American Journal of Epidemiology

Vol. 163, No. 9
DOI: 10.1093/aje/kwj093
Advance Access publication March 1, 2006

Commentary

Reproducible Epidemiologic Research

Roger D. Peng, Francesca Dominici, and Scott L. Zeger

From the Biostatistics Department, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

Received for publication November 4, 2005; accepted for publication January 10, 2006.

The replication of important findings by multiple independent investigators is fundamental to the accumulation of scientific evidence. Researchers in the biologic and physical sciences expect results to be replicated by independent data, analytical methods, laboratories, and instruments. Epidemiologic studies are commonly used to quantify small health effects of important, but subtle, risk factors, and replication is of critical importance where results can inform substantial policy decisions. However, because of the time, expense, and opportunism of many current epidemiologic studies, it is often impossible to fully replicate their findings. An attainable minimum standard is "reproducibility," which calls for data sets and software to be made available for verifying published findings and conducting alternative analyses. The authors outline a standard for reproducibility and evaluate the reproducibility of current epidemiologic research. They also propose methods for reproducible research and implement them by use of a case study in air pollution and health.

air pollution; information dissemination; models, statistical

Abbreviation: NMMAPS, National Morbidity, Mortality, and Air Pollution Study.

Determinants of human disease are commonly investigated by epidemiologic studies focused on a particular subpopulation, time frame, and geographic location. Findings from such studies can play an important role in policy decisions affecting public health (1). Yet epidemiologic research has been criticized as being increasingly unreliable. One review of the field a decade ago raised questions about the reliability of observational epidemiologic studies when quantifying the health effects of important, but subtle, risk factors such as second-hand smoke, air pollution, and diet (2).

Scientific evidence is strengthened when important findings are *replicated* by multiple independent investigators using independent data, analytical methods, laboratories, and instruments. Replication, as described here, has long been the standard in the biologic and physical sciences and is of critical importance in epidemiologic studies, particularly when they can impact broad policy or regulatory decisions. Because of the time and expense involved with epidemiologic studies, many are often not fully replicable, and policy decisions must be made with the evidence at hand.

An attainable minimum standard is *reproducibility*, where independent investigators subject the original data to their own analyses and interpretations. Reproducibility calls for data sets and software to be made available for 1) verifying published findings, 2) conducting alternative analyses of the same data, 3) eliminating uninformed criticisms that do not stand up to existing data, and 4) expediting the interchange of ideas among investigators. Ultimately, all scientific evidence should be held to the standard of full replication and the confirmation of important findings by independent investigators. However, the desire to quantify small health effects and the significant weight placed on epidemiologic findings in the policy-making process create a need for epidemiologic studies to meet a minimum standard. We propose reproducibility to be this minimum standard.

There are a number of new developments that are intensifying the need for reproducible epidemiologic research.

Correspondence to Dr. Roger Peng, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, E3535, Baltimore, MD 21205 (e-mail: rpeng@jhsph.edu).

The signal-to-noise ratio in today's epidemiologic studies tends to be smaller than it was in decades past simply because much of the ''low-hanging fruit'' has been picked. Factors with large relative risks, such as smoking, socioeconomic status, family history, and obesity, are well established for major diseases. The targets of current investigations tend to have smaller relative risks that are more easily confounded. For example, in air pollution epidemiology, the national relative risk of increased mortality is estimated to be 1.005 per 10 parts per billion of 24-hour ozone. Remarkably, an integrated analysis of mortality in 95 metropolitan areas can detect this signal, which translates into thousands of excess deaths per year given the universality of ozone exposure (3). Nevertheless, the potential for unexplained confounding cannot be denied for such a small relative risk (4, 5).

The explosion of new biologic measurements presents exciting opportunities for epidemiologic studies. We can now quantify DNA sequences, single nucleotide polymorphisms, and gene and protein expression. We can image the structure and function of the brain and other organs. We can quantify diet with lengthy dietary-recall questionnaires and can quantify disease symptoms and health conditions with multiitem instruments. However, because the data are inherently complex and high dimensional, there is an increased potential for identifying spurious associations between their components and risk factors or health outcomes (6).

