Computational and Policy Tools for Reproducible Research

Roger D. Peng, PhD

Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health

July 2011

Vancouver, BC

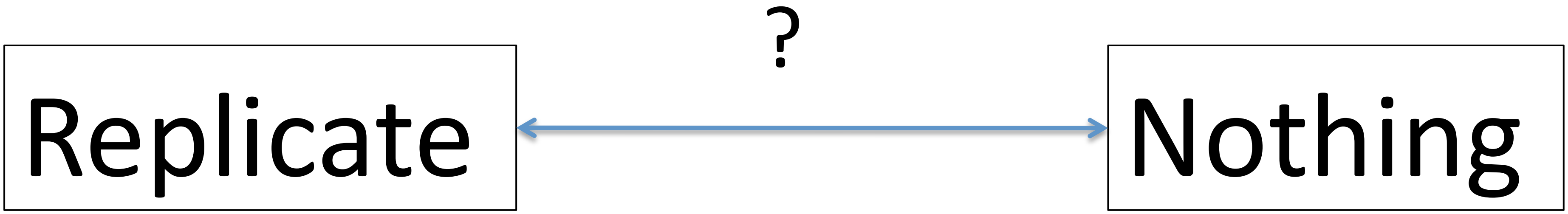
Replication

- 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?

- Some studies cannot be replicated
  - No time, opportunistic
  - No money
  - Unique
- 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"

How Can We Bridge the Gap?



Replicate

?

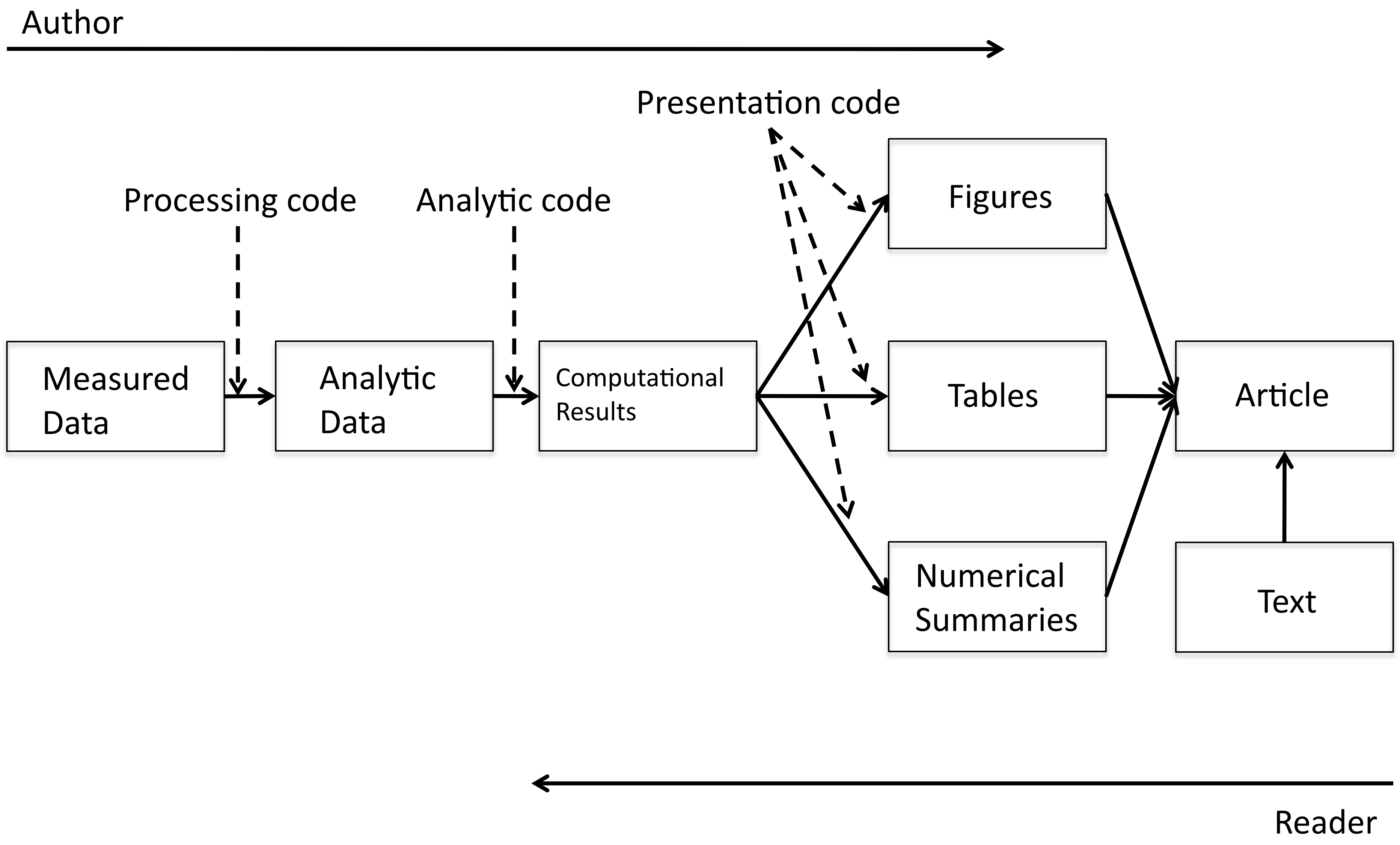
Nothing

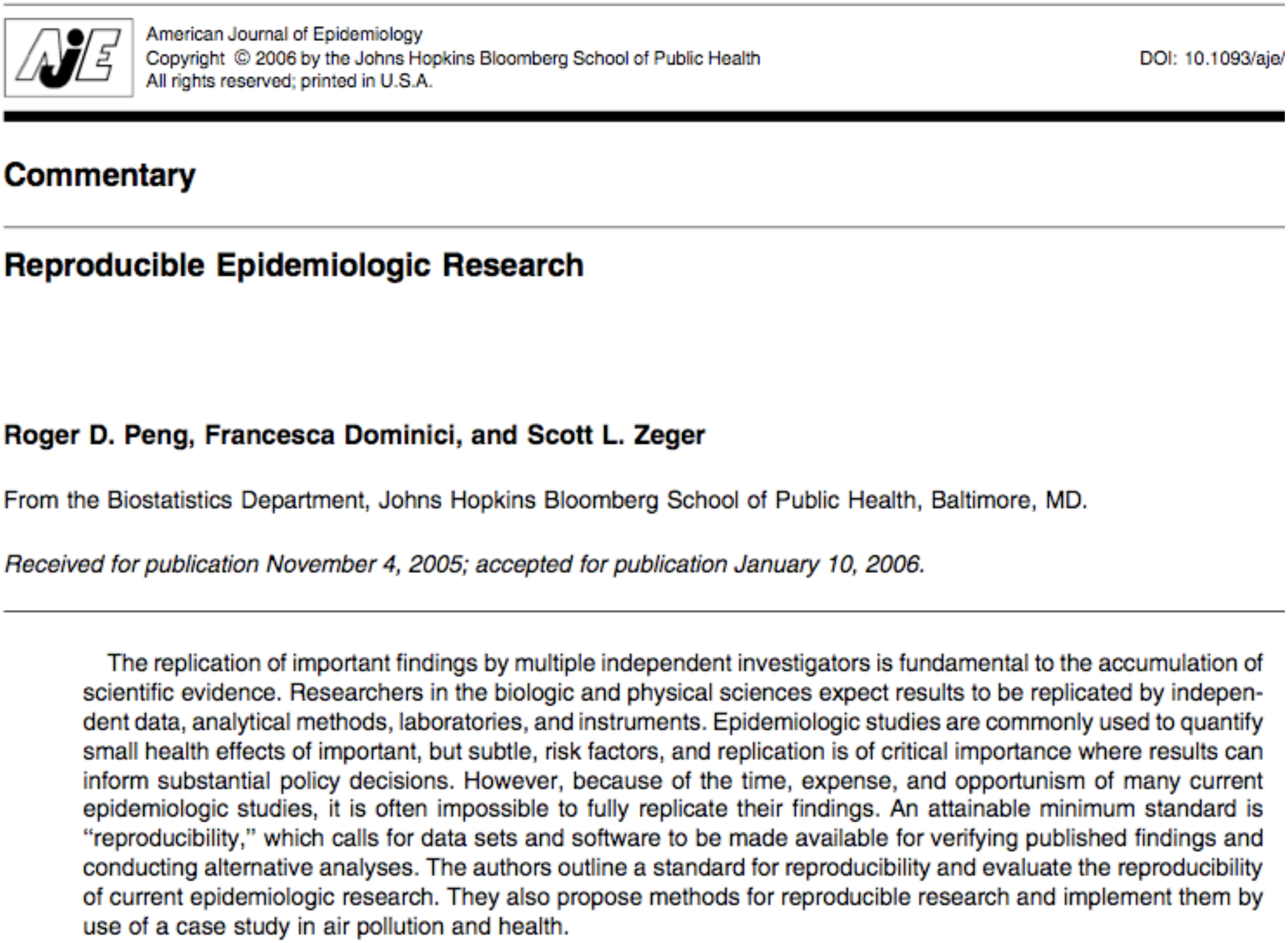
Research Pipeline

Article

Reader

Research Pipeline





American Journal of Epidemiology
Copyright © 2006 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved; printed in U.S.A.

