Reproducible Research in Scientific Journals

Roger D. Peng, PhD
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

CSE 2011
Replication

• The ultimate standard for strengthening scientific evidence is replication of findings and conducting studies with independent
  – Investigators
  – Data
  – Analytical methods
  – Laboratories
  – Instruments

• Replication is particularly important in studies that can impact broad policy or regulatory decisions
Why Do We Need Reproducible Research?

• Today, scientific papers published in journals represent the *advertising* of the research (Jon Claerbout)
• Some studies cannot be replicated
  – No time, No money, Unique/opportunistic
• New technologies increasing data collection throughput; data are more complex and extremely high dimensional
• Existing databases can be merged into new "megadatabases"
• Computing power is greatly increased, allowing more sophisticated analyses
• For every field "X" there is a field "Computational X"
Current State of Affairs

Share

? 

Don’t Share
Where’s the Research?
Where’s the Research?

Measured Data → Analytic Data → Computational Results → Figures → Tables → Numerical Summaries → Article

Processing code → Analytic code → Presentation code

Author → Reader

Text
Reproducibility as a Minimum Standard

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.
Reproducible Air Pollution and Health Research

• Estimating small (but important) health effects in the presence of much stronger signals
• Results inform substantial policy decisions, affect many stakeholders
  – EPA regulations can cost billions of dollars
• Complex statistical methods are needed and subjected to intense scrutiny
What is Reproducible Research?

<table>
<thead>
<tr>
<th>Research component</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Analytical data set is available.</td>
</tr>
<tr>
<td>Methods</td>
<td>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.</td>
</tr>
<tr>
<td>Documentation</td>
<td>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.</td>
</tr>
<tr>
<td>Distribution</td>
<td>Standard methods of distribution are used for others to access the software, data, and documentation.</td>
</tr>
</tbody>
</table>
Who are the Players?

• Authors
  – Want to make their research reproducible
  – Want tools for RR to make their lives easier (or at least not much harder)

• Readers
  – Want to reproduce, verify, and perhaps expand upon interesting findings
  – Want tools for RR to make their lives easier
Challenges

• Authors must undertake considerable effort to put data/results on the web (may not have resources like a web server)
• Readers must download data/results individually and piece together which data go with which code sections, etc.
• Authors/readers must manually interact with web sites via a browser
• There is no single document to integrate data analysis with textual representations; i.e. data, code, and text are not linked
The Reality

• Authors
  – Just put stuff on the web
  – Journal supplementary materials
  – There are some central databases for various fields (e.g. biology, ICPSR)

• Readers
  – Just download the data and (try to) figure it out
  – Get the software and run it
RR Policy at *Biostatistics*

**Reproducible research and *Biostatistics***

ROGER D. PENG

1. Introduction and Motivation

The replication of scientific findings using independent investigators, methods, data, equipment, and protocols has long been, and will continue to be, the standard by which scientific claims are evaluated. However, in many fields of study there are examples of scientific investigations that cannot be fully replicated because of a lack of time or resources. In such a situation, there is a need for a minimum standard that can fill the void between full replication and nothing. One candidate for this minimum standard is “reproducible research”, which requires that data sets and computer code be made available to others for verifying published results and conducting alternative analyses.

The need for publishing reproducible research is increasing for a number of reasons. Investigators are more frequently examining weak associations and complex interactions for which the data contain a low signal-to-noise ratio. New technologies allow scientists in all areas to compile complex high-dimensional databases. The ubiquity of powerful statistical and computing capabilities allows investigators to explore those databases and identify associations of potential interest. However, with the increase in data and computing power comes a greater potential for identifying spurious associations. In addition to these developments, recent reports of fraudulent research being published in the biomedical literature have highlighted the need for reproducibility in biomedical studies and have invited the attention of the major medical journals (Laine and others, 2007). Even without the presence of deliberate fraud, it should be noted that as analyses become more complicated, the possibility of inadvertant errors resulting in misleading findings looms large. In the examples of Baggerly and others (2005) and Coombes and others (2007), the errors discovered were not necessarily simple or obvious and the examination of the problem itself required...
Dimensions of Reproducibility

• **Data** ("D"): The analytic data from which the principal results were derived are made available on the journal’s Web site. The authors are responsible for ensuring that necessary permissions are obtained before the data are distributed.

• **Code** ("C"): Any computer code, software, or other computer instructions that were used to compute published results are provided. For software that is widely available from central repositories (e.g. CRAN, Statlib), a reference to where they can be obtained will suffice.

• **Reproducible** ("R"): An article is designated as reproducible if the AER succeeds in executing the code on the data provided and produces results matching those that the authors claim are reproducible. In reproducing these results, reasonable bounds for numerical tolerance will be considered.
What is Reproducible?

Table 3. Posterior medians and 95% credible intervals for the effects of PM$_{10}$ and NO$_2$ on respiratory hospital admissions. The results are shown on the relative risk scale for an increase of one standard deviation in pollution concentrations (1.7 $\mu$g m$^{-3}$ for PM$_{10}$ and 8 $\mu$g m$^{-3}$ for NO$_2$).