The widespread availability of statistical and computing technology is yet another factor contributing to the potential for false positive epidemiologic findings. It is now easy for a researcher to routinely engage in sophisticated optimizations across a large number of models and/or variables to identify associations that are of potential scientific interest. Even with a single risk factor and a single response, it is standard practice to consider a potentially large number of models in an effort to adjust for differences among the exposed and the unexposed. As the number of covariables measured increases, so do the degrees of freedom for influencing the association between the risk factor and outcome and for identifying subgroups in which the association is particularly strong.

The developments identified above also have the potential to increase the power and precision of epidemiologic research by enhancing our understanding of disease mechanisms and leading to studies with more targeted hypotheses. Modern computing makes possible the organization and analysis of large databases, so that we can look farther and wider for systematic patterns indicating the health effects of various risk factors. The reproducibility of epidemiologic findings from current and future studies will be crucial to providing the substance for informed debate regarding policies affecting the public's health.

DEFINING REPRODUCIBLE RESEARCH

Reproducibility is a minimum step that can be taken in any study. In the context of epidemiology, a study is reproducible when it satisfies the criteria in table 1, adapted from the paper by Schwab et al. (7) and others. We illustrate reproducibility requirements separately for each of the following research components: data, methods, documentation, and distribution.

TABLE 1. Criteria for reproducible epidemiologic research

Research component	Requirement
Data	Analytical data set is available.
Methods	Computer code underlying figures, tables, and other principal results is made available in a human-readable form. In addition, the software environment necessary to execute that code is available.
Documentation	Adequate documentation of the computer code, software environment, and analytical data set is available to enable others to repeat the analyses and to conduct other similar ones.
Distribution	Standard methods of distribution are used for others to access the software, data, and documentation.

Epidemiologic data sets can be difficult to define because of the complexity of the data collection and the preprocessing. In addition, a published finding may use only a subset of the data collected for a particular study, the other parts being irrelevant to the reproducibility of the finding. We therefore separate the data into ''analytical'' data, which serve as the input to a statistical analysis program to produce the results found in the table/figure supporting the paper's conclusions, and ''measured'' data, which consist of all data and functions thereof that are used to create the analytical data, whether or not they are part of the analytical data. This classification is crude and far from ideal, but it strikes a compromise between those data that are necessary for reproduction and those that may be of secondary interest. We propose as a first requirement that the analytical data set be made available to others for reproducing results.

With the increased use of advanced statistical methodology and larger data sets, analyses today are almost always implemented on a computer. Given that, the simplest way to reproduce the statistical methods is to examine the computer code or instructions for the analysis. While some analyses may be considered too rudimentary to warrant publishing computer code, most statistical software routines, for example, contain many options that need to be set by the user. Since it is not always clear from the outset which options can have an impact on the numerical results, this information can be critical for reproducing scientific findings, particularly when investigating small relative risks (8, 9).

REPRODUCIBILITY OF CURRENT EPIDEMIOLOGIC RESEARCH

To measure the reproducibility of recent epidemiologic research, we reviewed 90 epidemiology articles from the *American Journal of Epidemiology* and the *Journal of the American Medical Association*. We selected every article published in 2005 in the time period between January and the time at which we conducted the review (May). We developed a questionnaire to collect information relevant to

TABLE 2. Results from examining the epidemiologic literature: articles from the *American Journal of Epidemiology* and the *Journal of the American Medical Association* published between January 2005 and May 2005

	No. of papers
Total papers collected	90
Observational studies	69
Cross-sectional	20
Case-control	20
Cohort	29
Source of outcome data	
Original study	31
Ongoing study	29
Government	8
Other	1
Statistical analysis implementation	
Not reported	21
By hand	0
Use of software package	48
Method of processing measured data	
Not reported	43
By hand	1
Use of software package	13
Outcome data reported to be available	0
Exposure data reported to be available	1
Code for statistical analysis available	0
Code for processing measured data available	0

reproducibility that is available at http://www.biostat.jhsph.edu/~rpeng/reproducible/survey/. Some of the articles selected during this time period were excluded on the basis of the criteria developed in the questionnaire.

