Observational Studies 1 (2015) 294-306 Submitted 6/15; Published 11/15

Targeted Learning for Pre-Analysis Plans in Public Health and Health Policy Research

Sherri Rose **rose@hcp.med.harvard.edu**
Department of Health Care Policy
Harvard Medical School
Boston, MA 02115, U.S.A

Abstract

The publication of pre-analysis plans is gaining further mainstream attention as the open science movement within academic research expands. Additionally, some funding agencies require pre-analysis plans as a condition of their grant awards. This paper proposes the existing targeted learning roadmap as a framework for the construction of pre-analysis plans. It also discusses current standards for content and hurdles that may arise in public health and health policy settings, including the complexities of "Big Data," when writing pre-analysis plans for observational studies in practice.

Keywords: Pre-analysis plans, Targeted learning, Observational studies, Public health, Health policy

1. Introduction

There is a growing interest in adopting a standard for pre-analysis plans in public health and health policy studies that make use of observational data. An infrastructure and understood need for registering randomized studies in clinical research arose partly out of a demand to protect patients (Zarin and Keselman, 2007; Moher et al., 2010). Additionally, while not yet pervading the field, new tools, guidelines and resources have emerged for randomized and observational studies in the social sciences (Boutron et al., 2010; Humphreys et al., 2013; Miguel et al., 2014; Experiments in Governance and Politics, 2015; International Initiative for Impact Evaluation, 2015; American Economic Association, 2015). In contrast, a uniform recommendation has not been established for observational health studies; an area where, due to the lack of randomization, analysis choices can greatly impact the final results of studies that may have lasting policy implications.

The potential benefits of pre-analysis plans are generally widely acknowledged. This includes transparency of the research process, as scientists must document *a priori* their intended outcomes, measured covariates, parameters of interest, and estimation techniques. One goal is to reduce concerns about possible "data snooping" for research questions and statistically significant findings. Additional post hoc analyses undertaken at a later stage in the study are then clearly labelled as such in the results. Monogan (2013) and Gelman (2013) note that flaws in the pre-analysis plan may be identified by outside researchers before a study is conducted. Another key potential benefit of pre-analysis plans is the archiving of studies that ultimately have null results; studies that, without pre-analysis plans, often go unpublished. In the social sciences, a recent study of this phenomena found that 65% of

survey-based experiments with null results were not even drafted into manuscripts, while only 4% of studies with "strong" results were not drafted into a manuscript (Franco et al., 2014). Unreported findings distort the evidence on a topic in the published literature and can also lead to repeated replication in unfruitful areas.

However, it has been argued that incentivizing replication itself may be better for science than the introduction of widespread pre-analysis plans (Coffman and Niederle, 2014). Coffman and Niederle also asserted that the advantages of pre-analysis plans are limited as they have a nontrivial time cost to researchers and do not reduce the number of false positives published. Other arguments against mandatory registration of studies with pre-analysis plans include giving undue weight to less important unpublished findings and creating a culture where scientists are not trusted (Mervis, 2014). Gelman (2013) also discusses the concern that pre-analysis plans could incentivize "robotic" data analyses that do not evolve with the progression of the study. Instead, Gelman supports pre-analysis simulation studies that require detailed decision-making, as in Casey et al. (2011), for major studies.

This paper proposes the existing targeted learning roadmap (van der Laan and Rose, 2011) as a possible framework for the construction of pre-analysis plans. It also includes discussions on the current lack of a global standard for pre-analysis plan content in public health and health policy observational studies, and the barriers that emerge when attempting to write a robust pre-analysis plan for these types of studies. Of note, the much hyped era of "Big Data" introduces further complexities in the creation of a comprehensive pre-analysis plan. These issues have become increasingly relevant as funding agencies begin to adopt mandatory open science policies with pre-analysis plans. A more thorough understanding of the intricacies involved in drafting pre-analysis plans "in the trenches" may help lead to more uniform approaches and anticipation of previously unforeseen practical considerations.