DOI: 10.1093/aje/

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.

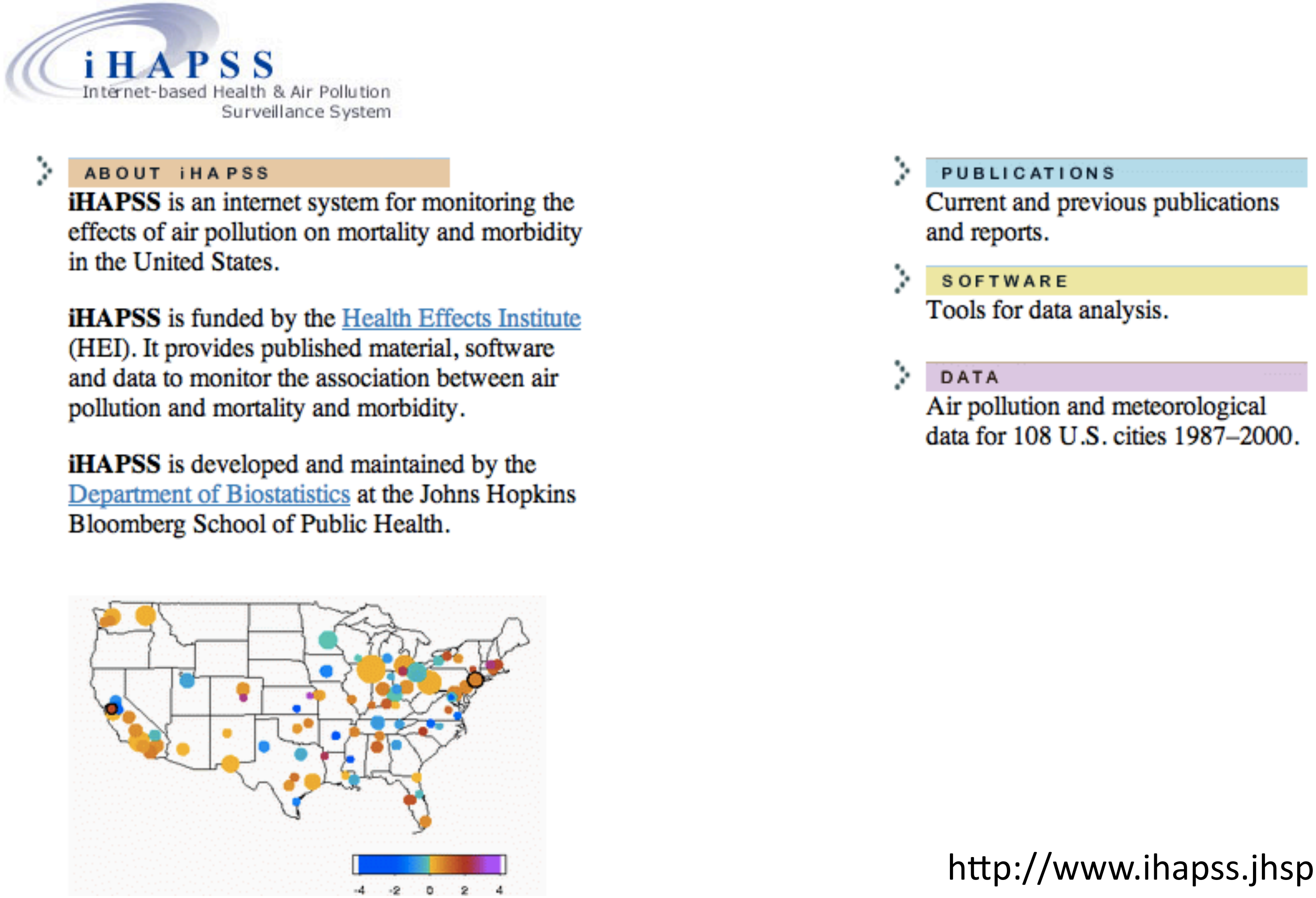
Received for publication November 4, 2005; accepted for publication January 10, 2006.

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

Internet-based Health and Air Pollution Surveillance System (iHAPSS)



iHAPSS
Internet-based Health & Air Pollution Surveillance System

ABOUT iHAPSS

iHAPSS is an internet system for monitoring the effects of air pollution on mortality and morbidity in the United States.

iHAPSS is funded by the Health Effects Institute (HEI). It provides published material, software and data to monitor the association between air pollution and mortality and morbidity.

iHAPSS is developed and maintained by the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health.

PUBLICATIONS

Current and previous publications and reports.

SOFTWARE

Tools for data analysis.

DATA

Air pollution and meteorological data for 108 U.S. cities 1987–2000.

http://www.ihapss.jhsph.edu

What is Reproducible Research?

- Analytic data are available
- Analytic code are available
- Documentation of code and data
- Standard means of distribution

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 (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.
- Readers may not have the same resources as authors

In 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
  - Piece together the software and run it

