Communicating Results

Specifying Levels of Detail

Roger D. Peng, Associate Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

tl;dr

- People are busy, especially managers and leaders
- Results of data analyses are sometimes presented in oral form, but often the first cut is presented via email
- It is often useful to breakdown the results of an analysis into different levels of granularity / detail
- Getting responses from busy people: http://goo.gl/sJDb9V

Hierarchy of Information: Research Paper

- Title / Author list
- Abstract
- Body / Results
- Supplementary Materials / the gory details
- Code / Data / really gory details

Hierarchy of Information: Email Presentation

- Subject line / Sender info
  - At a minimum; include one
  - Can you summarize findings in one sentence?
- Email body
  - A brief description of the problem / context; recall what was proposed and executed; summarize findings / results; 1–2 paragraphs
  - If action needs to be taken as a result of this presentation, suggest some options and make them as concrete as possible.
  - If questions need to be addressed, try to make them yes / no

Hierarchy of Information: Email Presentation

- Attachment(s)
  - R Markdown file
  - knitr report
  - Stay concise; don't spit out pages of code (because you used knitr we know it's available)
- Links to Supplementary Materials
  - Code / Software / Data
  - GitHub repository / Project web site

Reproducible Research Checklist

What to Do and What Not to Do

Roger D. Peng, Associate Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

DO: Start With Good Science

- Garbage in, garbage out
- Coherent, focused question simplifies many problems
- Working with good collaborators reinforces good practices
- Something that's interesting to you will (hopefully) motivate good habits

DON'T: Do Things By Hand

- Editing spreadsheets of data to "clean it up"
    - Removing outliers
    - QA / QC
    - Validating
- Editing tables or figures (e.g. rounding, formatting)
- Downloading data from a web site (clicking links in a web browser)
- Moving data around your computer; splitting / reformatting data files
- "We're just going to do this once...."

Things done by hand need to be precisely documented (this is harder than it sounds)

DON'T: Point And Click

- Many data processing / statistical analysis packages have graphical user interfaces (GUIs)
- GUIs are convenient / intuitive but the actions you take with a GUI can be difficult for others to reproduce
- Some GUIs produce a log file or script which includes equivalent commands; these can be saved for later examination
- In general, be careful with data analysis software that is highly *interactive*; ease of use can sometimes lead to non-reproducible analyses
- Other interactive software, such as text editors, are usually fine

DO: Teach a Computer

- If something needs to be done as part of your analysis / investigation, try to teach your computer to do it (even if you only need to do it once)
- In order to give your computer instructions, you need to write down exactly what you mean to do and how it should be done
- Teaching a computer almost guarantees reproducibilty

For example, by hand, you can

1. Go to the UCI Machine Learning Repository at http://archive.ics.uci.edu/ml/
2. Download the Bike Sharing Dataset by clicking on the link to the Data Folder, then clicking on the link to the zip file of dataset, and choosing "Save Linked File As..." and then saving it to a folder on your computer

DO: Teach a Computer

Or You can teach your computer to do the same thing using R:

```
download.file("http://archive.ics.uci.edu/ml/machine-learning-databases/00275/
              Bike-Sharing-Dataset.zip", "ProjectData/Bike-Sharing-Dataset.zip")
```

Notice here that

- The full URL to the dataset file is specified (no clicking through a series of links)
- The name of the file saved to your local computer is specified
- The directory in which the file was saved is specified ("ProjectData")
- Code can always be executed in R (as long as link is available)

DO: Use Some Version Control

- Slow things down
- Add changes in small chunks (don't just do one massive commit)
- Track / tag snapshots; revert to old versions
- Software like GitHub / BitBucket / SourceForge make it easy to publish results

DO: Keep Track of Your Software Environment

- If you work on a complex project involving many tools / datasets, the software and computing environment can be critical for reproducing your analysis
- **Computer architecture**: CPU (Intel, AMD, ARM), GPUs,
- **Operating system**: Windows, Mac OS, Linux / Unix
- **Software toolchain**: Compilers, interpreters, command shell, programming languages (C, Perl, Python, etc.), database backends, data analysis software
- **Supporting software / infrastructure**: Libraries, R packages, dependencies
- **External dependencies**: Web sites, data repositories, remote databases, software repositories
- **Version numbers**: Ideally, for everything (if available)

DO: Keep Track of Your Software Environment

```
sessionInfo()
```

```
## R version 3.0.2 Patched (2014-01-20 r64849)
## Platform: x86_64-apple-darwin13.0.0 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  base
##
## other attached packages:
## [1] slidify_0.3.3
##
## loaded via a namespace (and not attached):
## [1] evaluate_0.5.1 formatR_0.10   knitr_1.5      markdown_0.6.3
## [5] stringr_0.6.2  tools_3.0.2    whisker_0.3-2  yaml_2.1.8
```

DON'T: Save Output

- Avoid saving data analysis output (tables, figures, summaries, processed data, etc.), except perhaps temporarily for efficiency purposes.
- If a stray output file cannot be easily connected with the means by which it was created, then it is not reproducible.
- Save the data + code that generated the output, rather than the output itself
- Intermediate files are okay as long as there is clear documentation of how they were created

DO: Set Your Seed

- Random number generators generate pseudo-random numbers based on an initial seed (usually a number or set of numbers)
    - In R you can use the `set.seed()` function to set the seed and to specify the random number generator to use
- Setting the seed allows for the stream of random numbers to be exactly reproducible
- Whenever you generate random numbers for a non-trivial purpose, **always set the seed**

DO: Think About the Entire Pipeline

- Data analysis is a lengthy process; it is not just tables / figures / reports
- Raw data → processed data → analysis → report
- How you got the end is just as important as the end itself
- The more of the data analysis pipeline you can make reproducible, the better for everyone

Summary: Checklist

- Are we doing good science?
- Was any part of this analysis done by hand?
  - If so, are those parts *precisely* document?
  - Does the documentation match reality?
- Have we taught a computer to do as much as possible (i.e. coded)?
- Are we using a version control system?
- Have we documented our software environment?
- Have we saved any output that we cannot reconstruct from original data + code?
- How far back in the analysis pipeline can we go before our results are no longer (automatically) reproducible?