2. Standards

Even in non-clinical fields where pre-analysis plans have begun gaining traction, such as randomized studies in political science, there remains a lack of consensus on the standards for content. For example, the Experiments in Governance and Politics Network (EGAP), founded in 2009, facilitates study registration for experimental research, among other initiatives. Their study design registration form includes a series of questions about the investigators and the study, as well as three main content areas requiring longer responses:

1. Goals of the project;
2. List of hypotheses that will be tested; and
3. Description of how hypotheses will be tested.

Researchers also have an option to include a detailed analysis plan as a separate file (Experiments in Governance and Politics, 2015). It has been noted that the amount of detail supplied in studies registered at EGAP varies, despite using the same form. Dr. Macartan Humphreys, Executive Director at EGAP and Professor in the Department of Political Science at Columbia University, was quoted regarding an EGAP pre-analysis plan filed by Michael LaCour, saying it was "missing a lot of detail one might like to see in a registered

design" (Singal, 2015). It has also been asserted that LaCour may have fabricated a second pre-analysis plan after publication of his studies; one he claimed was filed a priori (Singal, 2015). While perhaps an extreme example, considering the later retraction of the paper associated with the LaCour studies due to scientific misconduct (McNutt, 2015; Broockman et al., 2015), it remains a stark illustration of how pre-analysis plans cannot prevent possible scientific fraud, but rather provide transparency in the spirit of open science.

The Laura and John Arnold Foundation (LJAF) funds research in multiple areas, including evidence-based policy and innovation. Research sponsored by LJAF must be pre-registered using the Center for Open Science's Open Science Framework, and LJAF provides guidelines for the content of pre-analysis plans (Laura and John Arnold Foundation, 2013). The guidelines are described as flexible regarding level of detail and include sections specific to randomized experiments, observational studies, and machine learning. For observational studies, LJAF directs researchers to include:

> 1. Scientific background and explanation of rationale;
> 2. Specific objectives or hypotheses;
> 3. Complete description of dataset to be used, including where it is available, what variables it contains, what years it covers, and settings and locations where the data were collected;
> 4. Completely defined pre-specified primary and secondary outcome measures;
> 5. A description of statistical methods to be used (e.g., instrumental variables, regression discontinuity, fixed effects, random effects or more general hierarchical models, propensity score matching or other matching variants, smoothing splines, etc.), including complete specifications such as covariates, transformations, functional form, etc.;
> 6. A description of methods for additional analyses, such as subgroup analyses and adjusted analyses; and
> 7. A description of whether the researcher will use data splitting measures.
>
> —Quoted text from Laura and John Arnold Foundation (2013), page 3.

The interface in the Open Science Framework also allows for what one might term "iterative" pre-analysis plans, where revised plans are uploaded in subsequent files when crucial details regarding the analysis of the study change. Not all of LJAF's funded grants are public health or health policy initiatives, they are currently funding a number of such programs.

While there are other guidelines for pre-analysis plans that have been presented for various scientific arenas, there are also roadmaps from the statistics literature that could be translated into a framework for pre-analysis plans in observational health research. Going back to the 1970s, George E.P. Box discusses iteration of the scientific method, and the

