American Journal of Epidemiology
Vol. 186, No. 4
DOI: 10.1093/aje/kwx066
Advance Access publication:
August 2, 2017

Practice of Epidemiology

A Pragmatic Approach for Reproducible Research With Sensitive Data

Bryan E. Shepherd*, Meridith Blevins Peratikos, Peter F. Rebeiro, Stephany N. Duda, and Catherine C. McGowan

* Correspondence to Bryan E. Shepherd, Department of Biostatistics, Vanderbilt University School of Medicine, 2525 West End, Suite 11000, Nashville, TN 37203 (e-mail: bryan.shepherd@vanderbilt.edu).

Initially submitted July 30, 2016; accepted for publication February 24, 2017.

Reproducible research is important for assessing the integrity of findings and disseminating methods, but it requires making original study data sets publicly available. This requirement is difficult to meet in settings with sensitive data, which can mean that resulting studies are not reproducible. For studies in which data cannot be shared, we propose a pragmatic approach to make research quasi-reproducible. On a publicly available website without restriction, researchers should post 1) analysis code used in the published study, 2) simulated data, and 3) results obtained by applying the analysis code used in the published study to the simulated data. Although it is not a perfect solution, such an approach makes analyses transparent for critical evaluation and dissemination and is therefore a significant improvement over current practice.

de-identification; HIV; observational data; reproducible research

Abbreviations: AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus.

***Editor's note:** A commentary on this article appears on page 393.*

Research is reproducible if, given a study's original data sets and analysis code, an independent scientist can obtain quantitative results identical to those obtained by the original researchers (1, 2). The journal *Biostatistics* defines papers as reproducible if 1) the data are made available on the journal's website, 2) any computer code or software used to compute the published results is provided, and 3) the journal's associate editor is able to execute the code on the available data and obtain results that match those in the paper (3). By these criteria, reproducible research clearly differs from replicable research: Research is replicable if independent scientists can obtain similar study results using the same methods with different data sets. Replicating research findings is a critical step in assessing study validity, ensuring generalizability, and translating research into practice, but the conduct of confirmatory studies involves extensive effort and cost. Reproducing research findings provides different benefits, but it can be done with comparatively minimal cost and effort.

Reproducible research bolsters confidence in analytic findings, ensures that results are correct, engenders trust in the honesty and competence of researchers, and serves as a protection against fraud (4–6). Public sharing of well-documented analysis code and reusable data sets allows researchers to elucidate complicated analyses and interested readers to critically examine what analysts have done. Because of publication page limits, the methods sections of articles rarely include the level of detail on data preparation and statistical approaches needed to replicate an analysis. Further text descriptions of procedures, such as those provided in supplemental materials, may still omit key steps. Data and code sharing not only make research reproducible but also permit efficient dissemination of statistical methods for reuse, because other groups may want to directly apply analysis code from a published paper to different studies.

Journals and funding agencies have begun to recognize the importance of reproducible research, particularly in sharing data. For example, the *PLOS* journals have a data policy in which they require authors to make all data underlying the findings described in their manuscript fully available without restriction. In rapid succession, the Institute of Medicine and International Committee of Medical Journal Editors released guidance on sharing clinical trial data (7, 8). The International Committee of Medical Journal Editors declared that publication of clinical trial results in affiliated journals will soon be contingent on the provision of related de-identified individual data within 6 months of publication. The recent attention has led to a wave of commentaries highlighting benefits and challenges of and strategies for data sharing (9–14). The United States National Institutes of Health has a data-sharing policy for studies that it funds (15); however, the specific requirements of the institutes vary according to the types of data involved and the funding institute (16–18). These mandates have not yet been extended to observational data by the International Committee of Medical Journal Editors, extended across all United States National Institutes of Health institutes, or required by other government regulatory agencies; however, this might simply be a matter of time.

Despite the importance of making data publicly available, there are some types of data that cannot be made fully available without restriction, even in de-identified forms (11, 19). Legal and ethical restrictions prevent the public disclosure of cohort data on stigmatizing conditions such as human immunodeficiency virus (HIV) infection. Many research data sets stem from the secondary use of clinical data, with patient informed consent waived by institutional review boards because of understandings that such data will not be made publicly available. Sensitive data can be de-identified; there are many ways to de-identify or anonymize data, each with differing probabilities of re-identification (20). The United States Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule provides a precise description for de-identifying data that appears to yield a very low probability of re-identification (21). However, researchers have demonstrated that partial re-identification of de-identified data sets is possible using publicly available information (21). In addition, many collaborative cohorts have by-laws that specifically prohibit publicly posting data, even in de-identified forms.