<table>
<thead>
<tr>
<th>Spatial model</th>
<th>PM$_{10}$</th>
<th>Spatial resolution</th>
<th>NO$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DZ</td>
<td>IG</td>
<td>DZ</td>
</tr>
<tr>
<td>Grampian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td>1.03 (0.95, 1.10)</td>
<td>1.05 (0.96, 1.13)</td>
<td>1.01 (0.89, 1.14)</td>
</tr>
<tr>
<td>Joint</td>
<td>1.00 (0.92, 1.09)</td>
<td>1.05 (0.97, 1.14)</td>
<td>1.04 (0.94, 1.15)</td>
</tr>
<tr>
<td>Conditional</td>
<td>1.03 (0.95, 1.12)</td>
<td>1.04 (0.97, 1.13)</td>
<td>1.04 (0.93, 1.17)</td>
</tr>
<tr>
<td>Tayside</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td>1.03 (0.92, 1.16)</td>
<td>1.06 (0.94, 1.20)</td>
<td>0.96 (0.84, 1.09)</td>
</tr>
<tr>
<td>Joint</td>
<td>1.04 (0.91, 1.21)</td>
<td>1.04 (0.90, 1.21)</td>
<td>0.97 (0.81, 1.16)</td>
</tr>
<tr>
<td>Conditional</td>
<td>1.04 (0.93, 1.18)</td>
<td>1.05 (0.92, 1.21)</td>
<td>1.01 (0.88, 1.14)</td>
</tr>
<tr>
<td>Lothian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td>1.06 (1.01, 1.11)</td>
<td>1.07 (1.01, 1.14)</td>
<td>1.04 (0.97, 1.12)</td>
</tr>
<tr>
<td>Joint</td>
<td>1.09 (1.01, 1.16)</td>
<td>1.09 (1.02, 1.16)</td>
<td>1.12 (1.02, 1.22)</td>
</tr>
<tr>
<td>Conditional</td>
<td>1.08 (1.02, 1.14)</td>
<td>1.07 (1.01, 1.15)</td>
<td>1.10 (1.01, 1.21)</td>
</tr>
<tr>
<td>Greater</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td>1.10 (1.06, 1.14)</td>
<td>1.08 (1.03, 1.14)</td>
<td>1.11 (1.07, 1.15)</td>
</tr>
<tr>
<td>Joint</td>
<td>1.09 (1.04, 1.15)</td>
<td>1.07 (1.02, 1.13)</td>
<td>1.10 (1.04, 1.15)</td>
</tr>
<tr>
<td>Conditional</td>
<td>1.07 (1.03, 1.11)</td>
<td>1.07 (1.01, 1.13)</td>
<td>1.08 (1.03, 1.12)</td>
</tr>
</tbody>
</table>

Second-order estimating equations for the analysis of clustered current status data

RICHARD J. COOK*, DAVID TOLUSSO
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1
rjcook@uwaterloo.ca

Air pollution and health in Scotland: a multicity study

DUNCAN LEE*, CLAIRE FERGUSON
Department of Statistics, University of Glasgow, Glasgow, G12 8QQ UK
duncan@stats.gla.ac.uk
RICHARD MITCHELL
Public Health and Health Policy, University of Glasgow, Glasgow, G12 8QQ UK
Air pollution and health in Scotland: a multicity study

Supplementary Data

Files in this Supplementary Material:

- Supplementary Data - Manuscript file of format pdf
- Supplementary Data - Manuscript file of format txt

Abstract

First published online: April 17, 2009, doi: 10.1093/biostatistics/kxp010
Biostatistics July 1, 2009 vol. 10 no. 3 409-423

Full Text (HTML)
Full Text (PDF)
Supplementary Data
Literate (Statistical) Programming

• An article is a stream of text and code
• Analysis code is divided into text and code “chunks”
• Each code chunk loads data and computes results
• Presentation code formats results (tables, figures, etc.)
• Article text explains what is going on
• Literate programs can be weaved to produce human-readable documents and tangled to produce machine-readable documents
Literate (Statistical) Programming

• Literate programming is a general concept that requires
  1. A documentation language (human readable)
  2. A programming language (machine readable)
• Sweave uses $\LaTeX$ and R as the documentation and programming languages
• Sweave was developed by Friedrich Leisch (member of the R Core) and is maintained by R core
• Main web site: http://www.statistik.lmu.de/~leisch/Sweave
• Alternatives to LATEX/R exist, such as HTML/R (package R2HTML) and ODF/R (package odfWeave).
Summary

• Reproducible research is important as a **minimum standard**, particularly for studies that are difficult to replicate

• Scientific culture needs to evolve to encourage greater **sharing** of datasets and methods

• Infrastructure is needed for **creating** and **distributing** reproducible documents, beyond what is currently available
Acknowledgments

• Joint work with
  – Duncan Temple Lang, Deb Nolan
  – Sandy Eckel
  – Scott Zeger, Peter Diggle
  – Francesca Dominici

• Funded by
  – NIEHS, NIA, Health Effects Institute
  – Johns Hopkins Faculty Innovation Fund