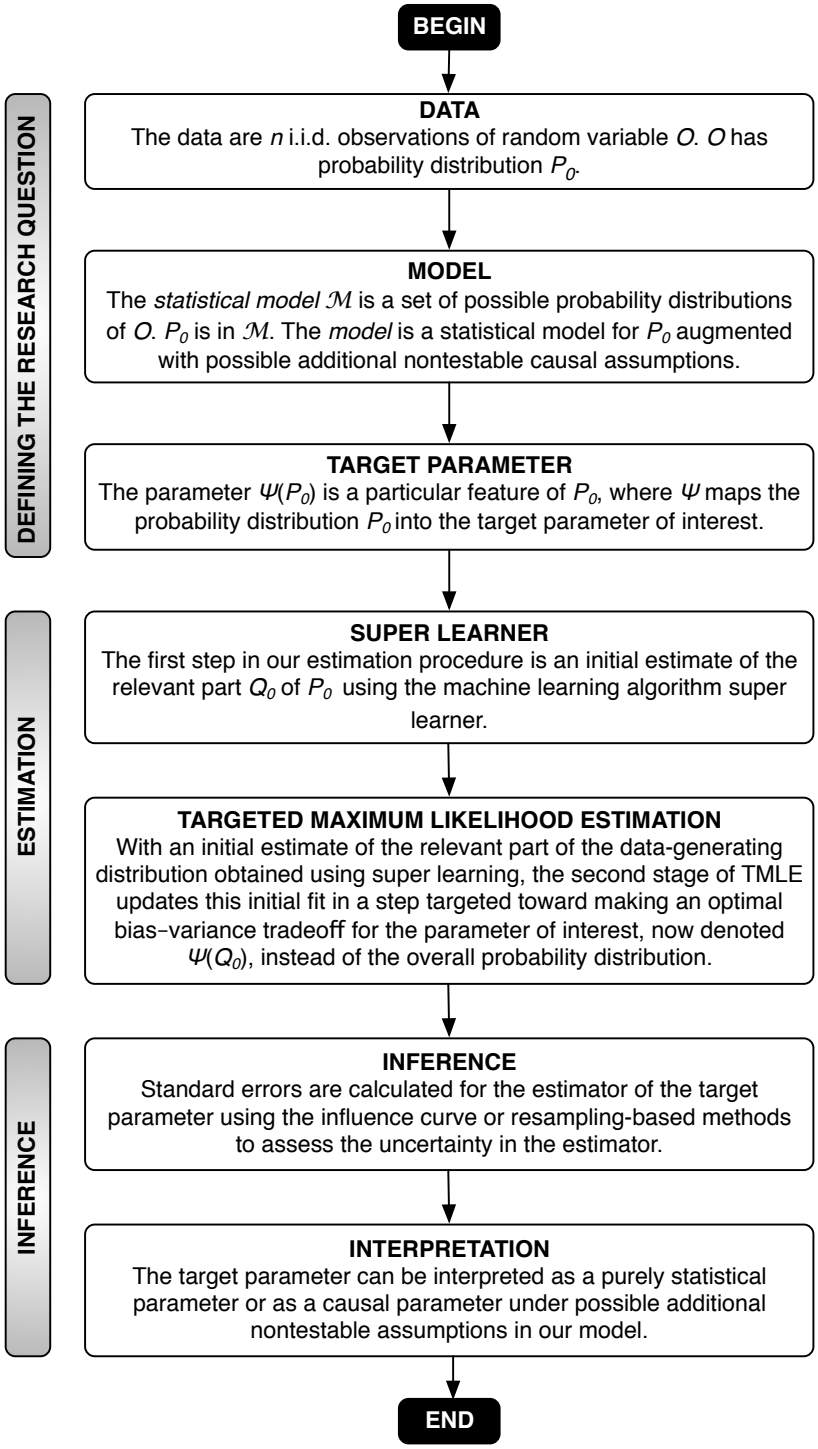

Figure 1: Roadmap for Targeted Learning. *Reprinted from van der Laan and Rose (2011) with permission from Springer via license #3647521094752.*

relationship of theory and practice for science and statistics (Box, 1976). A more recent and explicit example is the roadmap for targeted learning, displayed in Figure 1 (van der Laan and Rose, 2011), that we propose adopting as a possible option in this paper. The fundamentals in this statistical roadmap are also portable for use with other statistical methodology. The first stage is to carefully define the research question, including descriptions of the data, model, all statistical and causal assumptions, and the target parameter of interest. The middle stage is the estimation framework, discussed further in Section 3.2. The last stage is for inference, which includes confidence intervals and correct interpretation of the parameter of interest, i.e., as a purely statistical entity or as a causal parameter (see Section 3.3). A key tenet of the roadmap for targeted learning is that this roadmap is defined *a priori*, which is the exact principle of the pre-analysis plan. However, the "iterative" pre-analysis plan is not excluded here, as there would be continued transparency about when and where changes were made. Similarly, the detailed simulation approach supported by Gelman (2013) could also form the basis of the original pre-analysis plan. There are additional roadmaps for targeted learning when prediction is the research goal, or there are a large number of effect parameters of interest (van der Laan and Rose, 2011, Chapters 3-4). The necessity for differing methodology depending on the research question and data supports the consensus in much of the pre-analysis plan literature that a single mandatory structure for pre-analysis plans is untenable.

3. Road Map for Targeted Learning

We focus on the roadmap for targeted learning in (causal) effect estimation, which features estimation that incorporates machine learning and inference. The desirable statistical properties of such estimators have been discussed in detail elsewhere in the literature (e.g., van der Laan and Rubin, 2006; van der Laan et al., 2007; van der Laan and Rose, 2011), and include being well-defined, double robust, efficient substitution estimators. The roadmap, presented in Figure 1, has three major steps: defining the research question, estimation, and inference. We expand on and explain these steps in this section, describing adaptations for pre-analysis plans. There are also many practical considerations that arise in each step of a thorough pre-analysis plan, regardless of the exact guidelines one is following. These considerations are included in the relevant step.