There are many examples of investigators effectively sharing sensitive data within research communities. For example, there are multicohort HIV/acquired immunodeficiency syndrome (AIDS) collaborations in which sensitive, de-identified data sets are sent to other researchers, with permission, for joint analyses (e.g., the International Epidemiology Databases to Evaluate AIDS (22–24)). The process typically requires a formal request for data by an independent researcher who submits a proposed study concept sheet, approval by the directors of the cohort and relevant ethics committees, and assurances that the data will only be used for the proposed study and will not be shared outside of the collaboration. Several forward-thinking groups have posted data to the Internet at secure locations (e.g., the Data and Specimen Hub of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (https://dash.nichd.nih.gov) and the Multicenter AIDS Cohort Study (25)); access to these data requires permission that generally involves registering, submitting a study concept sheet for approval, and/or paying a fee. Although helpful, these data are not available without restriction, and therefore the study findings are not fully reproducible.

Noting both the importance of reproducible research and the impossibility of making some types of original data publicly available, we propose a pragmatic approach that makes research quasi-reproducible. On a publicly available website without restriction, investigators should post 1) analysis code used in the published study, 2) simulated data, and 3) results created by applying the analysis code used in the published study to the simulated data.

Following the steps listed above, we have begun making our research quasi-reproducible for studies using data from the Caribbean, Central and South America Network for HIV Epidemiology (CCASAnet) (22). The website URL is http://biostat.mc.vanderbilt.edu/ArchivedAnalyses. Publications are listed alongside analysis code used for the published results and quasi-reproduced analysis reports applying the analysis code to simulated data. Here, we briefly discuss each of the 3 steps.

1. Posting analysis code for a published study is simple and straightforward. Point-and-click analyses should be obsolete, and documentation of analyses in the form of an analysis script is the accepted standard (2). Posted analysis code should include any data manipulation required to get from an original data set to the analysis data set. We have been routinely posting our analysis code for the past 5 years and have seen some immediate benefits. First, it has been good for record keeping. Second, it has allowed us to easily share code both internally and externally for future studies. Third, reviewers like it. Recently, a conscientious reviewer followed the link to our code and verified that our analysis procedure was correct despite some confusing language in our submitted manuscript. Posting analysis code takes a bit of courage. Certainly our code is not always optimal and not consistently well documented. It is possible that there are mistakes in the code/analyses that will be uncovered. However, this transparency moves science forward by recognizing such errors.
2. The idea of sharing simulated or synthetic data in place of sensitive original data dates back to at least the 1990s (26, 27). Simulated observational medical data sets have been developed and posted online (e.g., 28, 29). For our purposes, simulated data need to be of exactly the same structure as the data used in the original analysis so that analysis code can be directly run without modification: that is, same variable names and types, same categories for categorical variables, and similar ranges for numerical variables. In general, it would be ideal to maintain relationships between variables in the simulated data set. However, this is not critical; of greatest import is that the data are simulated in a way so that the original analysis code can be run on the simulated data, not that the results using the simulated data match those of the original studies. There are many ways to create simulated data (30–33). It is important to simulate in a manner that preserves anonymity, which will generally involve more than just permuting rows of the original data. Our prosaic approach for simulating our cohort data is to fit simple parametric models of basic patient characteristics for each of the participating cohorts, randomly simulate new data from these models, and then model and simulate additional data conditional on these basic patient characteristics. For some quantities (e.g., days between CD4 measurements and antiretroviral regimens), we sampled directly from the observed data. Our resulting simulated data set, the code used to simulate this data set, and a second simulated data set (simulated from the first simulated data set using the simulation code) are posted at the bottom of the website described above.

3. Finally, we advocate posting results generated by applying the analysis code used in the published study to the simulated data. These results should be perfectly reproducible: An independent analyst should be able to run the analysis code using the publicly available simulated data and obtain exactly the same results as those posted on the website. The only difference between these analyses and those in the publication is the data used. Figures 1 and 2 provide an example. Figure 1 is a reproduction of a figure in a publication from our own cohort. In this time-to-event analysis, we used Cox regression with multiple imputation and restricted cubic splines to examine associations between variables at antiretroviral therapy initiation and the predicted survival probability after 5 years (34). Figure 2 is the figure resulting from the same analysis code used to produce Figure 1 applied to the simulated data. Notice that the results of Figures 1 and 2 are different, which is perfectly fine for our purposes. Our goal is to demonstrate exactly what was done in the original publication, not to generate identical findings using a simulated data set. A motivated reader may want to critically examine how we multiply imputed the data, for example, or they may want to perform a similar analysis in their cohort and can therefore use our code as a starting point.