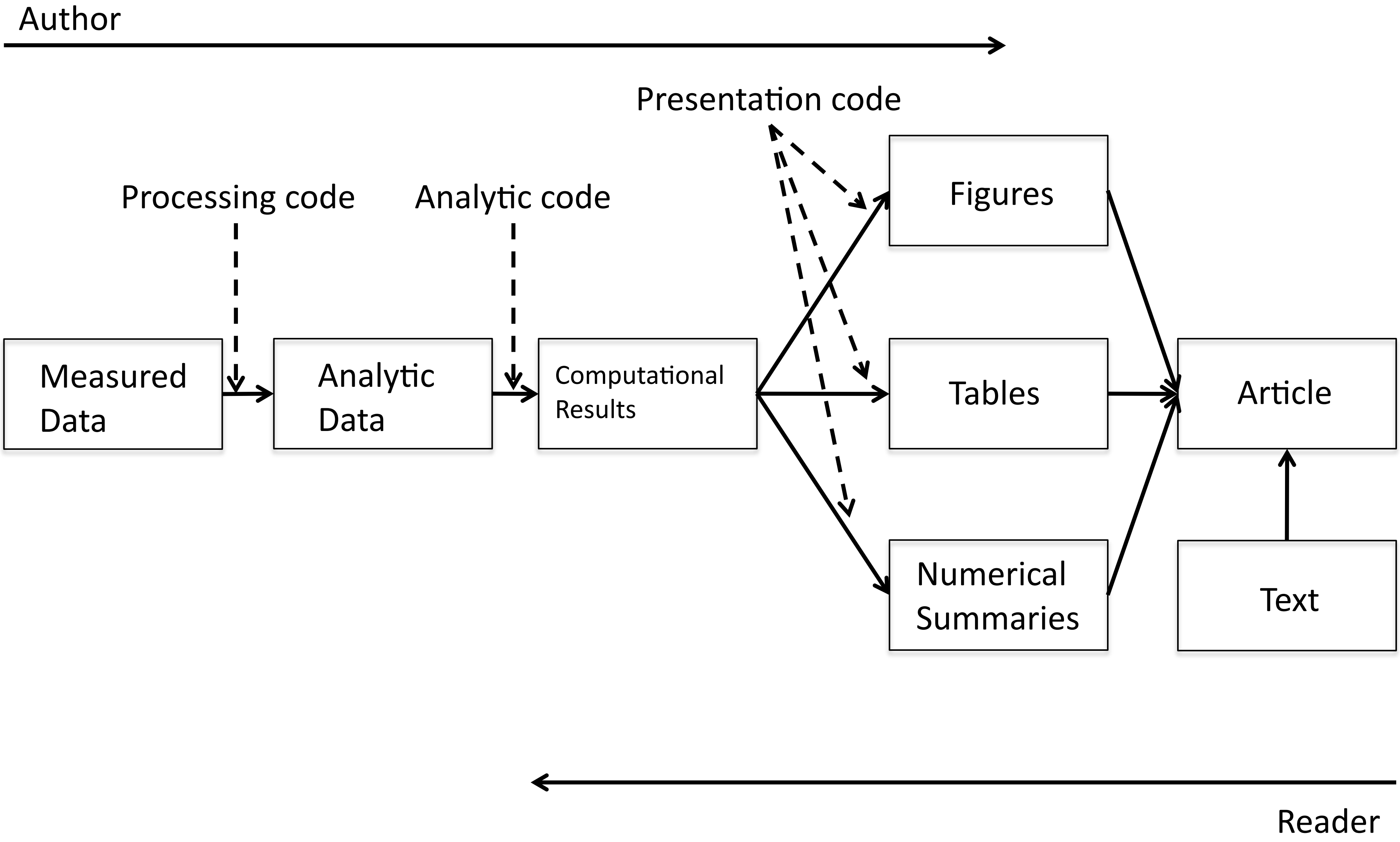
Literate (Statistical) Programming

- 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

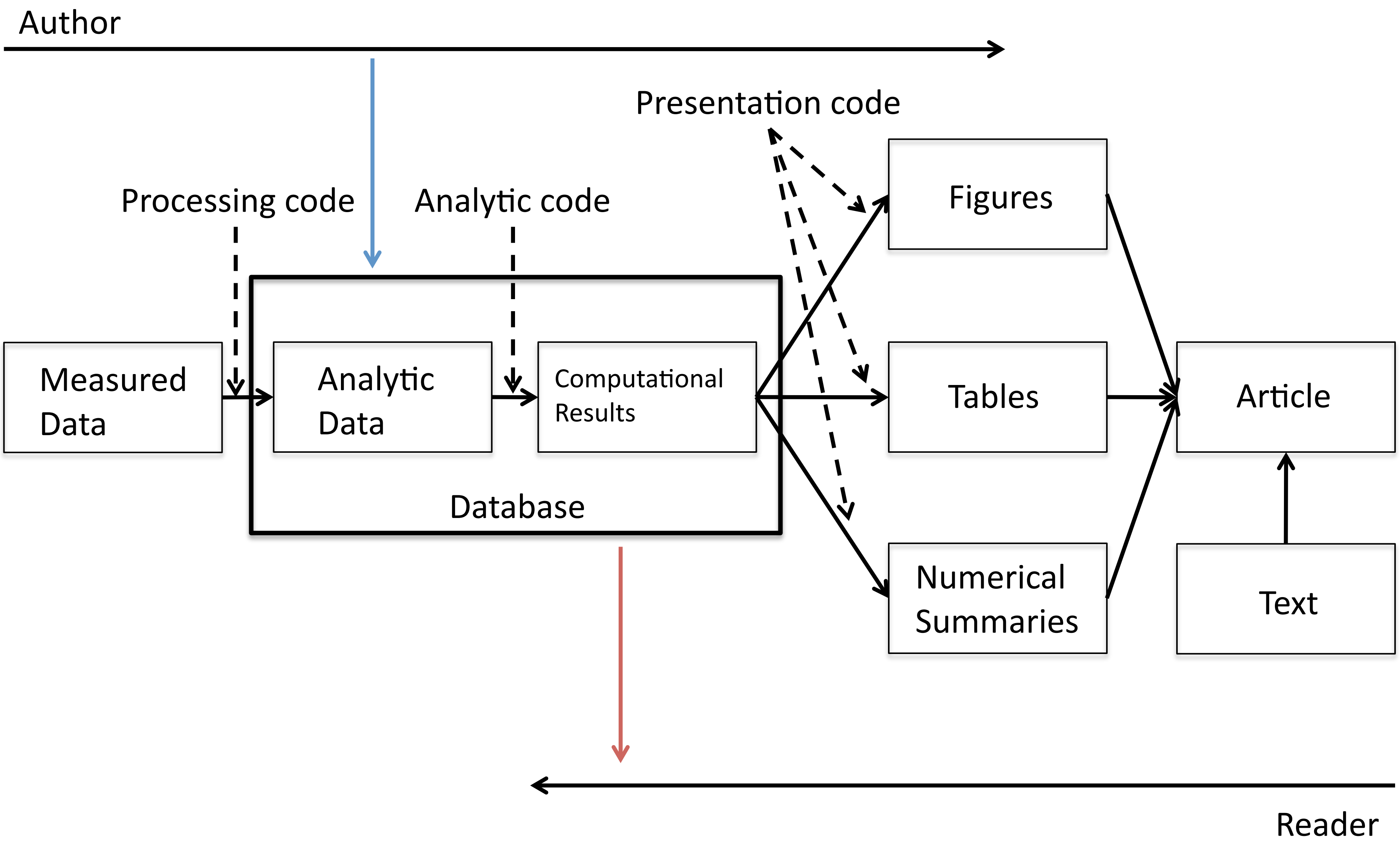
Literate (Statistical) Programming

- 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, suchas HTML/R (package R2HTML) and ODF/R (package odfWeave).