Reproducible Research with Evidence-based Data Analysis

Roger D. Peng, Associate Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Replication and Reproducibility

Replication

- Focuses on the validity of the scientific claim
- "Is this claim true?"
- The ultimate standard for strengthening scientific evidence
- New investigators, data, analytical methods, laboratories, instruments, etc.
- Particularly important in studies that can impact broad policy or regulatory decisions

Replication and Reproducibility

Reproducibility

- Focuses on the validity of the data analysis
- "Can we trust this analysis?"
- Arguably a minimum standard for any scientific study
- New investigators, same data, same methods
- Important when replication is impossible

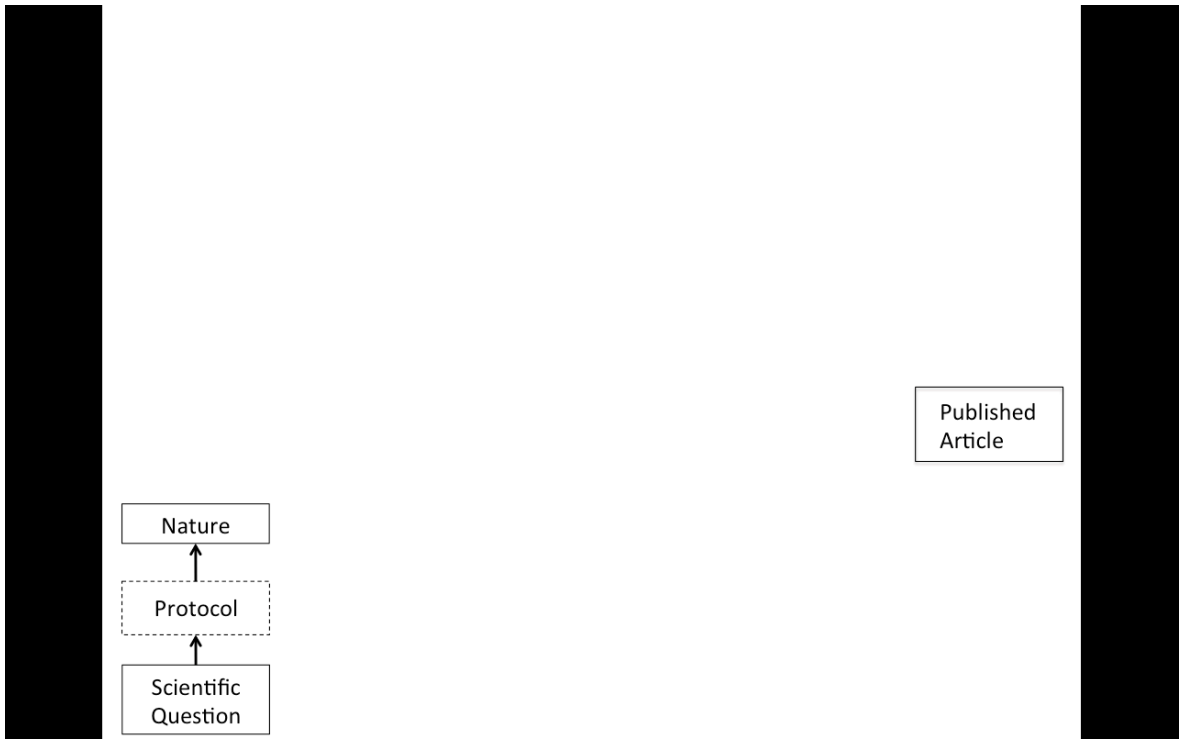
Background and Underlying Trends

- Some studies cannot be replicated: No time, No money, Unique/opportunistic
- Technology is increasing data collection throughput; data are more complex and high-dimensional
- Existing databases can be merged to become bigger databases (but data are used off-label)
- Computing power allows more sophisticated analyses, even on "small" data
- For every field "X" there is a "Computational X"

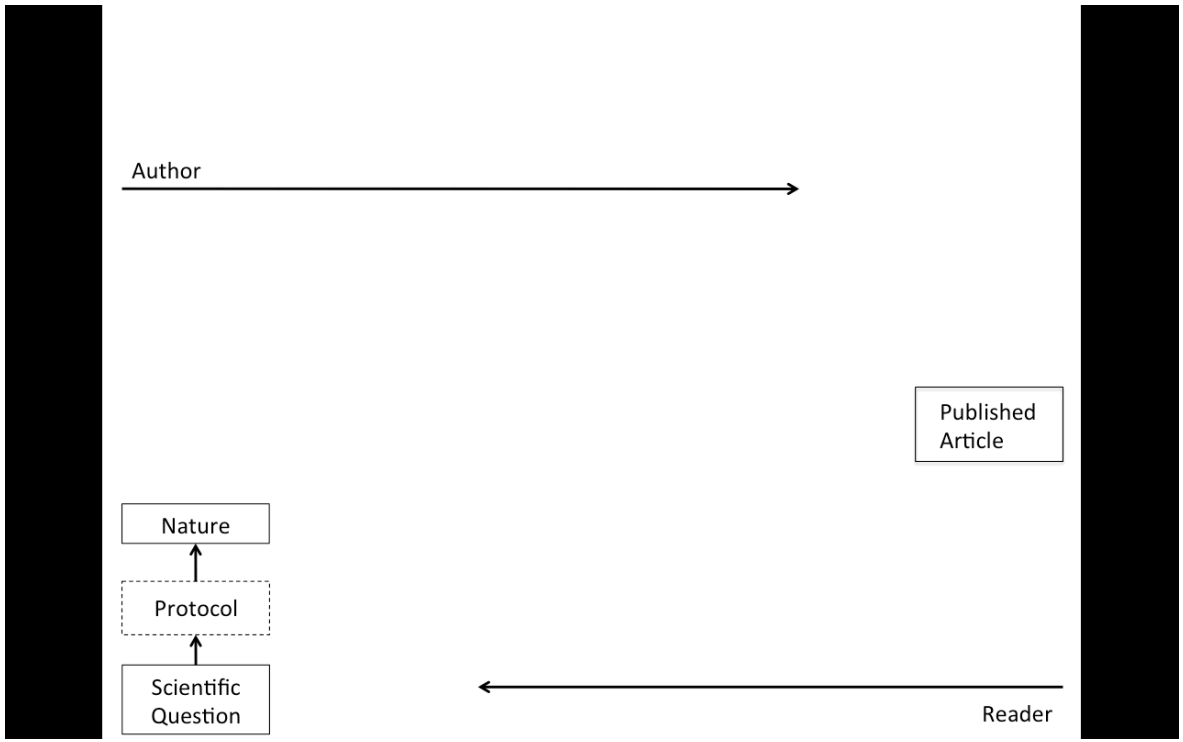
The Result?

