NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: A Workshop

February 26-27, 2015

National Academy of Sciences Building
2101 Constitution Ave NW
Washington, DC
NAS Lecture Room

Webinar link: http://sites.nationalacademies.org/DEPS/BMSA/DEPS_153236

Meeting Objectives

Address statistical challenges in assessing and fostering the reproducibility of scientific results by examining three issues from a statistical perspective: the extent of reproducibility, the causes of reproducibility failures, and potential remedies.

Specifically:

- What are appropriate metrics and study designs that can be used to quantify reproducibility of scientific results?
    - Variability across studies is a well-known phenomenon and has given rise to the field of research synthesis and meta-analysis. How should this variability be assessed? What degree of variability would lead to concerns about lack of reproducibility?
- How can the choice of statistical methods for study design and analysis affect the reproducibility of a scientific result?
    - How does routine statistical hypothesis testing with widely used thresholds for test significance affect the reproducibility of results? How do standard methods for study design and choice of sample size affect reproducibility?
- Are there analytical and infrastructural approaches that can enhance reproducibility, within disciplines and overall?
    - Do we need new conceptual/theoretical frameworks for assessing the strength of evidence from a study? Do we need broad adoption of practices for making study protocols and study data available to the scientific community? How can this be achieved?

Thursday, February 26

Session I - Overview and Case Studies

8:30am Introductions from the Workshop Co-Chairs
- Constantine Gatsonis, Brown University
- Giovanni Parmigiani, Dana Farber Cancer Institute

8:45am Perspectives from Stakeholders
- Lawrence Tabak, National Institutes of Health
- Irene Qualters, National Science Foundation
- Justin Esarey, Rice University and *The Political Methodologist*
- Gianluca Setti, University of Ferrara, Italy and IEEE
- Joelle Lomax, Science Exchange

9:45am Overview of the Workshop
- Victoria Stodden, University of Illinois at Urbana-Champaign

10:15am Break

10:30am Case Studies
Speakers:
- Yoav Benjamini, Tel Aviv University
- Justin Wolfers, University of Michigan

12:10pm Lunch

Session II - Conceptualizing, Measuring, and Studying Reproducibility

1:30pm Definitions and Measures of Reproducibility
Speaker: Steve Goodman, Stanford
Discussant: Yoav Benjamini, Tel Aviv University

2:30pm Reproducibility and "Statistical Significance"
Speaker: Dennis Boos, North Carolina State University
Discussants:
- Andreas Buja, Wharton, University of Pennsylvania
- Val Johnson, Texas A&M

3:30pm Break

3:45pm Assessment of Factors Affecting Reproducibility
Speaker: Marc Suchard, University of California, Los Angeles
Discussants:
- Courtney Soderberg, Center for Open Science
- John Ioannidis, Stanford University

4:45pm Reproducibility from the Informatics Perspective
Speaker: Mark Liberman, University of Pennsylvania
Discussant: Micah Altman, Massachusetts Institute of Technology

5:45pm Adjourn

Friday, February 27

Session III - The Way Forward: Using Statistics to Achieve Reproducibility

8:30am Panel Discussion: Open Problems, Needs and Opportunities for Methodologic Research
Moderator: Giovanni Parmigiani, Dana Farber Cancer Institute

- Lida Anestidou, National Research Council
- Tim Errington, Center for Open Science
- Xiaoming Huo, National Science Foundation
- Roger Peng, Johns Hopkins Bloomberg School of Public Health

9:45am Break

10:00am Panel Discussion: Reporting Scientific Results and Sharing Scientific Study Data
Moderator: Victoria Stodden, University of Illinois at Urbana-Champaign

- Keith Baggerly, MD Anderson Cancer Center
- Ron Boisvert, Association for Computing Machinery and National Institute of Standards and Technology
- Randy LeVeque, Society for Industrial and Applied Mathematics and University of Washington
- Marcia McNutt, Science

11:45am Panel Discussion: The Way Forward from the Data Sciences Perspective: Research
Moderator: Constantine Gatsonis, Brown University

- Chaitan Baru, National Science Foundation
- Phil Bourne, National Institutes of Health
- Rafael Irizarry, Harvard University
- Jeff Leek, Johns Hopkins University

1:00pm Adjourn

Suggested Literature

Session I - Overview and Case Studies

Alogna, V., Attaya, M., Aucoin, P., Bahník, Š., Birch, S., Bornstein, B., . . . Buswell, K. (2014). Contribution to alonga et al (2014). registered replication report: Schooler & engstler-schooler (1990). *Perspectives on Psychological Science, 9*(5), 556-578.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533.