3.1 Defining the Research Question

The guidelines discussed in this paper all involve the inclusion of background material. However, the role of this expository material is often ill defined beyond framing the scientific objectives. For the targeted learning roadmap, this background material is useful for translating the scientific question of interest into a statistical question. The extension of the roadmap for targeted learning proposed here recommends sufficiently detailed background material such that the observational unit, model, and target parameter can be completely described as discussed below.

3.1.1 (Big) Data

The observational unit O has true underlying probability distribution P_0. Our random variable O is observed n times. In a straightforward cross-sectional observational study without missingness, we might define our observational unit $O = (W, A, Y)$, where W is a vector of baseline covariates, A the intervention of interest, and Y the outcome. Many other definitions are possible depending on the data structure and scientific question. However, merely defining O in this fashion is not sufficient for our pre-analysis plan. There is increasing excitement over "Big Data" in many scientific areas within health research (Rudin et al., 2015). It is imperative that we acknowledge that not all large data sets are equally valuable for research, as many forms of "Big Data" are not collected for this purpose and quality may be poor (Rudin and Bates, 2014; Haneuse, 2015; Joyner and Paneth, 2015). Consider electronic medical records, obtained for billing claims, being used to predict health outcomes. Crucial variables may not be available given the nature of claims data. Additionally, with hundreds of potential covariates, descriptions of all data cleaning and screening steps requires additional careful consideration. Many of these same issues arise in smaller datasets as well. These important details are not included by simply defining W as a vector of covariates and should be included in this extension of the targeted learning roadmap for pre-analysis plans, or a revised pre-analysis plan when the full details become known to the research team.

Novel Data Sources. Another challenging aspect of pre-analysis plans in public health and health policy research is a detailed description of new data. This is especially true when the data have yet to be collected, and use of detailed simulation studies, as supported by Gelman (2013), may be suitable to continue with analytic decision-making. For example, simulation studies could be used to develop a priori rules for the number of dynamic regimes to study based on enrollment and projected power. Collaborations involving sensitive government data sources can also necessitate employing intermediaries to actually handle the data. Thus, without direct access to the data, it is difficult to truly understand the data, as it will require communication through multiple channels. A similar, but less anticipated situation when a formal description of the data can be difficult is when purchasing data from new partners. While these partner companies may have the best intentions in describing their data sources, it is not uncommon to later discover multiple issues. For example, a particular data field may be described as "existing," but when the database arrives, that variable is 80% missing. In severe cases, crucial variables may have such extreme measurement error that the entire research project must be abandoned. The quality of new data sources may be unknown and difficult to foresee at the stage one would write a pre-analysis plan. However, a pre-analysis plan followed by a summary detailing the reasons for aborting a project would be useful for informing the research community about the limitations of new data sources, and thus is recommended.

Inclusion and Exclusion Criteria. A final challenge presented by health databases considered in this extension of the targeted learning roadmap for pre-analysis plans is a transparent description of all inclusion and exclusion criteria. Often, there are multiple layers of criteria imposed on a set of data. For example, a partner company may make opaque restrictions on which subjects are released, and then the research team decides that they will further exclude additional subjects in order to isolate the effect of interest.

These inclusion and exclusion criteria can have a substantial impact on the definition of the parameter of interest and the effect estimates obtained (Hernán et al., 2008). Frequently, this level of detail is excluded from published papers and it is important to include this material in a pre-analysis plan, or revised plan, as it is part of comprehensively defining your population of interest.

3.1.2 Model & Parameter

The statistical model $\mathcal{M}$ is, formally, the set of possible probability distributions of O. Researchers will wish to consider a sufficiently large statistical model such that it contains the true underlying probability distribution P_0. Thus, the targeted learning roadmap recommends nonparametric or semiparametric statistical models representing only realistic assumptions about the data. In such a nonparametric statistical model, one might make only the assumption that there are n i.i.d. copies of O. While much of the analysis of observational data is performed within parametric statistical models, the unavoidable reality is that in "Big Data" our underlying knowledge about the system that generated the data will generally not support the strong assumptions required for parametric models. The parameter of interest $\Psi(P_0)$ is then defined as a feature of the probability distribution P_0 in a nonparametric or semiparametric model rather than a coefficient in a parametric model.