- Even basic analyses are difficult to describe
- Heavy computational requirements are thrust upon people without adequate training in statistics and computing
- Errors are more easily introduced into long analysis pipelines
- Knowledge transfer is inhibited
- Results are difficult to replicate or reproduce
- Complicated analyses cannot be trusted

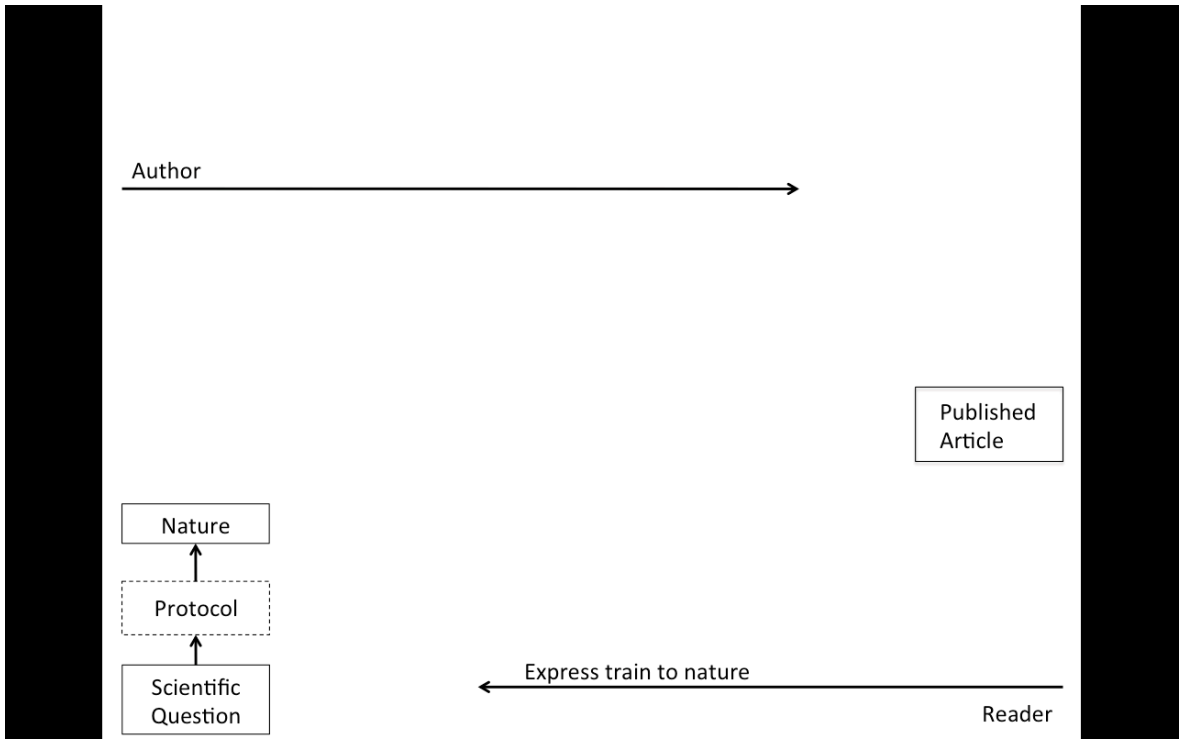
What is Reproducible Research?