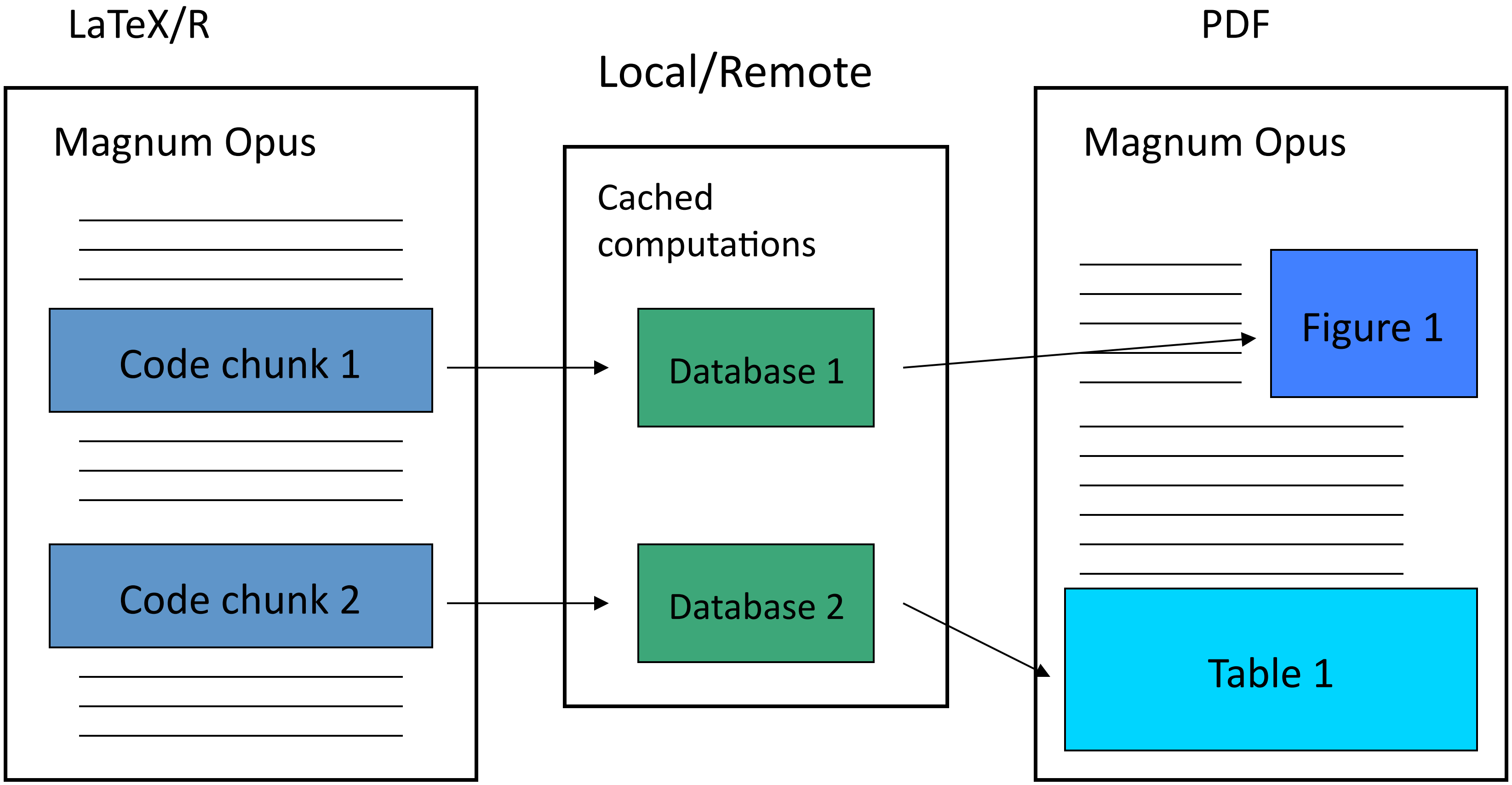
Research Pipeline



Research Pipeline



Caching Computations

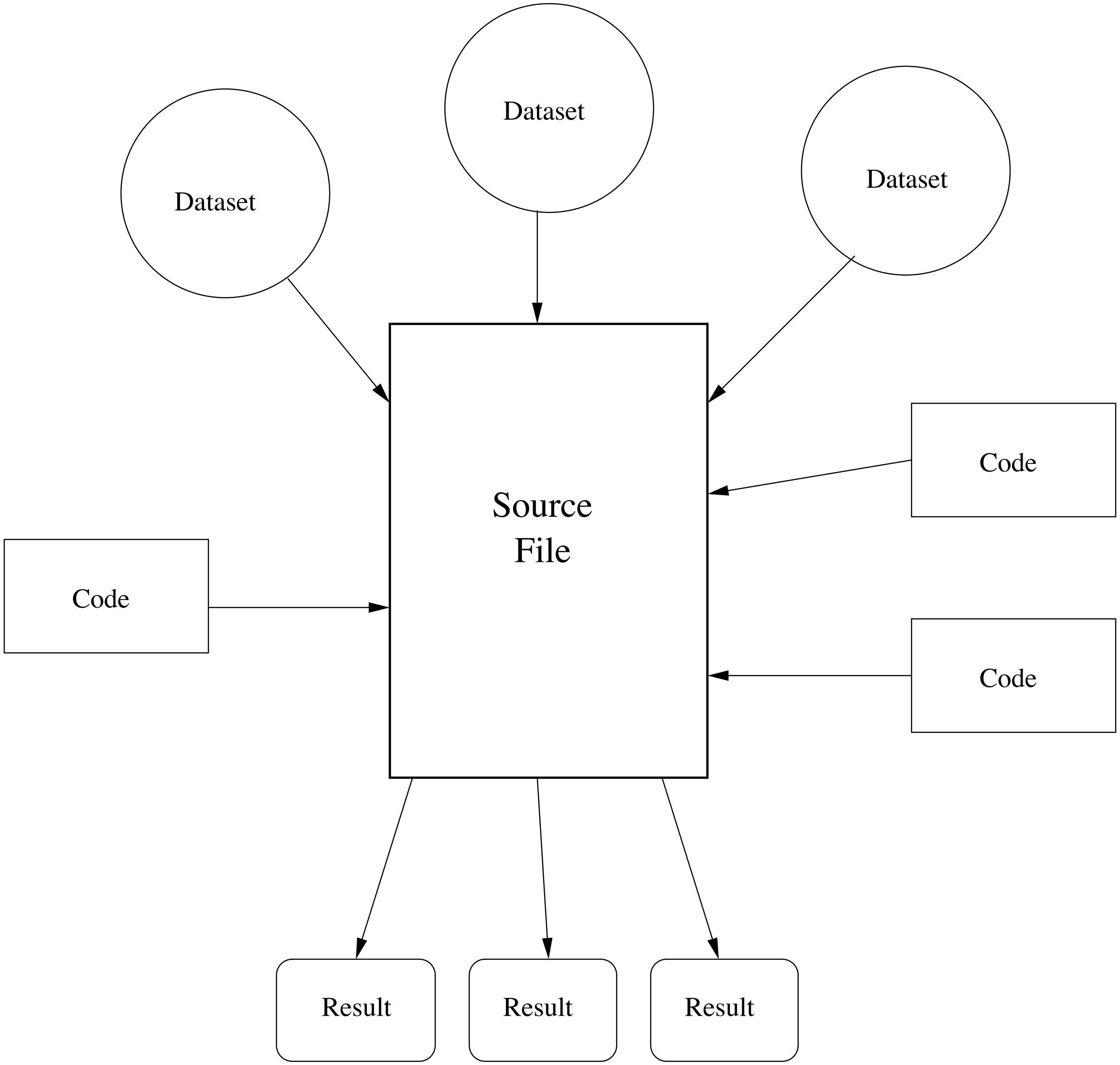


The `cacher` package for R

- Add-on package for R
- Evaluates code written in files and stores intermediate results in a key-value database
- R expressions are given SHA-1 hash values so that changes can be tracked and code reevaluated if necessary
- “Cacher packages” can be built for distribution
- Others can “clone” an analysis and evaluate subsets of code or inspect data objects

Journal of Statistical Software, 26 (7), 1—24

Conceptual Model



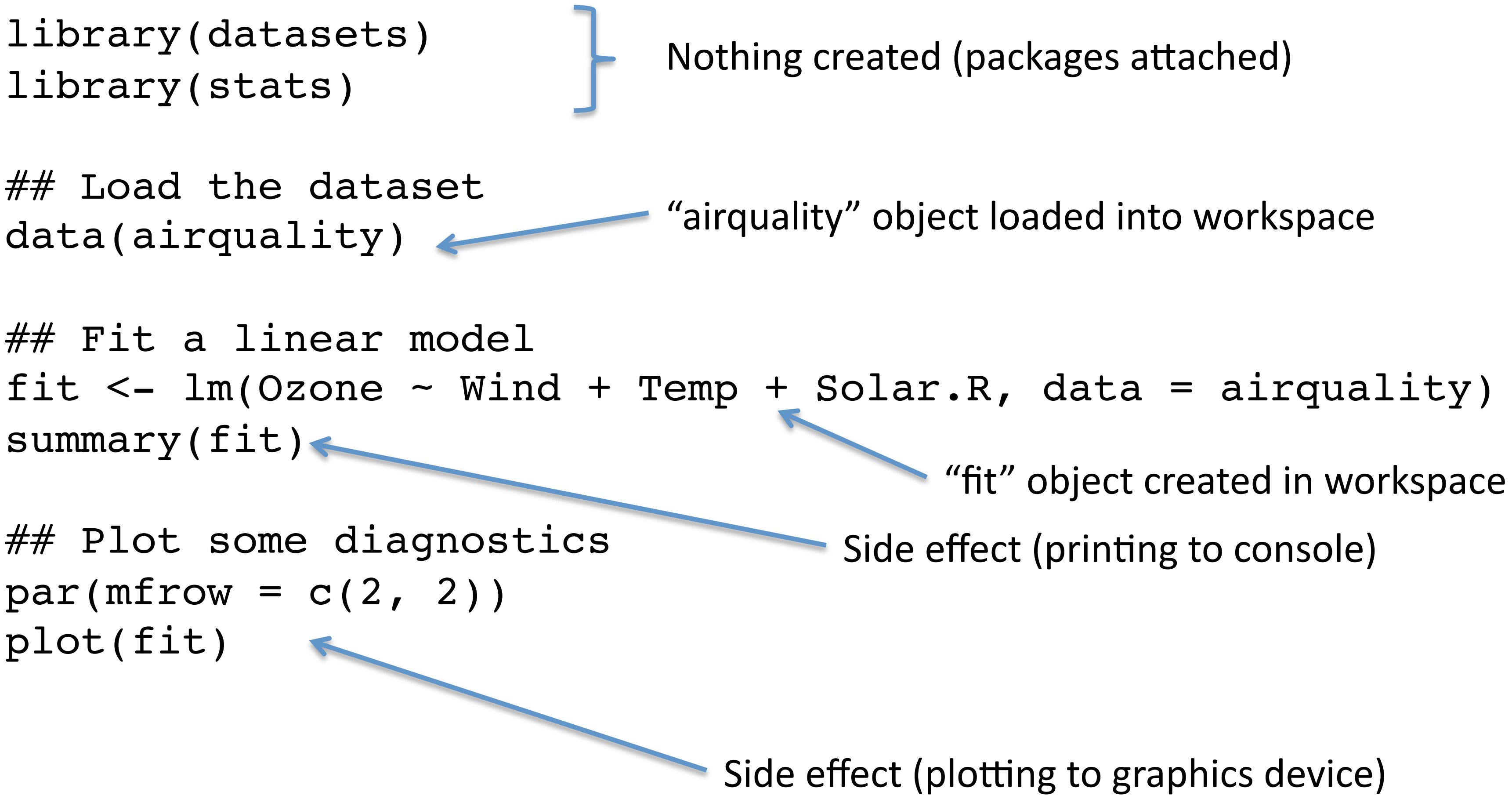
Using `cacher` as an Author

1. Parse the R source file; Create the necessary cache directories and subdirectories
2. Cycle through each expression in the source file:
   – If an expression has never been evaluated, evaluate it and store any resulting R objects in the cache database,
   – If a cached result exists, lazy-load the results from the cache database and move to the next expression,
   – If an expression does not create any R objects (i.e., there is nothing to cache), add the expression to the list of expressions where evaluation needs to be forced
   – Write out metadata for this expression to the metadata file.