We focused our review of an article on the abstract's *concluding statement* summarizing the main scientific findings. Each statement contained information about an outcome and an exposure variable or risk factor, as well as potential confounders and/or effect modifiers. For a given article, we first determined the type of study: randomized trial, methodology, literature review or meta-analysis, or original observational study. The survey addressed only the last category, as it forms the bulk of epidemiologic research. Articles describing other types of studies were excluded. Given an observational study, we determined the study design, the primary outcome and exposure variables, and the tables or figures providing the evidence supporting the concluding statement. We also recorded details of how the statistical analysis was implemented and whether or not data were reported to be available. The data availability was determined separately for the primary outcome, the exposure, any potential confounders, and any effect modifiers, since it is commonly the case that the different variables have different sources.

The results of our survey are summarized in table 2. In total, we examined 90 articles, 69 of which were observational studies and had either a cross-sectional, case-control, or cohort design: We focus only on these 69 articles. For 84 percent of the articles, the data for the outcome and exposure came from original studies or from large ongoing studies. None of the articles reported that the outcome data were available to others, either from the authors or from a third party. In only one study were the exposure data available. However, in this study, ''time'' was the relevant exposure variable, with the study examining trends in cardiovascular risk factors through a series of surveys. It should be noted that we did not attempt to contact the authors and to request the analytical data and computer code used for their published analyses. Had the authors been contacted, it is not known how many would have been willing or able to provide the data and code.

Thirty percent of the articles did not report how the statistical analyses were implemented, while the remaining 70 percent reported using a specific software package. Neither the software for the statistical analysis nor the software for processing the measured data into analytical data was reported to be available. Of the articles where measured data required processing, 93 percent did not report how this processing was implemented.

METHODS FOR REPRODUCIBLE RESEARCH

Articles printed in journals are still the primary means by which scientific results are presented. However, reproducible research as defined in the previous section requires that arrangements be made for the distribution of analytical data and methods. While journals typically govern the distribution of the scientific findings, the task of distributing data and methods is relegated to the authors.

Today, the World Wide Web is the most convenient medium for distributing information to other researchers and is already playing a central role in the implementation of reproducible research. Many journals now have websites that can host supplementary materials for published articles, such as data that can be downloaded along with computer code for reproducing specific analyses. In addition, more detailed explanations of methods and complementary figures can be provided to the reader who intends to reproduce the published findings and to conduct competing analyses of the same data set.

LITERATE PROGRAMMING

The practice of posting data and code on either personal or journal websites is a significant first step. While making data and code available is certainly necessary, it is typically insufficient for others to reproduce results. An author must additionally provide details about how the code is linked to the data and which code sections are applied to which data.

A *compendium* is an article linked together with the data and code necessary for producing all of the results in the article (10, 11). The tools for constructing a compendium are modeled on the idea of *literate programming*, a phrase coined by Donald Knuth (12) and a concept extended by many others (13). A literate program combines a documentation

language with a programming language to produce an overall program that is self-documenting and more ''literate.'' Knuth's original WEB system for writing literate programs combined the T_EX documentation language with the Pascal programming language. In a literate program, one *weaves* the text and code to produce a human-readable document and *tangles* the code to produce a machine-executable file. The advantage of the literate programming approach is that the code and text can provide a running commentary on each other.

The specific details of how a compendium is created depend on the computing environment and programming languages used by the author. Gentleman and Temple Lang (10) propose using the R software environment (15) coupled with the L^AT_EX document-formatting language. The general idea of a compendium is not tied to any one software package but, rather, the ideas of literate programming are easily applied to the documentation and programming language with which the researcher is familiar.

OPEN RESEARCH DATA LICENSES

It is understandable that authors do not make their research data available, if only because once data are released, there is little control over how the data will be used (16). A regime whereby partial rights to research data could be granted would allow some flexibility for authors to make data available without giving up complete control. For reproducible research to become the standard in epidemiology, limited access to data is a necessity.

We propose different classes of data licenses that provide partial rights to research data under prespecified conditions. In developing these classes, we borrow from standards created by the Creative Commons project (http://creativecommons.org), an organization devoted to creating licenses that provide partial rights to literary works. These ideas have also been discussed in the software community, where ''open source'' licenses are commonly used to provide partial rights to software products (e.g., the Open Source Initiative at http://opensource.org/).

The following list defines four possible classes of data licenses in order of increasing restrictiveness. We choose not to use precise legal definitions but rather outline the basic ideas.