Such a procedure does not result in fully reproducible research because the original data are not posted. We acknowledge that scientific misconduct could still occur by fabricating data or results or by applying different analyses. To minimize this possibility, there should still be a way for independent researchers to obtain the original de-identified data—that is, by formally requesting data in the usual manner. However, it is our belief that

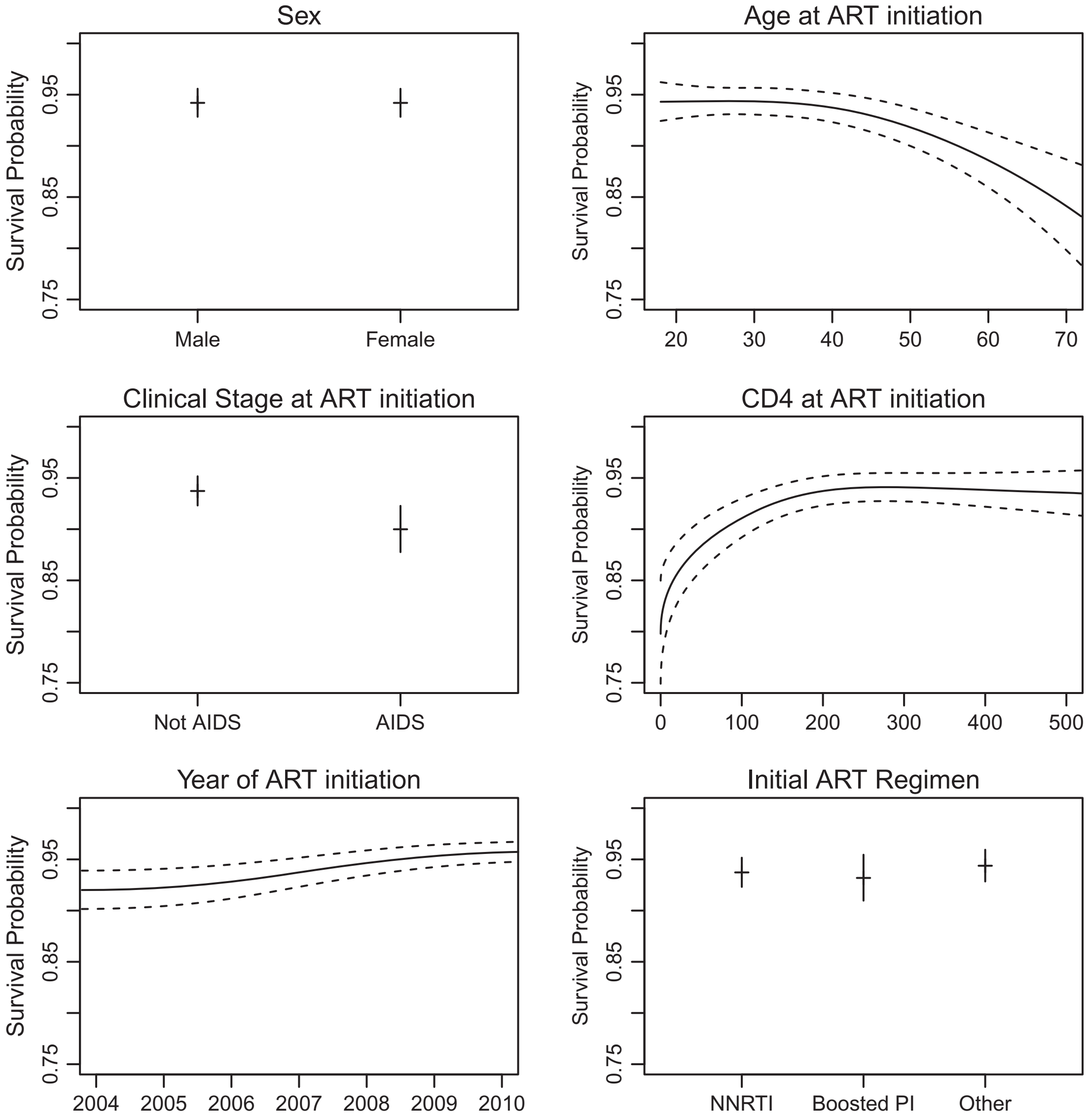

Figure 1. Association between variables at initiation of antiretroviral therapy (ART) and predicted 5-year survival probability in Latin America. This is a reproduction of a figure originally published in the *Journal of the International AIDS Society* (34).