Using `cacher` as an Author

- The `cachepackage` function creates a `cacher` package storing
  - Source file
  - Cached data objects
  - Metadata
- Package file is zipped and can be distributed
- Readers can unzip the file and immediately investigate its contents via `cacher` package

Example: Simple Analysis



```
library(datasets)
library(stats)

## Load the dataset
data(airquality)

## Fit a linear model
fit <- lm(Ozone ~ Wind + Temp + Solar.R, data = airquality)
summary(fit)

## Plot some diagnostics
par(mfrow = c(2, 2))
plot(fit)
```

Nothing created (packages attached)

“airquality” object loaded into workspace

“fit” object created in workspace

Side effect (printing to console)

Side effect (plotting to graphics device)

Using `cacher` as a Reader

A journal article says...

> “...the code and data for this analysis can be found in the cacher package 092dcc7dda4b93e42f23e038a60e1d44dbec7b3f.”

```
> library(cacher)
> clonecache(id = "092dcc7dda4b93e42f23e038a60e1d44dbec7b3f”)
> clonecache(id = “092d”)  ## Same as above
created cache directory '.cache'

> showfiles()
[1] "top20.R"
> sourcefile("top20.R")
```

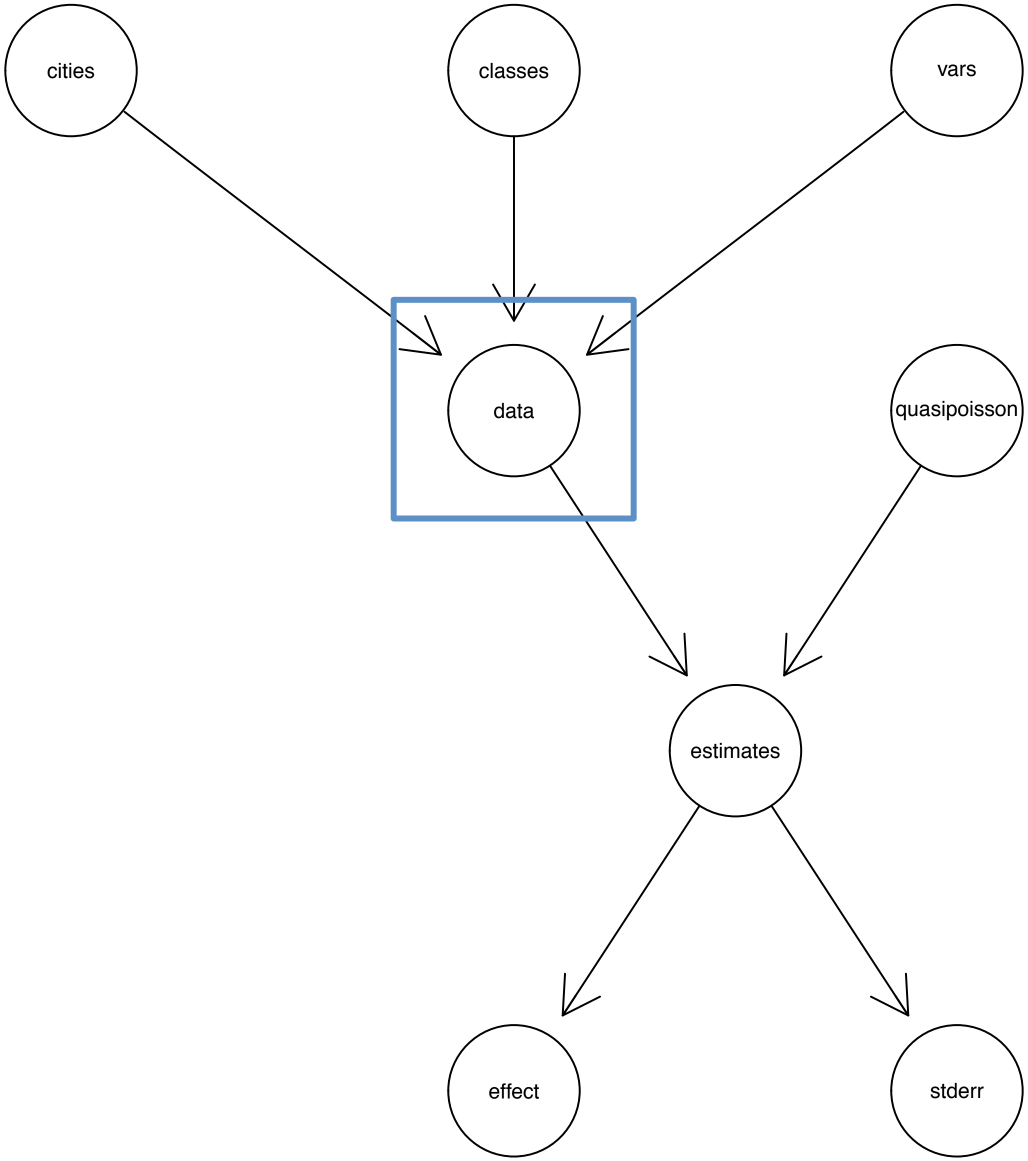
Cloning an Analysis

- Local directories created
- Source code files and metadata are downloaded
- Data objects are *not* downloaded by default
- References to data objects are loaded and corresponding data can be lazy-loaded on demand

Examining Code

```
> code()
source file: top20.R
1  cities <- readLines("citylist.txt")
2  classes <- readLines("colClasses.txt")
3  vars <- c("date", "dow", "death",
4  data <- lapply(cities, function(city) {
5  names(data) <- cities
6  estimates <- sapply(data, function(city) {
7  effect <- weighted.mean(estimates[1,
8  stderr <- sqrt(1/sum(1/estimates[2,

> graphcode()
```

Analysis Code Graphs



Tracing Code Backwards

```
> objectcode(“data”)
source file: top20.R
1  cities <- readLines("citylist.txt")
2  classes <- readLines("colClasses.txt")
3  vars <- c("date", "dow", "death", "tmpd", "rmtmpd", "dptp",
             "rmdptp", "l1pm10tmean")
4  data <- lapply(cities, function(city) {
           filename <- file.path("data", paste(city, "csv",
                                 sep = "."))
           d0 <- read.csv(filename, colClasses = classes,
                          nrow = 5200)
           d0[, vars]
   })
5  names(data) <- cities
```

Running Code

- The `runcode` function executes code in the source file
- By default, expressions that results in an object being created are *not* run and the resulting objects is lazy-loaded into the workspace
- Expressions not resulting in objects are evaluated

Checking Code and Objects

- The `checkcode` function evaluates all expressions from scratch (no lazy-loading)
- Results of evaluation are checked against stored results to see if the results are the same as what the author calculated
  - Setting RNG seeds is critical for this to work
- The integrity of data objects can be verified with the `checkobjects` function to check for possible corruption of data (i.e. in transit)

Inspecting Data Objects

```
> loadcache()

> ls()
[1] "cities"    "classes"   "data"      "effect"
[5] "estimates" "stderr"    "vars"

> cities
/ transferring cache db file b8fd490bcf1d48cd06...
 [1] "la"   "ny"   "chic" "dlft" "hous" "phoe"
 [7] "staa" "sand" "miam" "det"  "seat" "sanb"
[13] "sanj" "minn" "rive" "phil" "atla" "oakl"
[19] "denv" "clev"
```