1. *Full access*. The data can be used for any purpose.
2. *Attribution*. The data can be used for any purpose so long as the authors are cited (a specific citation should be provided).
3. *Share alike*. The data can be used to produce new findings or results. Any modifications to the data, including transformations, additions, or linkages to other data, which are used to produce the new findings, must be made available under the same terms.
4. *Reproduction*. The data can be used for the purpose of reproducing the results in the associated published article or for commenting on those results via a letter to the editor. No original findings based on the data may be published without explicit permission from the original investigators in a separate agreement.

Licenses providing partial rights to data can benefit both the donor and the recipient. The recipient obtains access to the data and an explicit understanding of the rights granted to him or her. The donor meets data disclosure obligations (from either granting agencies or journals) and is provided some measure of control over others' use of the data in an undesirable manner. In addition, the donor is relieved of having to negotiate potentially numerous requests for the data set. With the benefits also come the costs of such a licensing regime. The recipient must accept limitations to the data set by the donor, while the donor must initially invest time to arrange for data sharing and risks agreement violations by those using the data.

REPRODUCIBLE RESEARCH IN AIR POLLUTION AND HEALTH: A CASE STUDY

We demonstrate one implementation of reproducible research with a large observational study of the health effects of air pollution, the National Morbidity, Mortality, and Air Pollution Study (NMMAPS). NMMAPS is a national time-series study on the short-term health effects of air pollution, the goals of which are to 1) integrate multiple government databases that contain information on population health, ambient air pollution levels, weather variables, and socioeconomic variables for air pollution epidemiology; 2) develop statistical methods and computational tools for analyzing and interpreting the resulting database; and 3) estimate the short-term effects of air pollution on mortality and its uncertainty for the largest US metropolitan areas and for several geographic regions (17, 18).

Because of the regulatory context, quantification of air pollution risks is controversial. The assessments of risks are part of a highly charged policy debate and, consequently, statistical methods and data sources are subject to intense scrutiny by other scientists and interested parties. A recent review of the epidemiologic evidence on the health effects of fine particles described this debate (19), which will likely be revisited now that the Environmental Protection Agency (20) has promulgated its latest daily and annual standards for particulate matter. In the last few years, NMMAPS and several other large epidemiologic studies (21–23) have been a part of this policy debate (5).

As a first step toward developing higher standards of reproducibility, we created the Internet Health and Air Pollution Surveillance System for disseminating the entire NMMAPS database and the software for implementing all of our statistical analyses (http://www.ihapss.jhsph.edu/). Other scientists can fully reproduce our results, apply our methodology to their own data, or apply their methodology to the NMMAPS database. One of the goals of our approach is to raise the level of scientific debate by making all of our methods publicly available and to create new tools and standards that encourage others to do the same.

In addition to the Internet Health and Air Pollution Surveillance System website, we have created a compendium for a recent publication, ''Seasonal Analyses of Air Pollution and Mortality in 100 US Cities'' (24), which contains an analysis of the seasonal and regional variability of the health

TABLE 3. Making results from the National Morbidity, Mortality, and Air Pollution Study reproducible*

Research component	What we have done
Data	The entire NMMAPS† database is available to the public via the iHAPSS† website and the NMMAPS data package for R; the data are available under a "full access" class of license.
Methods	A full compendium written in L^AT_EX and R is available for download.
Documentation	We have outlined our data-processing pipeline on the iHAPSS website, and papers/technical reports are available for download.
Distribution	We use the World Wide Web to disseminate our data and software.

* Details at http://www.biostat.jhsph.edu/~rpeng/reproducible/.
† NMMAPS, National Morbidity, Mortality, and Air Pollution Study; iHAPSS, Internet Health and Air Pollution Surveillance System.

effects of particulate matter. To allow others to reproduce our findings, we have developed a simple webpage that contains links to all of the data and computer code for generating the figures and tables in the article. The compendium is written with the literate programming techniques described in the previous section, an excerpt of which is shown in the Appendix. Readers can inspect the code for producing the results and for creating the tables/figures, as well as download the code and data to their own computers to run other analyses or produce different figures. A summary of our efforts can be found in table 3, and we have posted complete information about the data and methods used in this paper at http://www.biostat.jhsph.edu/~rpeng/reproducible/. In a recent study of fine particulate matter and hospitalizations among the elderly (25), we have applied the same principles of reproducibility and have posted information at http://www.biostat.jhsph.edu/MCAPS/.