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature, 505*(7485), 612-613.

Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS One, 7*(1), e29081.

Esarey, J., Wu, A., Stevenson, R. T., & Wilson, R. K. (2014). Editorial statement. *The Political Methodologist, 22*, 2-3-23.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460.

Gerber, A. S., & Green, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review,94*(03), 653-663.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PloS One, 8*(8), e72467.

Hayes, D. N., Monti, S., Parmigiani, G., Gilks, C. B., Naoki, K., Bhattacharjee, A., . . . Meyerson, M. (2006). Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology, 24*(31), 5079-5090. doi:24/31/5079 [pii]

Hothorn, T., & Leisch, F. (2011). Case studies in reproducibility. *Briefings in Bioinformatics, 12*(3), 288-300. doi:10.1093/bib/bbq084

Imai, K. (2005). Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *American Political Science Review, 99*(02), 283-300.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? A direct replication of schnall, benton, and harvey (2008). *Social Psychology, 45*(3), 209.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., . . . Brumbaugh, C. C. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142.

Molina, H., Parmigiani, G., & Pandey, A. (2005). Assessing reproducibility of a protein dynamics study using in vivo labeling and liquid chromatography tandem mass spectrometry. *Analytical Chemistry, 77*(9), 2739-2744.

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? failure to replicate effects on social and food judgments. *PloS One, 7*(8), e42510.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery,10*(9), 712-712.

Rasko, J., & Power, C. (2015,). What pushes scientists to lie? the disturbing but familiar story of haruko obokata. *The Guardian*

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science, 9*(5), 552-555.

Stodden, V., Borwein, J., & Bailey, D. (2013). Setting the default to reproducible. *Computational Science Research.SIAM News, 46*, 4-6.

Waldron, L., Haibe-Kains, B., Culhane, A. C., Riester, M., Ding, J., Wang, X. V., . . . Parmigiani, G. (2014). Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *Journal of the National Cancer Institute, 106*(5), 10.1093/jnci/dju049. doi:10.1093/jnci/dju049

Session II - Conceptualizing, Measuring, and Studying Reproducibility

Altman, M., Gill, J., & McDonald, M. P. (2004). Sources of inaccuracy in statistical computation. *Numerical issues in statistical computing for the social scientist*. John Wiley & Sons.

Begley, C. G. (2013). Reproducibility: Six red flags for suspect work. *Nature, 497*(7450), 433-434.

Bernau, C., Riester, M., Boulesteix, A. L., Parmigiani, G., Huttenhower, C., Waldron, L., & Trippa, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics (Oxford, England), 30*(12), i105-12. doi:10.1093/bioinformatics/btu279

Berry, D. (2012). Multiplicities in cancer research: Ubiquitous and necessary evils. *Journal of the National Cancer Institute, 104*(15), 1124-1132. doi:10.1093/jnci/djs301

Clayton, J. A., & Collins, F. S. (2014). NIH to balance sex in cell and animal studies. *Nature, 509*(7500), 282-283.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science, 1*(3), 140216.

Cossins, D. (2014). Setting the record straight. *Scientist, 28*(10), 48-53.

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., & Stodden, V. (2009). Reproducible research in computational harmonic analysis. *Computing in Science & Engineering, 11*(1), 8-18.

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics (Oxford, England), 11*(3), 385-388. doi:10.1093/biostatistics/kxq028 [doi]

Donohue III, J. J., & Wolfers, J. (2006). *Uses and Abuses of Empirical Evidence in the Death Penalty Debate,*

Goodman, S. N., Altman, D. G., & George, S. L. (1998). Statistical reviewing policies of medical journals. *Journal of General Internal Medicine, 13*(11), 753-756.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Jager, L. R., & Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics (Oxford, England), 15*(1), 1-12. doi:10.1093/biostatistics/kxt007 [doi]

Li, Q., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics, 5*(3), 1752-1779.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Annals of Internal Medicine, 151*(4), W-65-W-94.

Peers, I. S., Ceuppens, P. R., & Harbron, C. (2012). In search of preclinical robustness. *Nature Reviews Drug Discovery, 11*(10), 733-734.