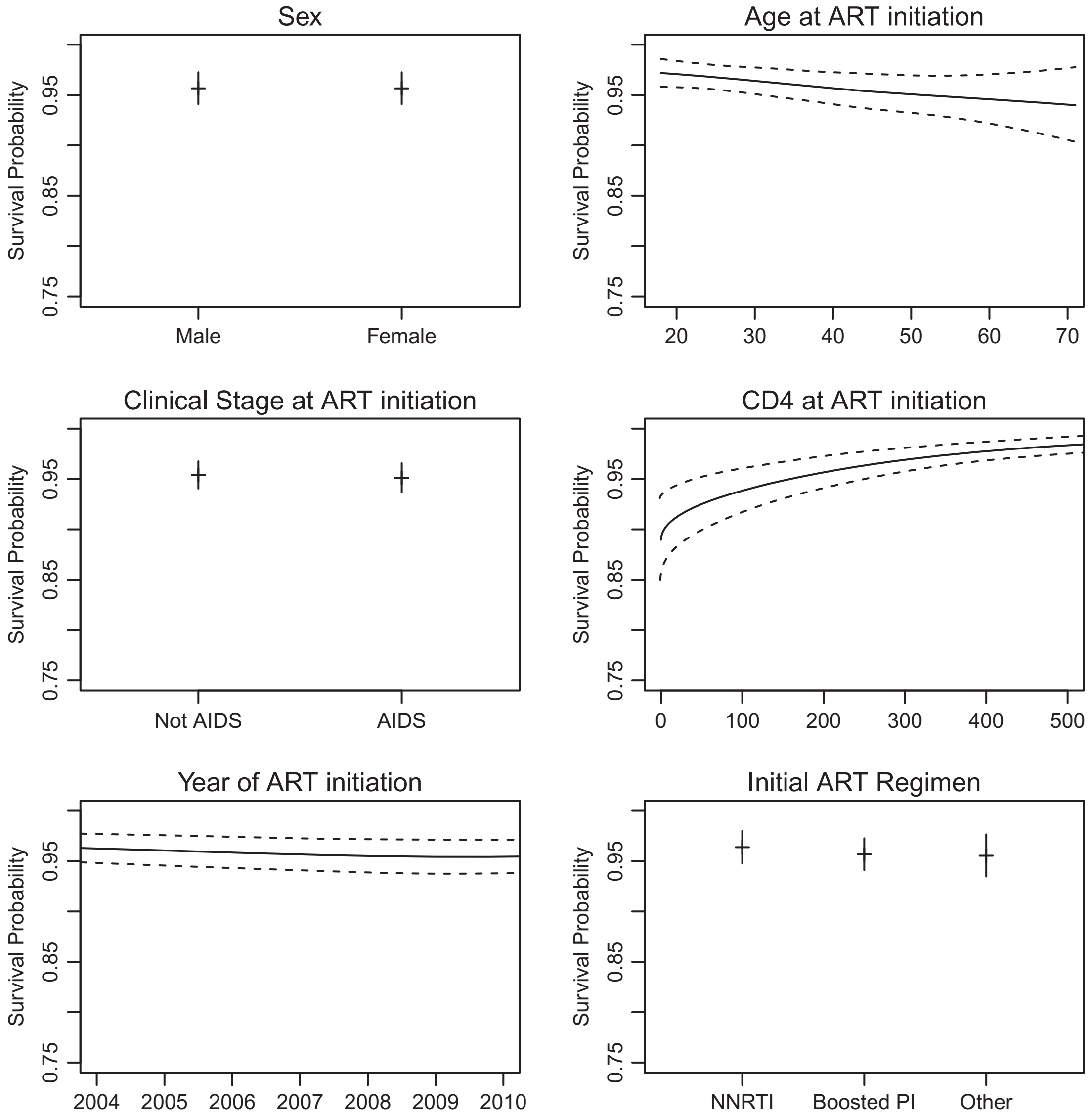

Figure 2. We applied the same analysis code used to generate Figure 1 to simulated data. ART, antiretroviral therapy.

the greatest benefits to making research reproducible will come not from preventing fraud but rather from making analyses transparent for critical evaluation and dissemination. Our proposal for quasi-reproducible research achieves the latter goals and is a significant improvement over current practice. Even simply posting analysis code, which is rarely done in practice, would be progress towards making findings more transparent.