DISCUSSION

In this article, we have proposed that reproducible research be the minimum standard in disseminating epidemiologic findings and have demonstrated the possibilities with a large ongoing study of air pollution and health. The policy implications of epidemiologic studies coupled with the investigation of smaller targets, the increasing use of complex databases, and the application of sophisticated statistical modeling can lead to research that is subject to intense scrutiny. The reproducibility of principal findings can foster rational discussions regarding the evidence in the data and serve as a bulwark against uninformed criticism.

The standard of reproducibility addresses some critical scientific issues, but its reach is still limited. In order to identify the issues that reproducibility can address, we must first agree on a model of the research process itself. One can think of an epidemiologic study as a sequence of stages, starting from the generation of the data by natural or other processes, to the collection of these data, to data processing and analysis, and then to the reporting of results. Prior to the generation of the data, one might also include the formulation of the hypotheses and the design of the study.

Reproducibility becomes meaningful in the stages of a study following the data collection. The processing of the data, as well as the analysis and subsequent presentation, can all be inspected if the research is reproducible. Beyond checking for statistical and programming errors, one can evaluate the sensitivity of the findings to certain modeling choices. By having the computer code used to process and analyze the data, others can obtain useful information regarding the many important choices made as part of the study.

Among the issues that cannot be addressed by reproducible research are those arising from stages of the research process prior to the data collection stage. Questions about the study design, the selection of subjects, the handling of nonrespondents, and many other facets of a study cannot be resolved with the analytical data alone. Similarly, it may not be possible to examine all relevant modeling choices, particularly those involving variables for which no data were originally collected. However, when we discuss these types of issues, we are moving closer to calling for full replication. If a particular study is fully replicable, then all aspects of the original study can be adjusted or modified. Clearly, full replication remains the ultimate standard by which we evaluate scientific evidence.

Data availability is the first and foremost challenge to reproducible epidemiologic research. Although this problem is not unique to epidemiology (26), being observational, evidence from epidemiologic studies is more often open to differing interpretations. It is exactly in such a circumstance that work needs to be open and reproducible. We have proposed a framework of "partial rights" to research data that would modulate the all-or-nothing scenario that exists today.

One impediment to making data available is preserving confidentiality. Health data are often obtained by making promises to individuals (either directly or through an intermediary) that confidential information about those individuals will not be released to the public. Under our definition, it would seem impossible to simultaneously honor those promises and make one's research reproducible. However, while the measured data often must be kept confidential, it may still be possible to provide summary statistics of the data upon which the analysis is based. For example, with time-series studies of air pollution and mortality, the individual mortality data are confidential, but the time series of aggregate counts for each county can be made available for large enough counties. Since the analysis is based entirely on those aggregate values, there is no need to release the individual-level data. This is a limited example, and although there is active discussion in the literature about disclosure limitation methods (27), the issue of releasing data for the purposes of reproducing scientific findings is in need of serious discussion.

The literature review served both to assess the state of reproducibility in the epidemiologic literature and to provide a "checklist" for producing a reproducible study. In addition to data availability, we identified a number of additional problems preventing epidemiologic findings from being reproduced. The sparse reporting of analytical methods

and the lack of computer code describing those methods are of concern. We have demonstrated how to use literate programming techniques to produce a reproducible document and the Web for distributing data and software. The reproducibility of the document is ensured by the use of tools that allow text and code to be intermingled to form a common source for the finished paper. In general, programming languages and statistical packages tend to change, and we do not presume that there is a single ''best'' environment. Rather, we describe the general concept of literate programming and highlight some specific tools that are available for encouraging such practice.

The call for reproducible research has already been echoed in other fields where computation and complex statistical methodology are critical for obtaining substantive results (7, 10, 11, 28–31). Biologists have made enormous progress toward integrating databases, sharing software, and making their analyses reproducible. Journals such as *Science* and *Nature* require deposition of biologic data into public repositories at the time of publication, and organizations such as the Microarray Gene Expression Data Society have developed rigorous standards for the reporting of microarray data (32). The Inter-University Consortium for Political and Social Research is a vast repository for social science data, providing archiving resources as well as standardization of data sets for a number of software environments. Social science investigators intent on making their research reproducible have a clear resource for sharing their data.