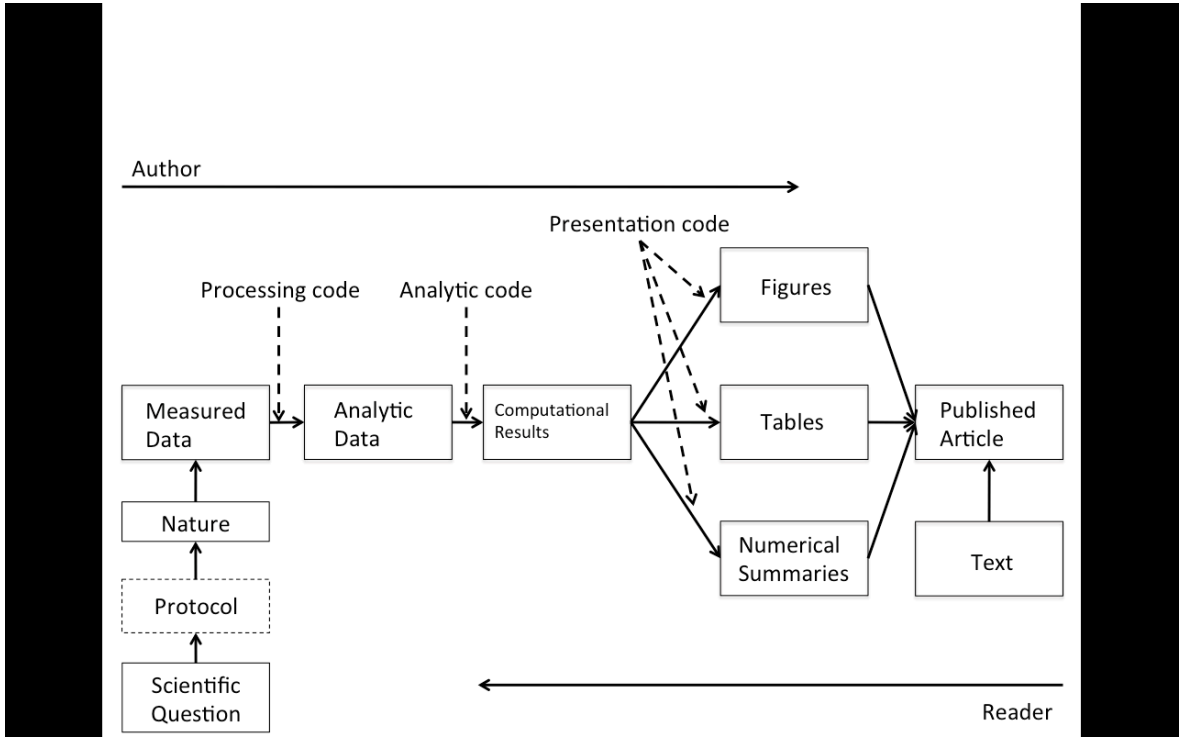
What is Reproducible Research?



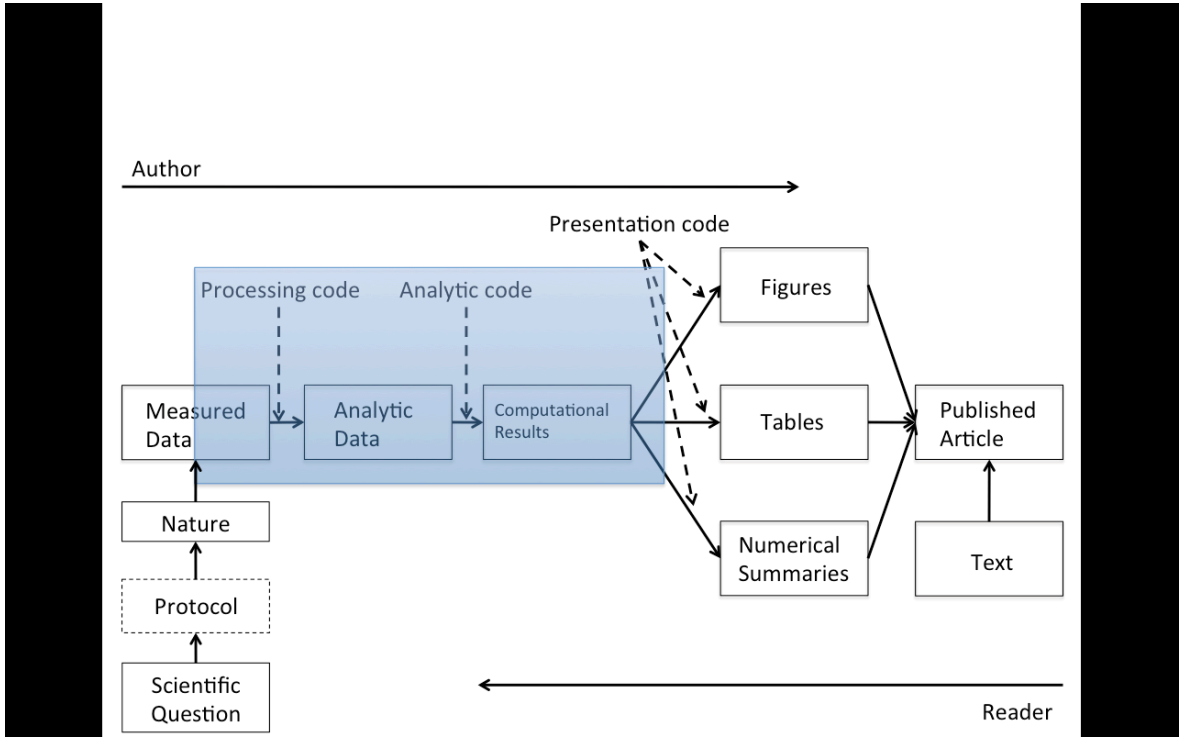
What is Reproducible Research?



What is Reproducible Research?



What is Reproducible Research?



What Problem Does Reproducibility Solve?

What we get

- Transparency
- Data Availability
- Software / Methods Availability
- Improved Transfer of Knowledge

What Problem Does Reproducibility Solve?

What we get

- Transparency
- Data Availability
- Software / Methods Availability
- Improved Transfer of Knowledge

What we do NOT get

- Validity / Correctness of the analysis

What Problem Does Reproducibility Solve?

What we get

- Transparency
- Data Availability
- Software / Methods Availability
- Improved Transfer of Knowledge

What we do NOT get

- Validity / Correctness of the analysis

An analysis can be reproducible and still be wrong

We want to know “can we trust this analysis?”