We acknowledge that there are other challenges to the approach described above. Although we have been posting analysis code for years, we only recently simulated data and began posting results generated by applying our original code to the simulated data. By retrospectively performing some of our previous analyses using our simulated data, we have discovered that variable names have changed over the years and that software packages have been modified so that old code no longer runs as it once did. These discoveries have highlighted our need to carefully design and conform to data standards and to accurately document what versions of software packages and data sets we use for analyses. A potential barrier to implementation may be the lack of in-house expertise to simulate a data set. Research groups interested in contributing to reproducible research are invited to build tools that allow for easy data simulation. Our approach does not touch on “pre-registration,” in which researchers publish their intended analysis plans before analysis to avoid fishing through the data for interesting findings (35, 36). Certainly, our website could be enhanced by pre-registration of analysis plans.

In recent years, the reproducible research workflow has been streamlined through the development of helpful tools that allow for version control, collaborative development, and the weaving of analysis code/input/output into one document (2). Many version-control systems are open-sourced, and they allow researchers to track changes to project files; examples include Git

(git-scm.com), Subversion (subversion.apache.org), and CVS (nongnu.org/cvs). Git projects can be placed in an online repository for distribution or collaboration using GitHub (github.com). Finally, reproducible statistical analysis reports are made easier using Sweave or knitr packages in R (r-project.org) to produce LaTeX/PDF or HTML output. A single script reads the data, performs the analysis, and prints the results. SAS (sas.com) has a similar capability called the Output Delivery System.

In conclusion, we have proposed an approach to make research quasi-reproducible in settings in which it is not possible to publicly post data. Specifically, research is quasi-reproducible if analysis code, simulated data sets, and reports applying analysis code to the simulated data are posted to a publicly available website and the code can be run by independent researchers to obtain identical results. Although it is not a perfect solution, we believe such an approach is a suitable compromise between reproducibility and data protection, and we encourage its use.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, School of Medicine, Vanderbilt University, Nashville, Tennessee (Bryan E. Shepherd, Meridith Blevins Peratikos); Department of Medicine, School of Medicine, Vanderbilt University, Nashville, Tennessee (Peter F. Rebeiro, Catherine C. McGowan); and Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, Tennessee (Stephany N. Duda).

This work was supported by the National Institutes of Health (grants U01AI069923, R01 AI093234, and P30 AI110527).

Portions of this work were presented at the 20th International Workshop on HIV Observational Databases, April 7–9, 2016, Budapest, Hungary.

Conflict of interest: none declared.