Inspecting Data Objects

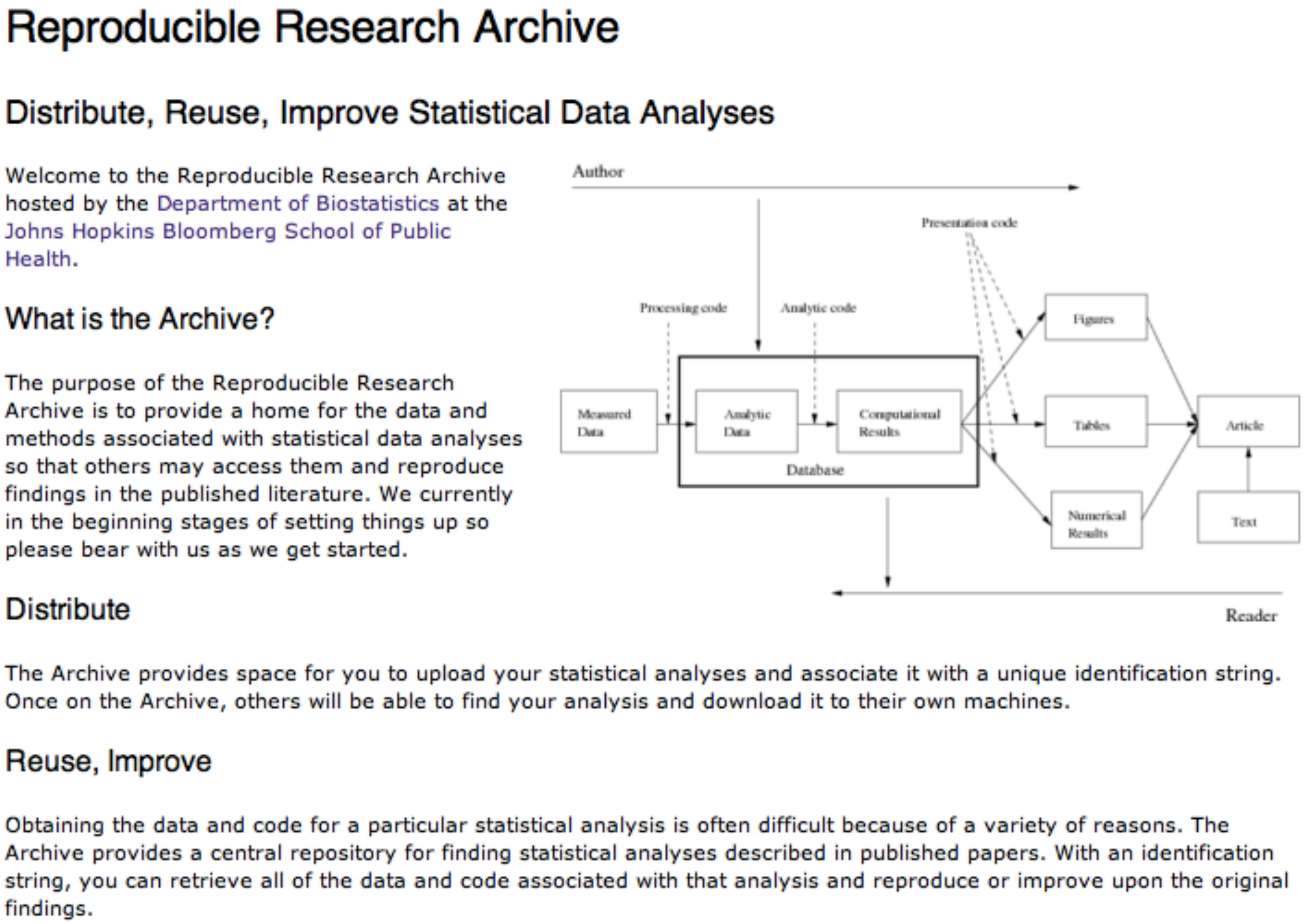
```
> effect
/ transferring cache db file 584115c69e5e2a4ae5...
[1] 0.0002313219

> stderr
/ transferring cache db file 81b6dc23736f3d72c6...
[1] 0.000052457
```

A 10 unit increase in PM_{10} is associated with a 0.23% increase in daily mortality

`cacher` Summary

- The `cacher` package can be used by authors to create cache packages from data analyses for distribution
- Readers can use the `cacher` package to inspect others' data analyses by examining cached computations
- `cacher` is mindful of readers' resources and efficiently loads only those data objects that are needed

A Central Archive for Reproducible Data Analyses



Reproducible Research Archive

Distribute, Reuse, Improve Statistical Data Analyses

Welcome to the Reproducible Research Archive hosted by the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health.

What is the Archive?

The purpose of the Reproducible Research Archive is to provide a home for the data and methods associated with statistical data analyses so that others may access them and reproduce findings in the published literature. We currently in the beginning stages of setting things up so please bear with us as we get started.

Distribute

The Archive provides space for you to upload your statistical analyses and associate it with a unique identification string. Once on the Archive, others will be able to find your analysis and download it to their own machines.

Reuse, Improve

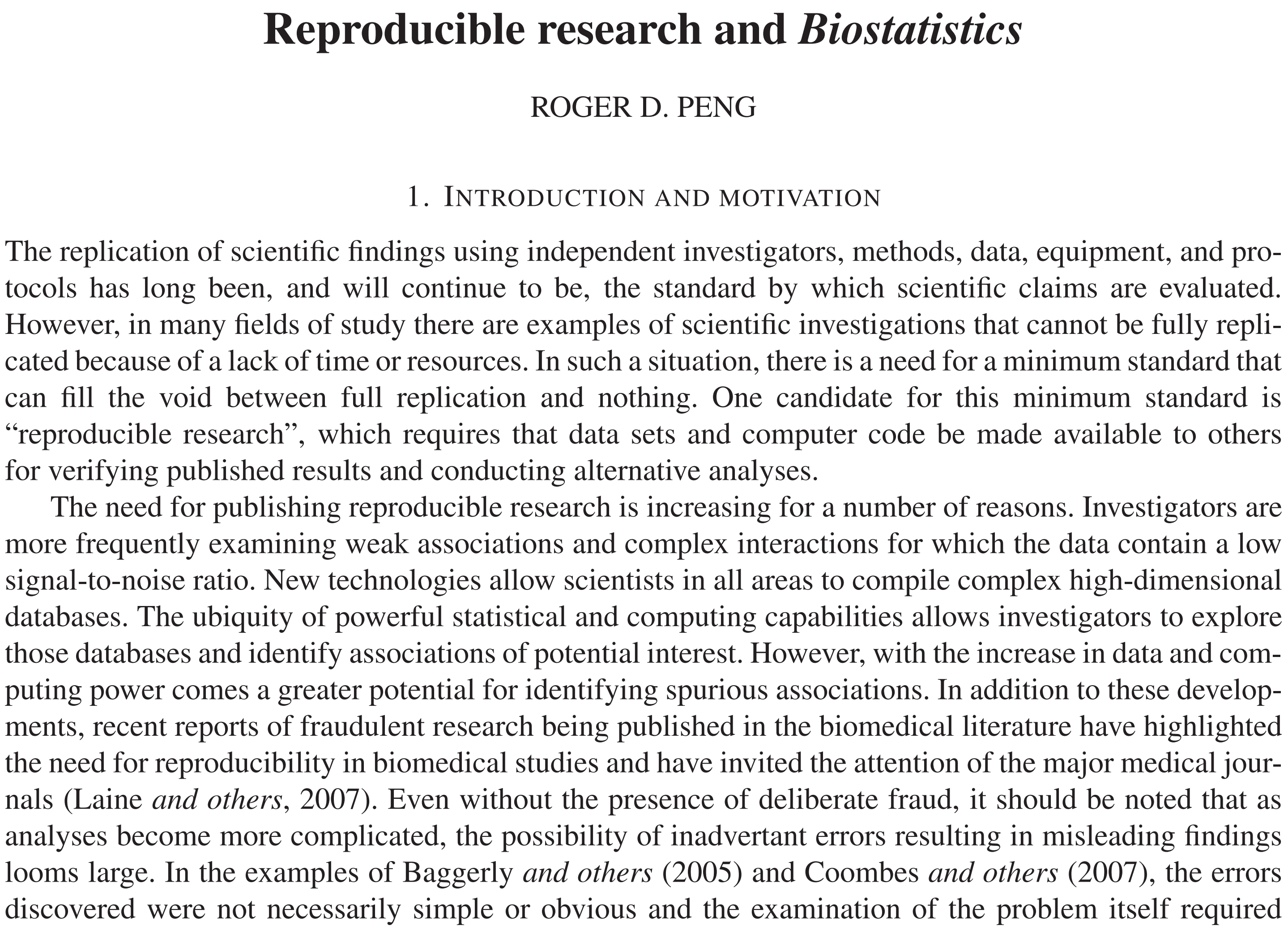
Obtaining the data and code for a particular statistical analysis is often difficult because of a variety of reasons. The Archive provides a central repository for finding statistical analyses described in published papers. With an identification string, you can retrieve all of the data and code associated with that analysis and reproduce or improve upon the original findings.

http://penguin.biostat.jhsph.edu/

Reproducible Research and Journals

- What policies can journals implement to make published research reproducible?
- Carrot or stick?