In addition to various field-specific efforts, the US National Institutes of Health now requires many of its grantees to implement a data-sharing policy for any research sponsored by the Institutes. Even more broadly, the federal government, via the Shelby Amendment of 1999 and the subsequent revision to the Office of Management and Budget Circular A110, requires data from any federally sponsored research to be made available upon request if the data were used in ''developing an agency action that has the force and effect of law'' (33, p. 220). It is not yet known what the full impact of either of these policies will be on the reproducibility of all biomedical research.

In the absence of full replication, reproducibility should become the minimum standard for epidemiologic research. In particular, studies with potential policy impact should be made reproducible to allow others to verify published findings and to conduct alternative analyses of the data. We have demonstrated through our case study that the standard of reproducibility can be achieved and have proposed a framework in which the results can be disseminated. The apparent unreliability of epidemiologic investigations predicted 10 years ago can be thwarted today by adopting new standards and embracing a more open research environment.

ACKNOWLEDGMENTS

Funding was provided by the National Institute of Environmental Health Sciences (grants R01ES012054 and R01ES012054-03S1), the National Institute of Environmental Health Sciences Center in Urban Environmental Health (grant PES003819), and the Health Effects Institute (grant HEI 025).

The authors thank Dr. Ronald S. Brookmeyer for his comments and suggestions and the faculty and students of the departments of Biostatistics and Epidemiology who helped with the survey.

Conflict of interest: none declared.

REFERENCES

1. Samet JM. Epidemiology and policy: the pump handle meets the new millenium. Epidemiol Rev 2000;22:145–54.
2. Taubes G. Epidemiology faces its limits. Science 1995;269:164–9.
3. Bell ML, McDermott A, Zeger SL, et al. Ozone and short-term mortality in 95 US urban communities, 1987–2000. JAMA 2004;292:2372–8.
4. Kaiser J. Researchers and lawmakers clash over access to data. Science 1997;277:467. (DOI: 10.1126/science.277.5325.467).
5. Kaiser J. Evidence mounts that tiny particles can kill. Science 2000;289:22–3. (DOI: 10.1126/science.289.5476.22).
6. The PLoS Medicine Editors. Why bigger is not yet better: the problems with huge datasets. PLoS Med 2005;2:e55. (DOI: 10.1371/journal.pmed.0020055).
7. Schwab M, Karrenbach N, Claerbout J. Making scientific computations reproducible. Comput Sci Eng 2000;2:61–7. (http://sepwww.stanford.edu/research/redoc/).
8. Kaiser J. Software glitch threw off mortality estimates. Science 2002;296:1945–7. (DOI: 10.1126/science.296.5575.1945).
9. Colburn KA, Johnson PRS. Air pollution concerns not changed by S-PLUS flaw. Science 2003;299:665–6. (DOI: 10.1126/science.1082105).
10. Gentleman R, Temple Lang D. Statistical analyses and reproducible research. Berkeley, CA: Berkeley Electronic Press, 2004. (http://www.bepress.com/bioconductor/paper2).
11. Ruschhaupt M, Huber W, Poustka A, et al. A compendium to ensure computational reproducibility in high-dimensional classification tasks. Stat Appl Genet Mol Biol 2004;3:article 37. (http://www.bepress.com/sagmb/vol3/iss1/art37).
12. Knuth DE. Literate programming. Stanford, CA: Center for the Study of Language and Information, 1992.
13. Rossini A. Literate statistical practice. In: Hornik K, Leisch F, eds. Presented at the 2nd International Workshop on Distributed Statistical Computing, Vienna, Austria, March 15–17, 2001. (http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings/Rossini.pdf).
14. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314.
15. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2005. (http://www.Rproject.org).
16. Rowen L, Wong GKS, Lane RP, et al. Intellectual property. Publication rights in the era of open data release policies. Science 2000;289:1881. (DOI: 10.1126/science.289.5486.1881).
17. Samet JM, Zeger SL, Dominici F, et al. The National Morbidity, Mortality, and Air Pollution Study, part I: methods and methodological issues. Cambridge, MA: Health Effects Institute, 2000.
18. Samet JM, Zeger SL, Dominici F, et al. The National Morbidity, Mortality, and Air Pollution Study, part II: morbidity and mortality from air pollution in the United States. Cambridge, MA: Health Effects Institute, 2000.

