute of Mathematical Statistics. For multiple copies or reprint permission, contact The Copyright Clearance Center, 222 Rosewood Drive, Danvers, Massachusetts 01923. Telephone (978) 750-8400. http://www.copyright.com. If the permission is not found at the Copyright Clearance Center, please contact the IMS Business Office: ims@imstat.org

Correspondence. Mail concerning membership, subscriptions, nonreceipt claims, copyright permissions, advertising or back issues should be sent to the IMS Dues and Subscription Office, 9650 Rockville Pike, Suite L 2310, Bethesda, Maryland 20814-3998. Mail concerning submissions or editorial content should be sent to the Editor of the appropriate journal. Addresses are listed on the inside back cover. Mail concerning the production of this journal should be sent to: Patrick Kelly, IMS Production Editor, Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340 USA

Individual Memberships: http://www.imstat.org/individual-membership/

Institutional Journal Subscriptions: https://imstat.org/journals-and-publications/subscriptions-and-orders/

Electronic Access to IMS Journals: http://www.imstat.org/journals-and-publications/electronic-access/

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.

IMS OFFICERS

President: Susan Murphy, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

President-Elect: Regina Y. Liu, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854-8019, USA

Past President: Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

Executive Secretary: Edsel Peña, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208-001, USA

Treasurer: Zhengjun Zhang, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1510, USA

Program Secretary: Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027-5927, USA

IMS EDITORS

The Annals of Statistics. *Editors*: Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK. Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027, USA

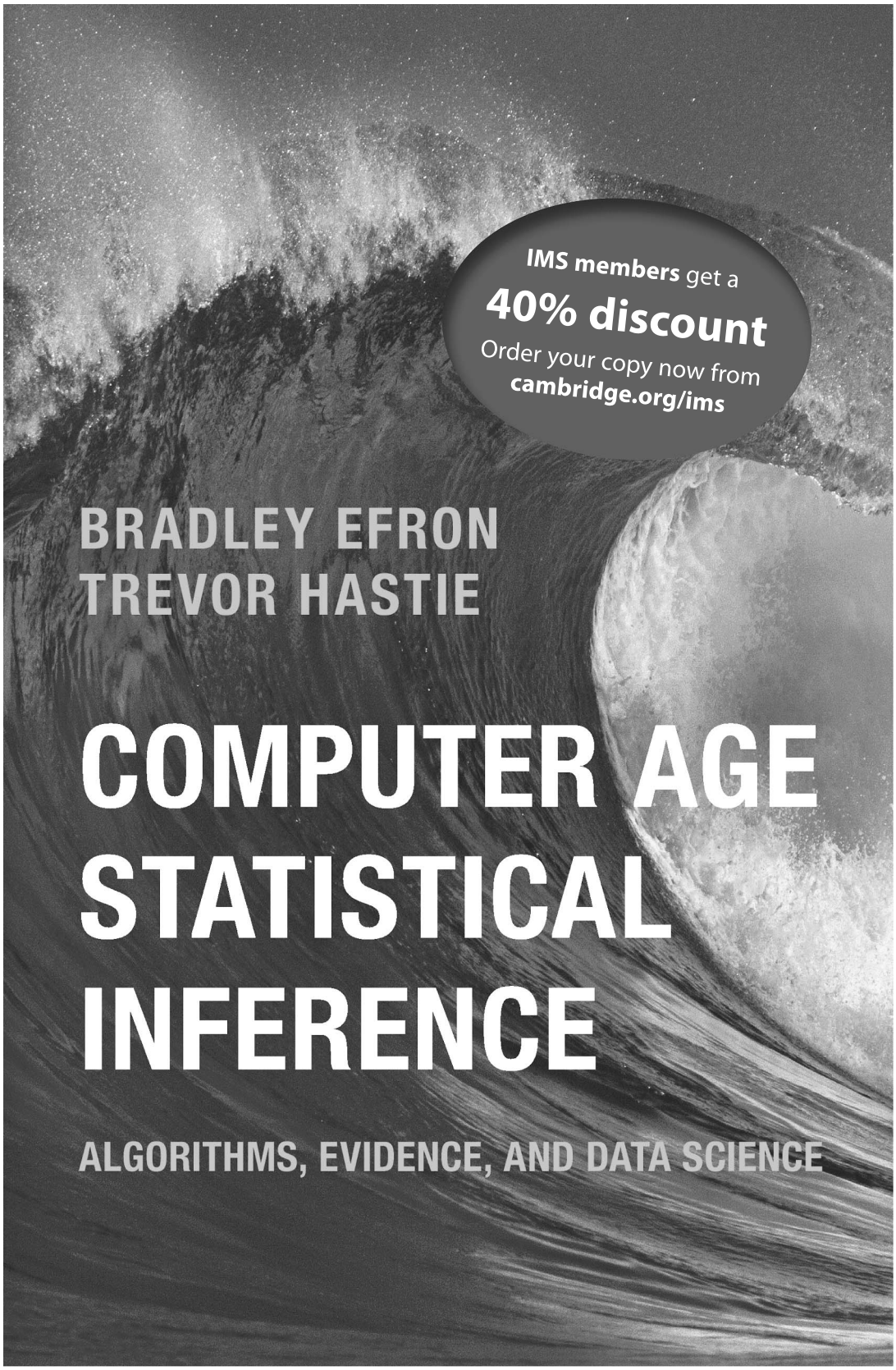
The Annals of Applied Statistics. *Editor-in-Chief*: Karen Kafadar, Department of Statistics, University of Virginia, Heidelberg Institute for Theoretical Studies, Charlottesville, VA 22904-4135, USA

The Annals of Probability. *Editor*: Amir Dembo, Department of Statistics and Department of Mathematics, Stanford University, Stanford, California 94305, USA

The Annals of Applied Probability. *Editors*: François Delarue, Laboratoire J. A. Dieudonné, Université de Nice Sophia-Antipolis, France-06108 Nice Cedex 2. Peter Friz, Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany and Weierstrass-Institut für Angewandte Analysis und Stochastik, 10117 Berlin, Germany

Statistical Science. *Editor*: Sonia Petrone, Department of Decision Sciences, Università Bocconi, 20100 Milano MI, Italy

The IMS Bulletin. *Editor*: Vlada Limic, UMR 7501 de l'Université de Strasbourg et du CNRS, 7 rue René Descartes, 67084 Strasbourg Cedex, France


IMS members get a
40% discount
Order your copy now from
cambridge.org/ims
BRADLEY EFRON
TREVOR HASTIE
COMPUTER AGE STATISTICAL INFERENCE
ALGORITHMS, EVIDENCE, AND DATA SCIENCE