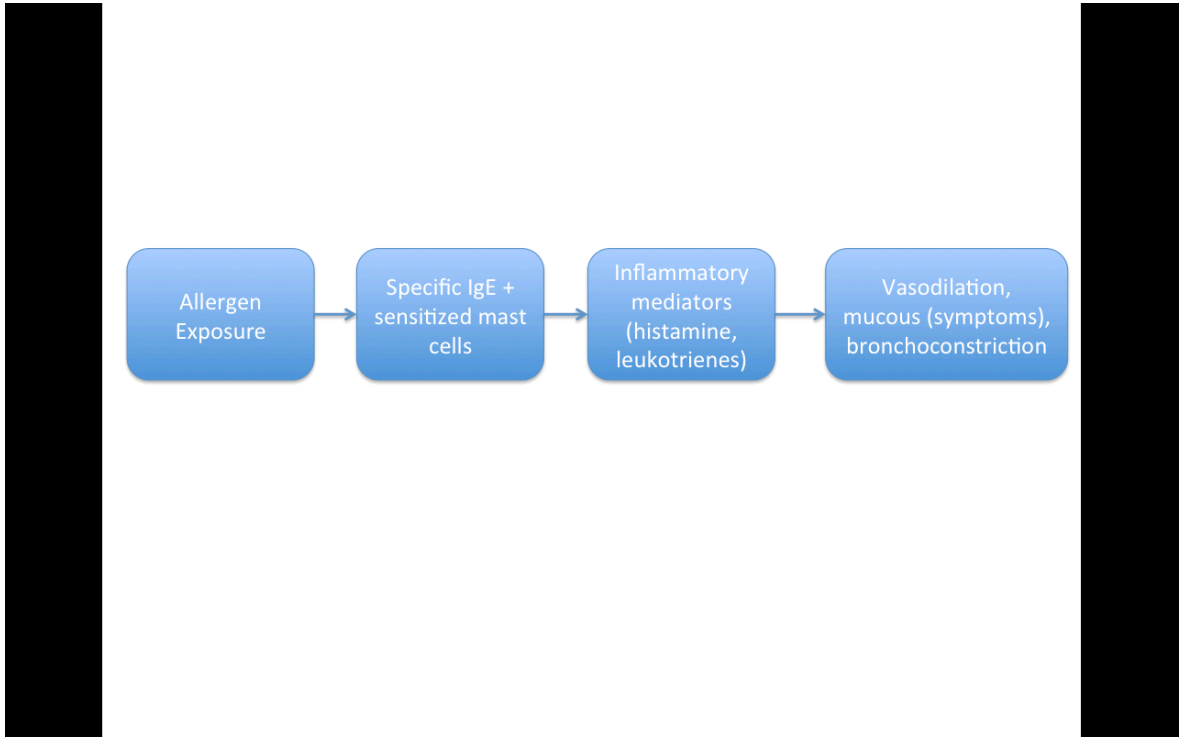
Does requiring reproducibility deter bad analysis?

Problems with Reproducibility

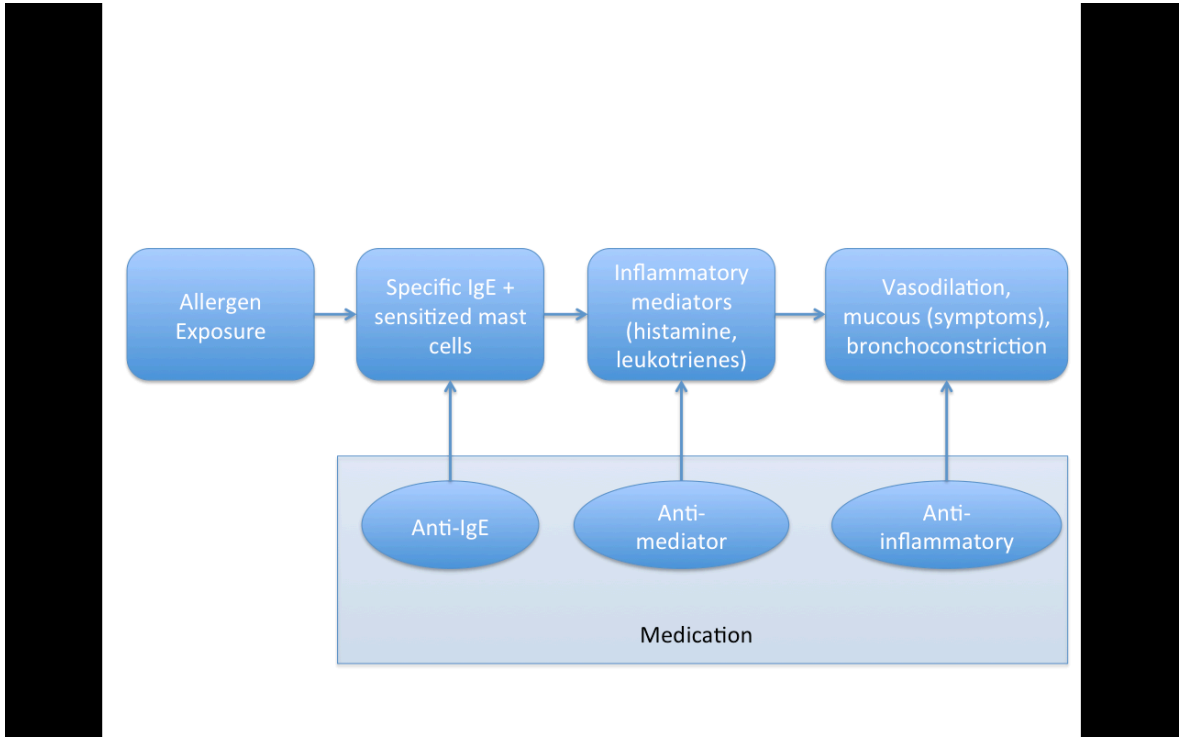
The premise of reproducible research is that with data/code available, people can check each other and the whole system is self-correcting

- Addresses the most “downstream” aspect of the research process – post-publication
- Assumes everyone plays by the same rules and wants to achieve the same goals (i.e. scientific discovery)