Additional untestable causal assumptions can augment the statistical model $\mathcal{M}$. Although, it is important to stress, that these additional causal assumptions do not change the statistical model, but rather allow an enriched causal interpretation of the parameter of interest. Whether it is possible to estimate causal effects in observational data is routinely discussed, as it is a topic of great importance, considering the significance of the issues studied in public health and health policy research (Taubes, 2007). Including an explicit causal framework in the pre-analysis plan has the benefit of highlighting what needs to be the case for the parameters of interest to have causal interpretations (Neyman, 1923; Holland, 1986; Rubin, 1990; Pearl, 2009).

3.2 Estimation

One major criticism of the analysis of observational data is that researchers may run multiple (parametric) regressions, select the regression that supports their hypothesis, and report a single regression as if it had been *a priori* specified. This is, of course, problematic, and explicit pre-analysis plans identifying the primary parametric regression, when parametric regression is the estimation method chosen, is an important first step. (Additional parametric regressions with varying functional forms would them be listed as supplementary sensitivity analyses.) Estimation in health research often involves the use of these common parametric regression methods, but also includes developing new methods or integrating approaches from related quantitative disciplines. Estimators for the larger model spaces described in Section 3.1.2 can be implemented, including targeted learning, as recommended in the roadmap for targeted learning. However, as noted earlier, the general components of the roadmap for targeted learning is portable to other methods, and this estimation step would then contain alternative techniques. Targeted learning is compared to other estimators based on their statistical properties in Chapter 6 of van der Laan and Rose (2011).

3.2.1 Super Learner

We first need an initial estimate of P_0, or the relevant component Q_0 of P_0 to estimate $\Psi(P_0)$. Super learner provides a template for using multiple algorithms (e.g., regression, tree-based methods, neural nets) to generate an estimator of the relevant part Q_0 that is the optimal weighted average of all considered algorithms. Thus, here, researchers have the flexibility to *run multiple regressions*, while protecting the research from arbitrary ad hoc choices about algorithm selection. The criterion for selection of the optimal weighted average is established *a priori*, as is the library of algorithms, and other choices, such as tuning parameters, variables, and screening methods (van der Laan et al., 2007). One common criterion is the squared error loss function.

3.2.2 Targeted Maximum Likelihood Estimation

For effect estimation, targeted maximum likelihood estimation (van der Laan and Rubin, 2006) integrates super-learning-based estimates of the outcome regression and exposure mechanism (as well as censoring mechanism and missingness mechanism if relevant). Formally, the targeted maximum likelihood estimator is a two-stage estimator that requires an initial estimate of the data-generating distribution P_0, or Q_0, as discussed above. The second step perturbs this initial estimate, targeting the estimator to make an optimal bias–variance tradeoff for the parameter $\Psi(P_0)$, instead of the overall density P_0. This is done through the definition of a one-dimensional working model to fluctuate the initial estimator. The implementation of targeted maximum likelihood estimation is described in detail for a risk difference parameter elsewhere (van der Laan and Rose, 2011, Chapter 4), as well as other settings and parameters (e.g., Rubin and van der Laan, 2008; Moore and van der Laan, 2009a,b,c; van der Laan, 2010; van der Laan and Rose, 2011; van der Laan, 2014).

3.2.3 Remarks on Implementation

Despite a desire to use large model spaces and flexible modern techniques in the statistical analysis of an observational health data set, the logistics of the research program again must be considered. As previously discussed, some collaborative initiatives will not involve the investigator's in-house team obtaining the actual data, with analyses performed by partners several steps removed from the investigator's authority and oversight. Additionally, the programming language may be restricted to SAS or STATA, rather than R, where many of the newest statistical advances have been implemented (e.g., Polley and van der Laan, 2013; Gruber and van der Laan, 2012). Thus, when writing a pre-analysis plan, it is essential to factor in the programming environment and capabilities of the team who will perform the work. This may involve forgoing a multi-step machine learning technique. For example, implementing g-computation in a parametric model (Robins, 1986) instead of targeted maximum likelihood estimation in a nonparametric model, while including transparent detail on required assumptions. Careful consideration, foresight, and detailed discussions with research partners will ideally take place early in the process in order to produce an accurate pre-analysis plan that is feasible to implement. That said, the estimation procedure can and should be updated upon discovery of such complications later in the research process.