Rekdal, O. B. (2014). Academic urban legends. *Social Studies of Science, 44*(4), 638-654.

Schooler, J. W. (2014). Turning the lens of science on itself verbal overshadowing, replication, and metascience. *Perspectives on Psychological Science, 9*(5), 579-584.

Spence, D. (2014). Evidence based medicine is broken. *Bmj, 348*

Stodden, V. (2013). Resolving irreproducibility in empirical and computational research. *IMS Bulletin Online,*

Vandewalle, P., Kovacevic, J., & Vetterli, M. (2009). Reproducible research in signal processing. *Signal Processing Magazine, IEEE, 26*(3), 37-47.

Session III - The Way Forward: Using Statistics to Achieve Reproducibility

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. Circulation Research, 116(1), 116-126. doi:CIRCRESAHA.114.303819 [pii]

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. The Annals of Statistics, 41(2), 802-837.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . de Vet, H. C. (2003). Toward complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. Academic Radiology, 10(6), 664-669.

Couzin-Frankel, J. (2015). Trust me, I'm a medical reseracher. Science, 347(6221), 501-502-503.

Donoho, D. L., & Huo, X. (2004). Beamlab and reproducible research. International Journal of Wavelets, Multiresolution and Information Processing, 2(04), 391-414.

Goodman, S. N. (1992). A comment on replication, p-values and evidence. Statistics in Medicine, 11(7), 875-879.

Heller, R., & Yekutieli, D. (2014). Replicability analysis for genome-wide association studies. The Annals of Applied Statistics, 8(1), 481-498.

Heller, R., Bogomolov, M., & Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. Proceedings of the National Academy of Sciences of the United States of America, 111(46), 16262-16267. doi:10.1073/pnas.1314814111

Karr, A. F. (2014). Why data availability is such a hard problem. Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 30(2), 101-107.

Laine, C., Goodman, S. N., Griswold, M. E., & Sox, H. C. (2007). Reproducible research: Moving toward research the public can really trust. Annals of Internal Medicine,146(6), 450-453.

Leek, J. T., & Peng, R. D. (2015). Opinion: Reproducible research can still be wrong: Adopting a prevention approach. Proceedings of the National Academy of Sciences of the United States of America, 112(6), 1645-1646. doi:10.1073/pnas.1421412111

LeVeque, R. J., Mitchell, I. M., & Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. Computing in Science and Engineering, 14(4), 13.

Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. British Journal of Pharmacology

Political Science Journal Editors. (2014). Data access and research transparency (DA-RT): A joint statement.

Reiter, J. P., & Kinney, S. K. (2011). Sharing confidential data for research purposes: A primer. Epidemiology (Cambridge, Mass.), 22(5), 632-635. doi:10.1097/EDE.0b013e318225c44b

Stodden, V. (2009). Enabling reproducible research: Open licensing for scientific innovation. International Journal of Communications Law and Policy, Forthcoming,

Stodden, V. (2009). The legal framework for reproducible scientific research: Licensing and copyright. Computing in Science & Engineering, 11(1), 35-40.

Stodden, V. (2013). Resolving irreproducibility in empirical and computational research. IMS Bulletin Online

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. PloS One, 8(6), e67111.

Stodden, V., Leisch, F., & Peng, R. D. (2014). Implementing reproducible research CRC Press.

Stodden, V., Miguez, S., & Seiler, J. (2015). ResearchCompendia. org: Cyberinfrastructure for reproducibility and collaboration in computational science. Computing in Science & Engineering, 17(1), 12-19.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PloS One, 6(11), e26828.

Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. Journal of the American Statistical Association, 103(481), 309-316.

Planning Committee

Co-Chairs

Constantine Gatsonis, Brown University
Giovanni Parmigiani, Dana Farber Cancer Institute

Members

Stephen Fienberg (NAS), Carnegie Mellon University
Steven N. Goodman, Stanford University School of Medicine
John H. Holmes, University of Pennsylvania Perelman School of Medicine
Alan F. Karr, RTI International
Jelena Kovačević, Carnegie Mellon University
Xihong Lin, Harvard University
Roger Peng, Johns Hopkins Bloomberg School of Public Health
Victoria Stodden, University of Illinois at Urbana-Champaign

*Staff officer***:**
Michelle Schwalbe
mschwalbe@nas.edu
202.334.1682