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

Biostatistics (2009), 10, 3, pp. 405–408

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.

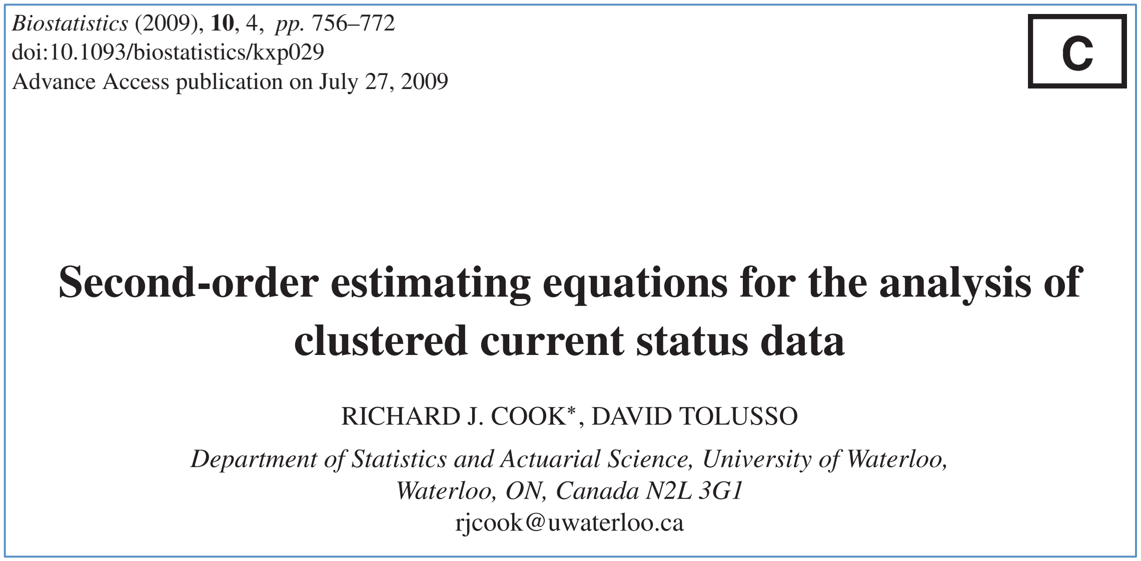
Dimensions of Reproducibility

- **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.

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.

Kite Marking



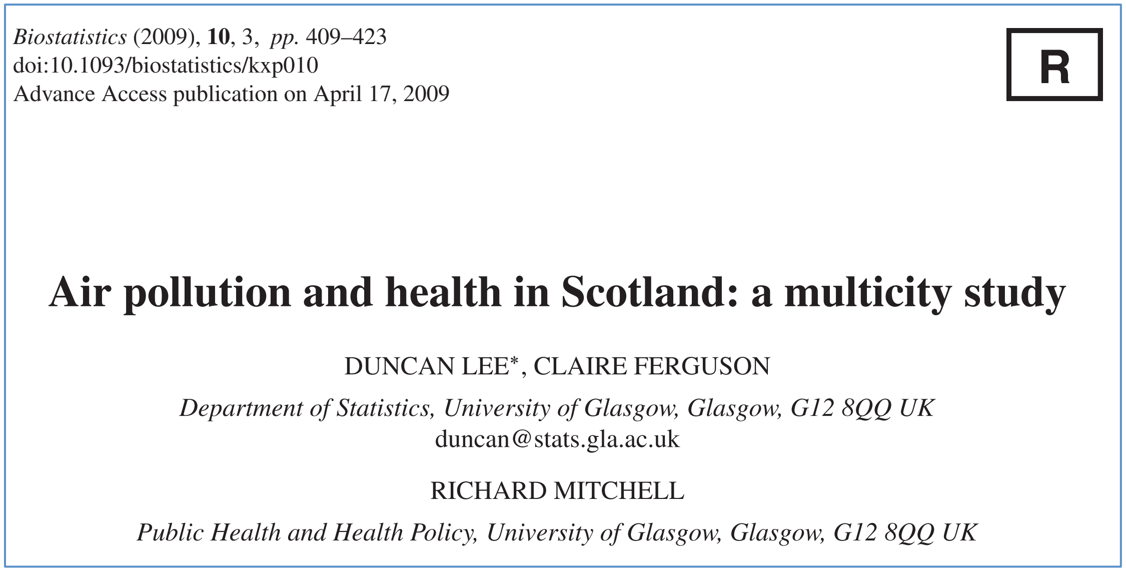
Biostatistics (2009), **10**, 4, *pp.* 756–772
doi:10.1093/biostatistics/kxp029
Advance Access publication on July 27, 2009

C

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



Biostatistics (2009), **10**, 3, *pp.* 409–423
doi:10.1093/biostatistics/kxp010
Advance Access publication on April 17, 2009

R

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

What is Reproducible?