3.3 Inference

Standard errors for targeted learning estimators are calculated using influence curves or bootstrapping (van der Laan and Rose, 2011). These standard errors can then be used to compute standard confidence intervals and p-values. As discussed in Section 3.1.2, interpreting the parameter of interest as a causal estimand requires a set of untestable causal assumptions. In observational data the assumption of no unmeasured confounding is, rightly so, a frequent source of concern. Suspected violations of this assumption may lead researchers to conclude that their parameter does not have a causal interpretation. Nevertheless, the exercise of walking through a formal treatment of causality in the pre-analysis plan adds another layer of transparency to the work, both for the investigator and consumers of the research. Without causal assumptions, the target parameter has a purely statistical interpretation, for example, as a W-adjusted effect parameter when $O = (W, A, Y)$ and $\Psi(P_0) = E_W[E[Y \mid A = 1, W] - E[Y \mid A = 0, W]]$.

4. Discussion

The goal of this paper was to introduce the targeted learning roadmap, with some extensions, as a possible framework for pre-analysis plans. Additionally, it presents related guidelines for pre-analysis plans and illuminates some unexpected hurdles that may arise for public health and health care policy researchers when drafting pre-analysis plans. Pre-analysis plans, whether filed publicly or used internally, can be beneficial, as they demand consideration of critical issues *before* researchers are influenced by the data. This does not preclude "iterative" pre-analysis plans or the use of detailed simulation studies to create the analysis plan. These are both supported by this extension of the targeted learning road map for pre-analysis plans. For large research teams, pre-analysis plans also offer a roadmap for group members to follow. Pre-analysis plans should therefore explicitly define primary and secondary research questions, data descriptions with inclusion and exclusion criteria, and detailed statistical methods for estimation, including contingency plans for common issues found in observational data, such as missing data.

More widespread use of pre-analysis plans may have the unintended but indeed useful side effect of integrating statisticians into collaborative teams at an earlier stage. As "Big Data" begins to proliferate health fields, and data science becomes the larger umbrella that encapsulates statistical considerations, the role of statisticians is even more critical. It is certainly a waste of research dollars and energy when data is collected in a study design that ultimately cannot answer the research question of interest; discovered only at the analysis stage when a statistician was brought on board.

This is not to say that constructing an exhaustive pre-analysis plan for public health and health policy research is necessarily easy. Particularly when the data are new and unknown and the research question is novel or pushing boundaries, pre-analysis plans are less straightforward. However, in these cases, the utility of the research plan may even be greater than less complicated projects, as nuances are discovered and given consideration formally before mistakes are made.

Disclosures

The author is a faculty member in the Healthcare Markets and Regulation Lab at Harvard Medical School, which has research projects funded by the Laura and John Arnold Foundation. The Laura and John Arnold Foundation requires that investigators register pre-analysis plans, and the author is currently partially funded by this organization. The Laura and John Arnold Foundation did not fund this paper nor are they responsible for any of its content.

Acknowledgments

The author thanks Michael Chernew, Alexandra Maher, and Isabel Ostrer for helpful discussion.

References

American Economic Association (2015). AEA RCT registry. socialscienceregistry.org. Accessed: 06/13/15.

Boutron, I., John, P., and Torgerson, D. (2010). Reporting methodological items in randomized experiments in political science. *The Annals of the American Academy of Political and Social Science*, 628(1):112–131.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Broockman, D., Kalla, J., and Aronow, P. (2015). Irregularities in LaCour (2014). Technical report, Stanford Working Paper.

Casey, K., Glennerster, R., and Miguel, E. (2011). Reshaping institutions: Evidence on aid impacts using a pre-analysis plan. Technical report, National Bureau of Economic Research.

Coffman, L. and Niederle, M. (2014). Pre-analysis plans are not the solution, replications might be. Technical report, Working Paper.

Experiments in Governance, E. and Politics (2015). Design registration form. egap.org/design-registration/standards-project-registration. Accessed: 06/13/15.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505.

Gelman, A. (2013). Preregistration of studies and mock reports. *Political Analysis*, 21(1):40–41.

Gruber, S. and van der Laan, M. (2012). tmle: An r package for targeted maximum likelihood estimation. *J Stat Softw*, 51(13).

Haneuse, S. (2015). On the use of electronic health records for CER. In Gatsonis, C. and Morton, S., editors, *Methods in Comparative Effectiveness Research*. Chapman and Hall/CRC, Boca Raton.