19. Kaiser J. Mounting evidence indicts fine-particle pollution. Science 2005;307:1858–61.
20. National ambient air quality standards for particulate matter. Research Triangle Park, NC: US Environmental Protection Agency Office of Air Quality Planning and Standards, 2005.
21. Dockery D, Pope CA, Xu X, et al. An association between air pollution and mortality in six U.S. cities. N Engl J Med 1993;329:1753–9.
22. Pope CA, Thun M, Namboodiri M, et al. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am J Respir Crit Care Med 1995;151:669–74.
23. Pope CA, Burnett RT, Thun MJ, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA 2002;287:1132–41.
24. Peng RD, Dominici F, Pastor-Barriuso R, et al. Seasonal analyses of air pollution and mortality in 100 US cities. Am J Epidemiol 2005;161:585–94.
25. Dominici F, Peng RD, Bell ML, et al. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA 2006;295:1127–35.
26. Campbell EG, Clarridge BR, Gokhale M, et al. Data withholding in academic genetics: evidence from a national survey. JAMA 2002;287:473–80.
27. Fienberg SE. Conflicts between the needs for access to statistical information and demands for confidentiality. J Off Stat 1994;10:115–32.
28. Gentleman R. Reproducible research: a bioinformatics case study. Stat Appl Genet Mol Biol 2005;4:article 2. (http://www.bepress.com/sagmb/vol4/iss1/art2).
29. Johnson DH. Three objections to databases answered. In: APS observer. Vol 14. Washington, DC: Association for Psychological Science, 2001. (http://www.psychologicalscience.org/observer/1101/database.html).
30. Buckheit J, Donoho DL. Wavelab and reproducible research. In: Antoniadis A, ed. Wavelets and statistics. New York, NY: SpringerVerlag, 1995.
31. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80. (DOI: 10.1186/gb-2004-5-10-r80).
32. Ball CA, Sherlock G, Parkinson H, et al. Standards for microarray data. (Letter). Science 2002;298:539. (DOI: 10.1126/science.298.5593.539b).
33. 45CFR74.36. Washington, DC: US Government Printing Office, 2003:220–1. (http://www.peo7.com/CFRFiles/PEOusCFR_45PUBLICWELFARE_119259.htm).

APPENDIX

Sweave Example

The following is taken from a vignette for reproducing the results reported in the paper by Peng et al. (24). The document is written in the L^AT_EX document-formatting language, and the portions between the <<>>= and @ symbols are written in the R language.

```
The national average estimates of the overall and seasonal short-term effects of \PMTen\
on mortality for lags 0, 1, and 2 are summarized in Table 2. These estimates were obtained by
pooling the city-specific maximum likelihood estimates from the main effect and pollutant-
season interaction models according to the hierarchical normal model.

<<nationalAverageEstimates,results=tex,echo=false>>=
Seasons <- c(''Winter'', ''Spring'', ''Summer'', ''Fall'', ''All Seasons'')
Lags <- paste(''Lag'', 0:2); exclude <- c(''hono'', ''anch'')

## Load non-seasonal estimates
load(file.path(''results'', ''city-specific-est.pm10.rda''))
results <- lapply(results, function(x) x[setdiff(names(x), exclude)])

## Pool estimates
betacovTotal <- lapply(results, extractBetaCov, pollutant = poll)
pooledTotal <- lapply(betacovTotal, poolCoef)

## Load seasonal estimates
load(file.path(''results'', ''seasonal.factor2.lag.012.pm10.rda''))
results <- lapply(results, function(x) x[setdiff(names(x), exclude)])

## Pool estimates by season
pooledSeas <- lapply(results, coefSeasonal, pollutant=poll,
method=method)
pooled <- lapply(seq(along = pooledSeas), function(i) {
    m <- rbind(pooledSeas[[i]], pooledTotal[[i]])
    rownames(m) <- Seasons
    m
})
@
```