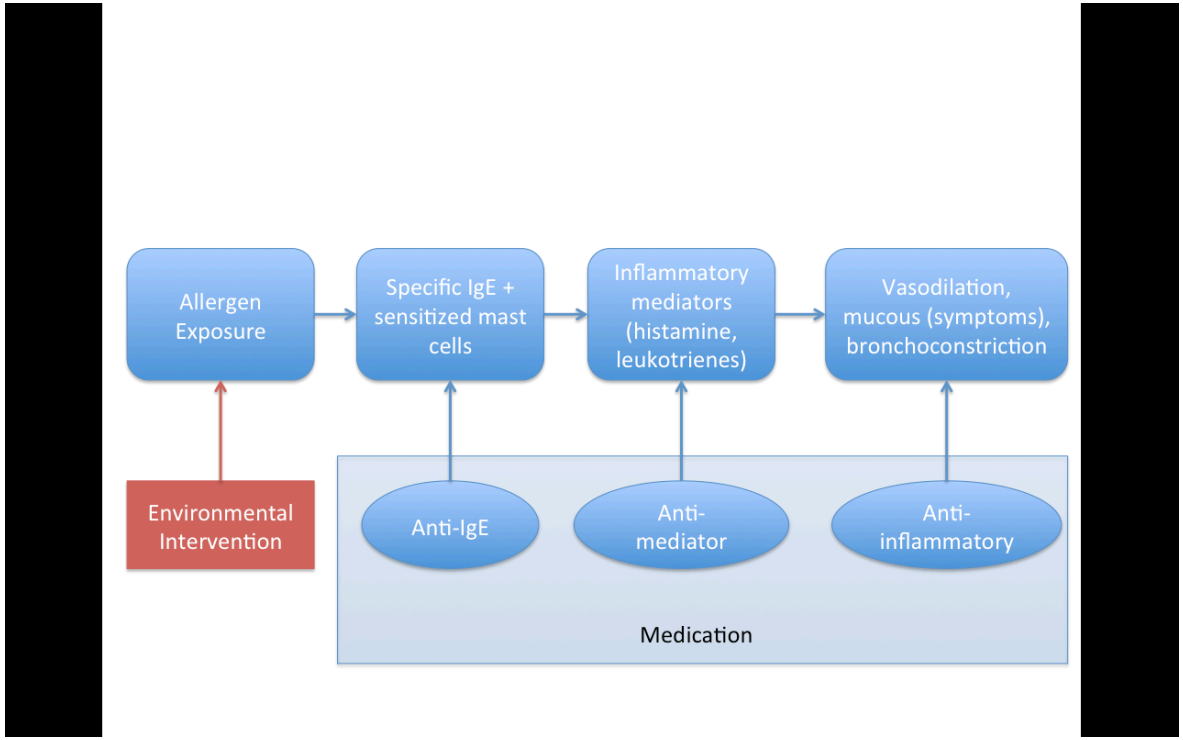
An Analogy from Asthma



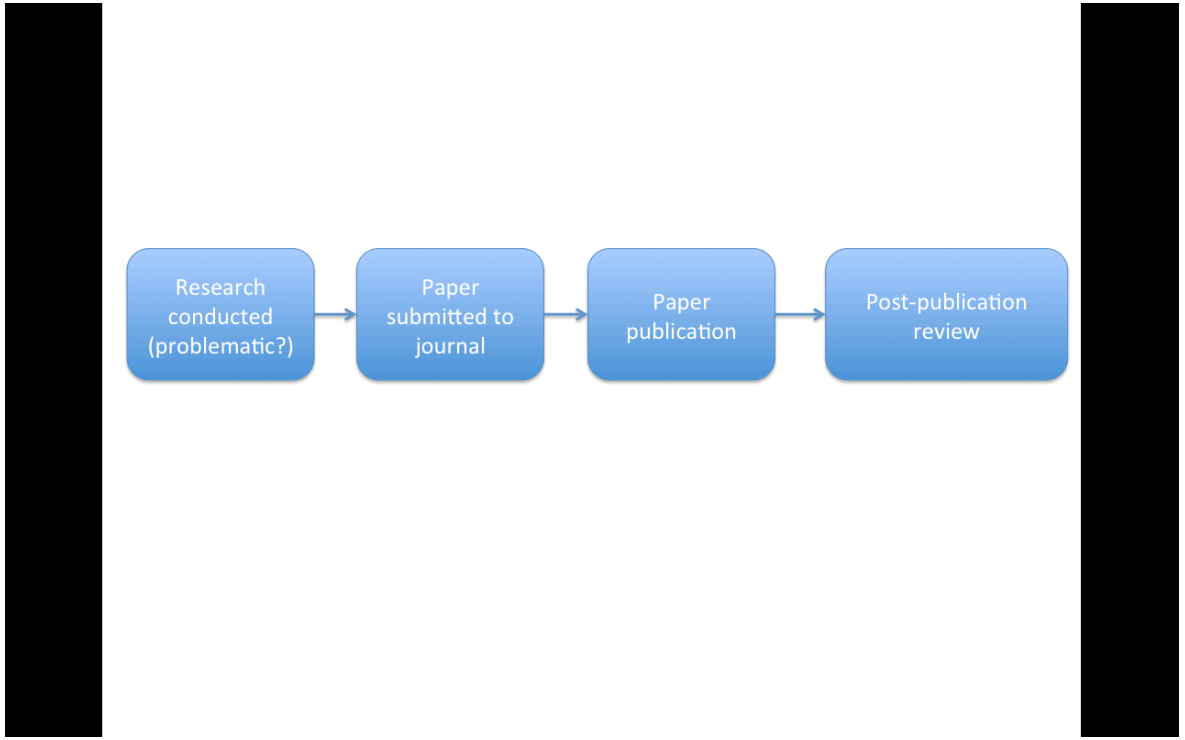
An Analogy from Asthma