Hernán, M., Alonso, A., Logan, R., Grodstein, F., Michels, K., Stampfer, M., Willett, W., Manson, J., and Robins, J. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, 19(6):766.

Holland, P. (1986). Statistics and causal inference. *J Am Stat Assoc*, 81(396):945–960.

Humphreys, M., Sanchez de la Sierra, R., and van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, 21(1):1–20.

International Initiative for Impact Evaluation (2015). Registry for international development impact evaluations. 3ieimpact.org/en/evaluation/ridie/. Accessed: 06/13/15.

Joyner, M. and Paneth, N. (2015). Seven questions for personalized medicine. *Journal of the American Medical Association*, Online First:doi:10.1001/jama.2015.7725.

Laura and John Arnold Foundation (2013). Guidelines for investments in research. Technical report, Laura and John Arnold Foundation Research Integrity.

McNutt, M. (2015). Editorial retraction. *Science*, page aaa6638.

Mervis, J. (2014). Why null results rarely see the light of day. *Science*, 345(6200):992–992.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K., Gerber, A., Glennerster, R., Green, D., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B., Petersen, M., Sedlmayr, R., Simmons, J., Simonsohn, U., and van der Laan, M. (2014). Promoting transparency in social science research. *Science*, 343(6166):30–31.

Moher, D., Hopewell, S., Schulz, K., Montori, V., Gøtzsche, P., Devereaux, P., Elbourne, D., Egger, M., and Altman, D. (2010). Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology*, 63(8):e1–e37.

Monogan, J. E. (2013). A case for registering studies of political outcomes: An application in the 2010 house elections. *Political Analysis*, 21(1):21–37.

Moore, K. and van der Laan, M. (2009a). Application of time-to-event methods in the assessment of safety in clinical trials. In Peace, K. E., editor, *Design, Summarization, Analysis & Interpretation of Clinical Trials with Time-to-Event Endpoints*, Boca Raton. Chapman & Hall.

Moore, K. and van der Laan, M. (2009b). Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Stat Med*, 28(1):39–64.

Moore, K. and van der Laan, M. (2009c). Increasing power in randomized trials with right censored outcomes through covariate adjustment. *J Biopharm Stat*, 19(6):1099–1131.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). English translation by D.M. Dabrowska and T.P. Speed (1990). *Stat Sci*, 5:465–480.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge, New York, 2nd edition.

Polley, E. and van der Laan, M. (2013). *SuperLearner: Super Learner Prediction*. R package version 2.0-10. Available from: `http://CRAN.R-project.org/package=SuperLearner`.

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods–application to control of the healthy worker survivor effect. *Math Mod*, 7:1393–1512.

Rubin, D. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Stat Sci*, 5(4):472–480.

Rubin, D. B. and van der Laan, M. (2008). Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Int J Biostat*, 4(1):Article 5.

Rudin, C., Dunson, D., Irizarry, R., Ji, H., Laber, E., Leek, J., McCormick, T., Rose, S., Schafer, C., van der Laan, M., Wasserman, L., and Xue, L. (2015). Discovery with data: Leveraging statistics with computer science to transform science and society. Technical report, Working Paper of the American Statistical Association.

Rudin, R. and Bates, D. (2014). Let the left hand know what the right is doing. *Journal of the American Medical Informatics Association*, 21(1):13–16.

Singal, J. (2015). Michael LaCour probably fabricated a document about research integrity. `nymag.com/scienceofus/2015/06/lacour-probably-fabricated-an-integrity-document.html`. June 1.

Taubes, G. (2007). Do we really know what makes us healthy? *The New York Times*.

van der Laan, M. (2010). Estimation of causal effects of community-based interventions. Technical Report 268, Division of Biostatistics, University of California, Berkeley.

van der Laan, M. (2014). Causal inference for networks. *J Causal Inference*, 2(1):13–74.

van der Laan, M., Polley, E., and Hubbard, A. (2007). Super learner. *Stat Appl Genet Mol*, 6(1):Article 25.

van der Laan, M. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, Berlin Heidelberg New York.

van der Laan, M. and Rubin, D. (2006). Targeted maximum likelihood learning. *Int J Biostat*, 2(1):Article 11.

Zarin, D. and Keselman, A. (2007). Registering a clinical trial in ClinicalTrials. gov. *CHEST Journal*, 131(3):909–912.