REFERENCES

1. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. *Am J Epidemiol*. 2006;163(9):783–789.
2. Stodden V, Leisch F, Peng RD. *Implementing Reproducible Research*. Boca Raton, FL: CRC Press; 2014.
3. Peng RD. Reproducible research and biostatistics. *Biostatistics*. 2009;10(3):405–408.
4. Coombes KR, Wang J, Baggerly KA. Microarrays: retracing steps. *Nat Med*. 2007;13(11):1276–1277.
5. Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann Appl Stat*. 2009;3(4): 1309–1334.
6. Deception at Duke: Fraud in Cancer Care. *60 Minutes*. CBS Television. February 12, 2012.
7. Institute of Medicine of the National Academies. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington, DC: National Academies Press; 2015.
8. Taichman DB, Backus J, Baethge C, et al. Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors. *Ann Intern Med*. 2016;164(7):505–506.
9. Longo DL, Drazen JM. Data sharing. *N Engl J Med*. 2016; 374(3):276–277.
10. Lewandowsky S, Bishop D. Research integrity: don't let transparency damage science. *Nature*. 2016;529(7587): 459–461.
11. Haug CJ. From patient to patient–sharing data from clinical trials. *N Engl J Med*. 2016;374(25):2409–2411.
12. Bierer BE, Li R, Barnes M, et al. A global, neutral platform for sharing trial data. *N Engl J Med*. 2016;374(25):2411–2413.
13. Merson L, Gaye O, Guerin PJ. Avoiding data dumpsters–toward equitable and useful data sharing. *N Engl J Med*. 2016; 374(25):2414–2415.
14. Jumbe NL, Murray JC, Kern S. Data sharing and inductive learning–toward healthy birth, growth, and development. *N Engl J Med*. 2016;374(25):2415–2417.
15. Stebbins M. Expanding public access to the results of federally funded research. http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research. Published February 22, 2013. Accessed July 7, 2017.
16. National Institutes of Health. National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data From NIH Funded Scientific Research. https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf. Published February 2015. Accessed July 7, 2017.
17. National Institutes of Health. NIH genomic data sharing policy. Notice number NOT-OD-14-124. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html. Published August 27, 2014. Accessed July 7, 2017.
18. Coady SA, Wagner E. Sharing individual level data from observational studies and clinical trials: a perspective from NHLBI. *Trials*. 2013;14:201.
19. van Puhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. *BMC Public Health*. 2014;14:1144.
20. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. *BMJ*. 2015;350:h1139.
21. El Emam K, Jonker E, Arbuckle L, et al. A systematic review of re-identification attacks on health data [published correction appears in *PLoS One*. 2015;10(4):e0126772]. *PLoS One*. 2011; 6(12):e28071.
22. McGowan CC, Cahn P, Gotuzzo E, et al. Cohort profile: Caribbean, Central and South America Network for HIV research (CCASAnet) collaboration within the International Epidemiologic Databases to Evaluate AIDS (IeDEA) programme. *Int J Epidemiol*. 2007;36(5):969–976.
23. Gange SJ, Kitahata MM, Saag MS, et al. Cohort profile: the North American AIDS Cohort Collaboration on Research and Design (NA-ACCORD). *Int J Epidemiol*. 2007;36(2):294–301.
24. Egger M, Ekouevi DK, Williams C, et al. Cohort profile: the International epidemiological Databases to Evaluate AIDS (IeDEA) in sub-Saharan Africa. *Int J Epidemiol*. 2012;41(5): 1256–1264.
25. The Multicenter AIDS Cohort Study. MACS Public Data Set. https://statepi.jhsph.edu/macs/pdt.html. Updated January 2017. Accessed February 13, 2017.
26. Rubin DB. Discussion: statistical disclosure limitation. *J Off Stat*. 1993;9(2):461–468.
27. Reiter JP. Satisfying disclosure restrictions with synthetic datasets. *J Off Stat*. 2002;18:531–543.
28. Observational Medical Outcomes Partnership. Simulated Data. http://omop.org/OSIM2. Updated April 11, 2012. Accessed February 13, 2017.
29. HIV Cohorts Data Exchange Protocol. Public Example Data. http://hicdep.org. Published September 2, 2014. Accessed February 13, 2017.

30. Murray RE, Ryan PB, Reisinger SJ. Design and validation of a data simulation model for longitudinal healthcare data. *AMIA Annu Symp Proc*. 2011;2011:1176–1185.
31. Gruber S. A causal perspective on OSIM2 Data Generation, with Implications for Simulation Study Design and Interpretation. *J Causal Inference*. 2015;3(2):177–187.
32. Phillips AN, Sabin C, Pillay D, et al. HIV in the UK 1980–2006: reconstruction using a model of HIV infection and the effect of antiretroviral therapy. *HIV Med*. 2007;8(8): 536–546.
33. Guru M, Gupta T, Kohler J, et al. A synthetic-data-creation tool based on cross sectional panel data. AMIA; 2016. https://knowledge.amia.org/amia-59309-tbi2016-1.3013007/t005-1.3013815/f005-1.3013816/a113-1.3013845/ap119-1.3013848. Accessed July 7, 2017.
34. Carriquiry G, Fink V, Koethe JR, et al. Mortality and loss to follow-up among HIV-infected persons on long-term antiretroviral therapy in Latin America and the Caribbean. *J Int AIDS Soc*. 2015;18:20016.
35. Humphreys M, Sanchez de la Sierra R, van der Windt P. Fishing, commitment, and communication: a proposal for comprehensive non binding research registration. *Pol Anal*. 2013;21(1):1–20.
36. Monogan JE. A case for registering studies of political outcomes: an application in the 2010 elections. *Pol Anal*. 2013; 21(1):21–37.