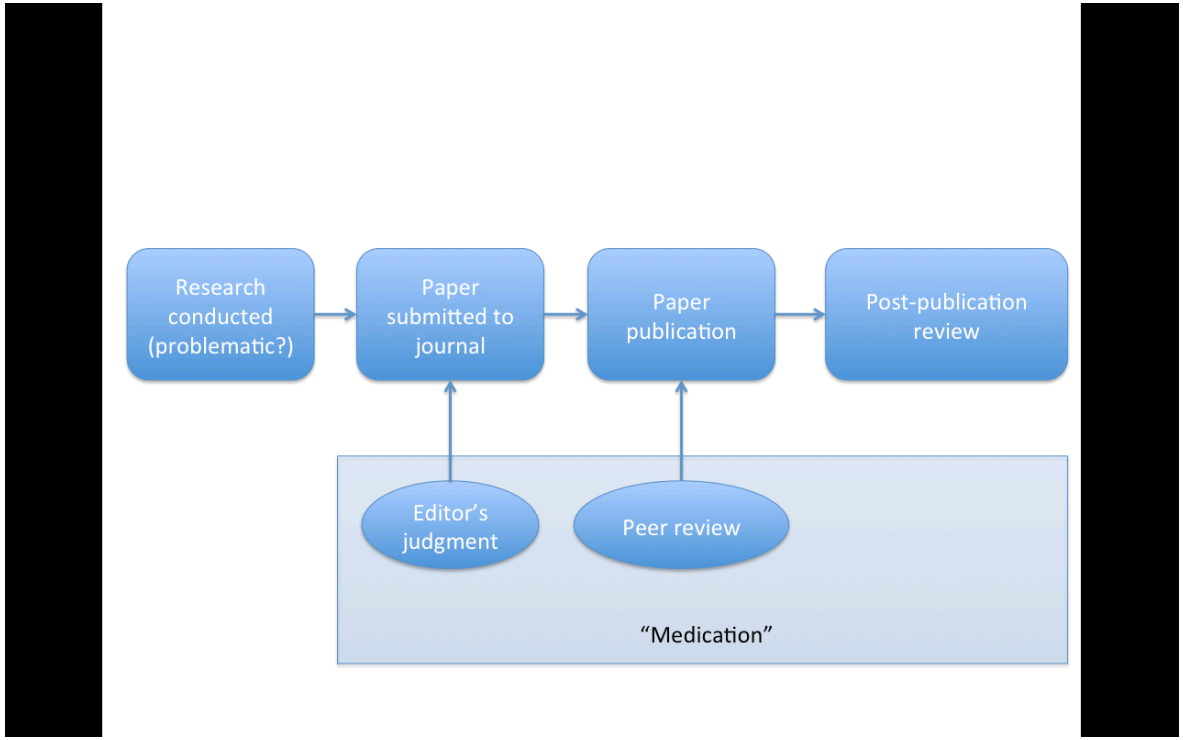
An Analogy from Asthma



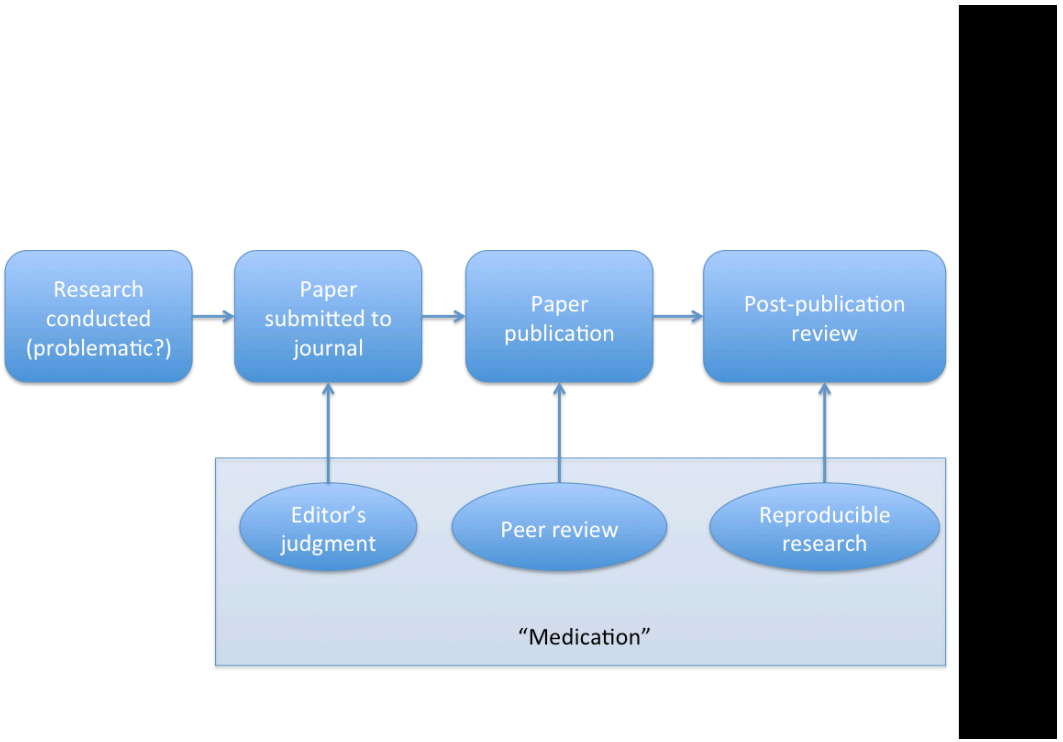
Scientific Dissemination Process