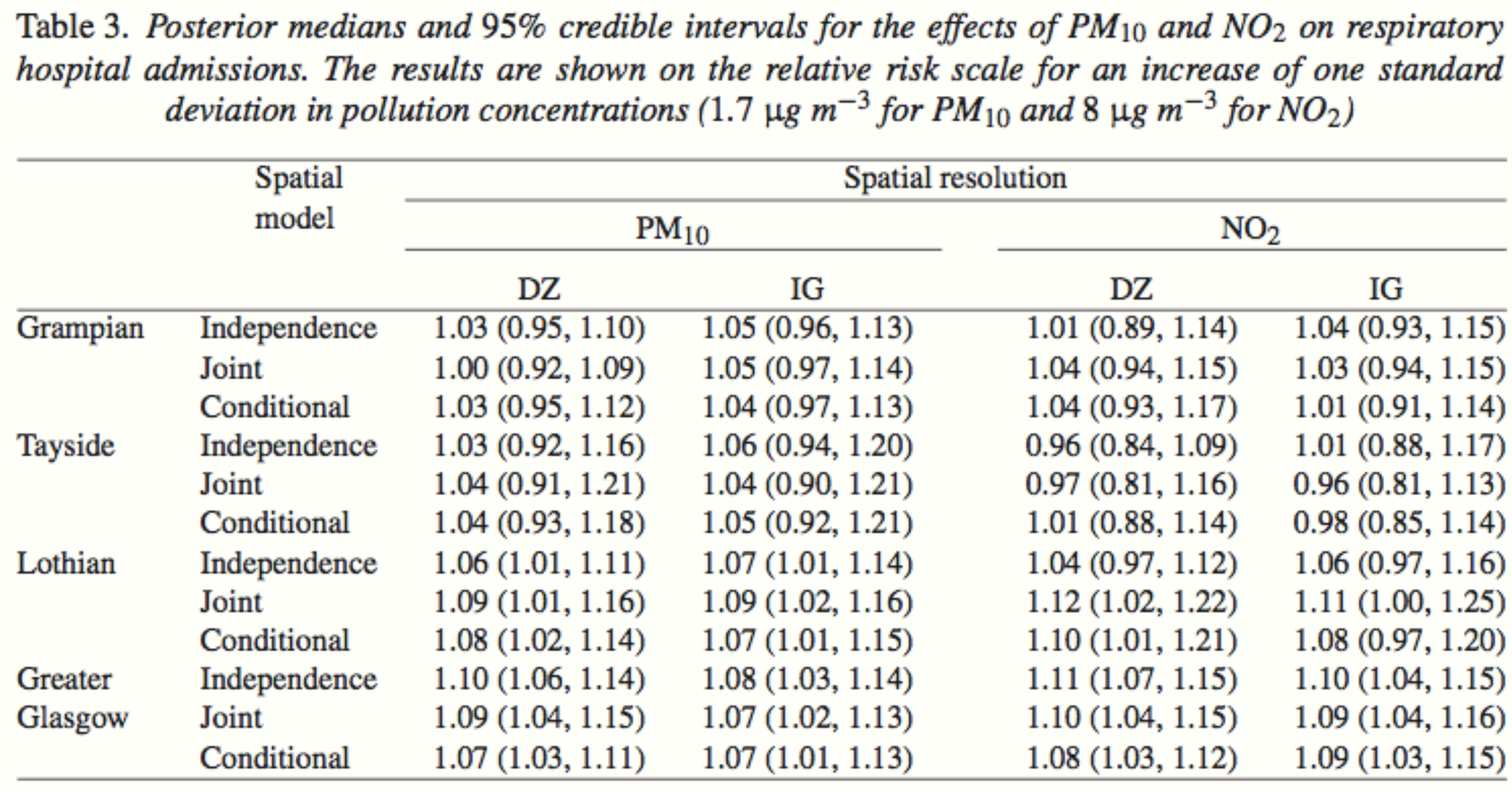
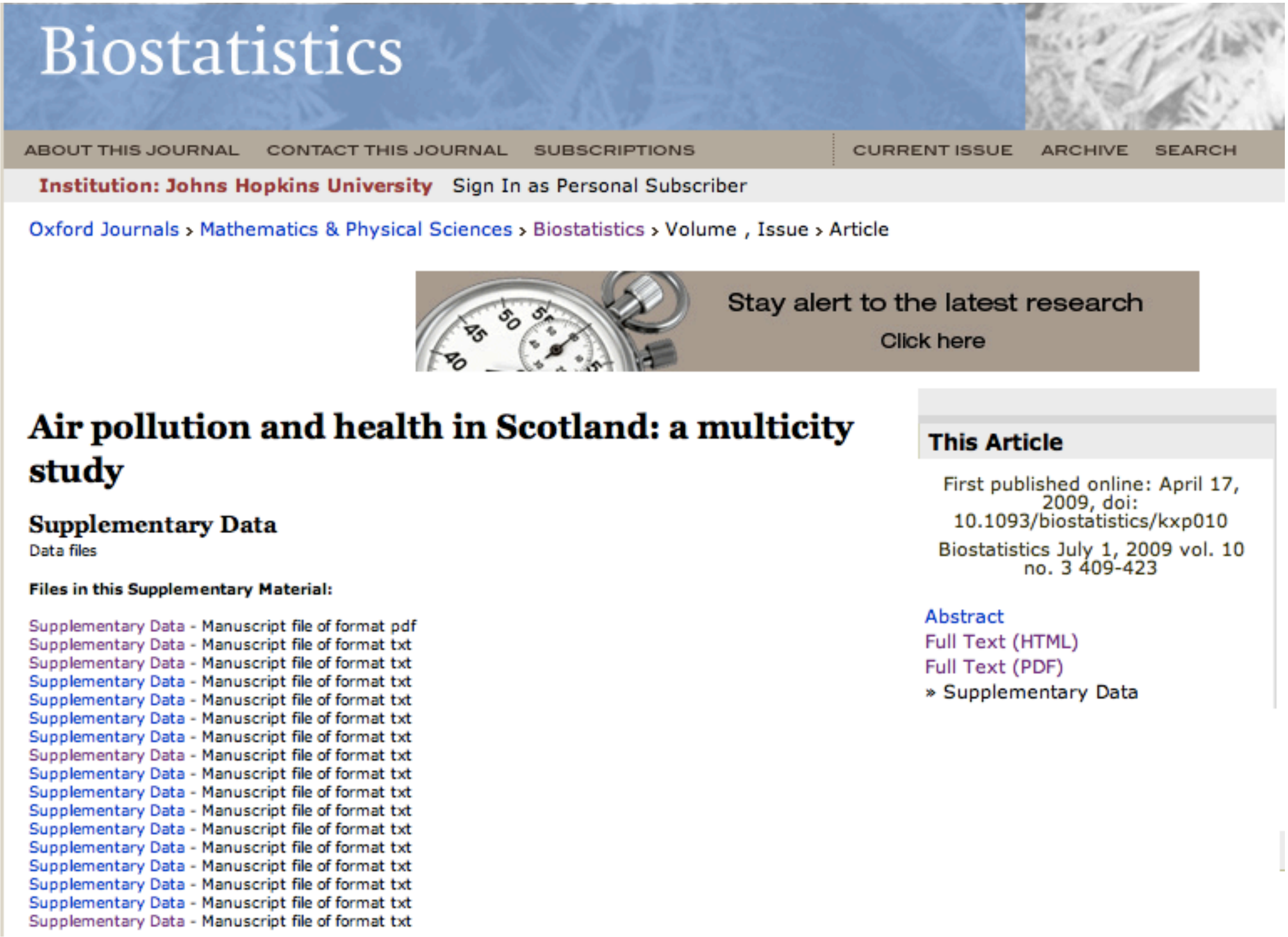


Table 3. *Posterior medians and 95% credible intervals for the effects of PM_{10} and NO_2 on respiratory hospital admissions. The results are shown on the relative risk scale for an increase of one standard deviation in pollution concentrations* (1.7 $\mu g\ m^{-3}$ *for* PM_{10} *and* 8 $\mu g\ m^{-3}$ *for* NO_2*)*

	Spatial model	Spatial resolution			
		PM_{10}		NO_2	
		DZ	IG	DZ	IG
Grampian	Independence	1.03 (0.95, 1.10)	1.05 (0.96, 1.13)	1.01 (0.89, 1.14)	1.04 (0.93, 1.15)
	Joint	1.00 (0.92, 1.09)	1.05 (0.97, 1.14)	1.04 (0.94, 1.15)	1.03 (0.94, 1.15)
	Conditional	1.03 (0.95, 1.12)	1.04 (0.97, 1.13)	1.04 (0.93, 1.17)	1.01 (0.91, 1.14)
Tayside	Independence	1.03 (0.92, 1.16)	1.06 (0.94, 1.20)	0.96 (0.84, 1.09)	1.01 (0.88, 1.17)
	Joint	1.04 (0.91, 1.21)	1.04 (0.90, 1.21)	0.97 (0.81, 1.16)	0.96 (0.81, 1.13)
	Conditional	1.04 (0.93, 1.18)	1.05 (0.92, 1.21)	1.01 (0.88, 1.14)	0.98 (0.85, 1.14)
Lothian	Independence	1.06 (1.01, 1.11)	1.07 (1.01, 1.14)	1.04 (0.97, 1.12)	1.06 (0.97, 1.16)
	Joint	1.09 (1.01, 1.16)	1.09 (1.02, 1.16)	1.12 (1.02, 1.22)	1.11 (1.00, 1.25)
	Conditional	1.08 (1.02, 1.14)	1.07 (1.01, 1.15)	1.10 (1.01, 1.21)	1.08 (0.97, 1.20)
Greater Glasgow	Independence	1.10 (1.06, 1.14)	1.08 (1.03, 1.14)	1.11 (1.07, 1.15)	1.10 (1.04, 1.15)
	Joint	1.09 (1.04, 1.15)	1.07 (1.02, 1.13)	1.10 (1.04, 1.15)	1.09 (1.04, 1.16)
	Conditional	1.07 (1.03, 1.11)	1.07 (1.01, 1.13)	1.08 (1.03, 1.12)	1.09 (1.03, 1.15)

Lee, Ferguson & Mitchell, *Biostatistics*, 2009

Supplementary Data (not ideal)



Biostatistics

Institution: Johns Hopkins University Sign In as Personal Subscriber

Oxford Journals › Mathematics & Physical Sciences › Biostatistics › Volume , Issue › Article

Air pollution and health in Scotland: a multicity study

Supplementary Data

Data files

Files in this Supplementary Material:

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

This Article

First published online: April 17, 2009, doi: 10.1093/biostatistics/kxp010

Biostatistics July 1, 2009 vol. 10 no. 3 409-423

Abstract
Full Text (HTML)
Full Text (PDF)
» Supplementary Data

Some Sparse Data

Data so far (a little old…)

- 4 papers have requested and received the “R” kite mark
- 4 papers received a “C”
- 2 papers received a “D”
- 1 paper with “DC”

Further Work

- Need a better system at journal for tracking and highlighting papers with kite-marks
- Infrastructure for hosting data is limited
- Infrastructure for reproducing results is limited
- Need better *advertising* of this policy

Summary

- Reproducible research is important as a **minimum standard**, particularly for studies that are difficult to replicate
- Infrastructure is needed for **creating** and **distributing** reproducible documents, beyond what is currently available
- The `cacher` package caches intermediate computations for future inspection
- Scientific culture needs to evolve to encourage greater **sharing** of datasets and methods
- Journals can play a key role by providing both carrots and sticks to authors

Acknowledgments

- Joint work with
  - Duncan Temple Lang (UC Davis)
  - Deb Nolan (Berkeley)
  - Sandy Eckel (USC)
- Funded by
  - National Institute of Environmental Health Science
  - National Institute on Aging
  - Johns Hopkins Faculty Innovation Fund
  - Health Effects Institute