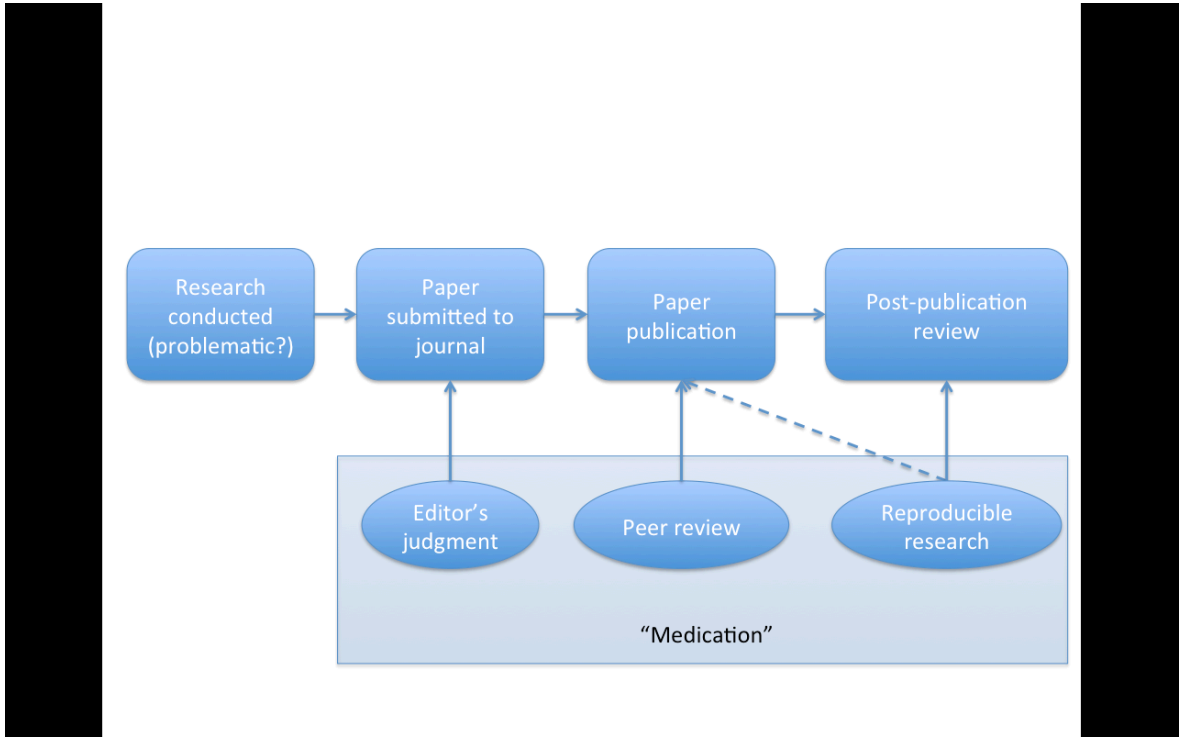
Scientific Dissemination Process



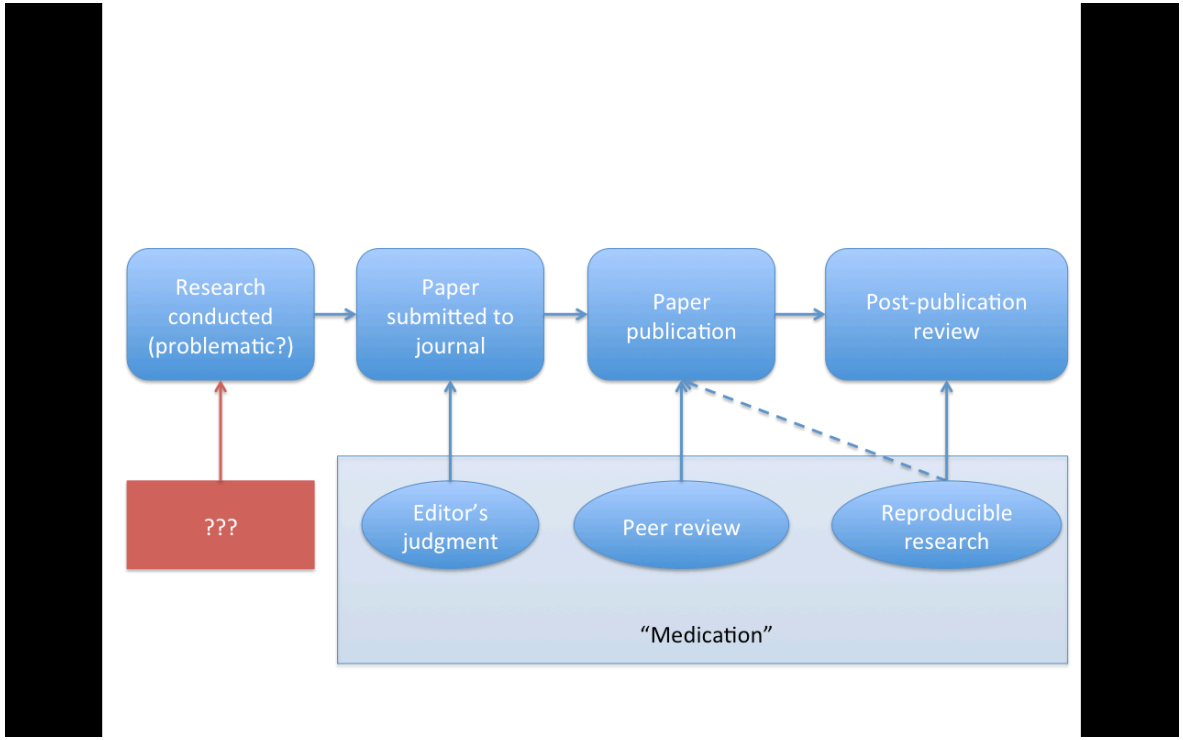
Scientific Dissemination Process



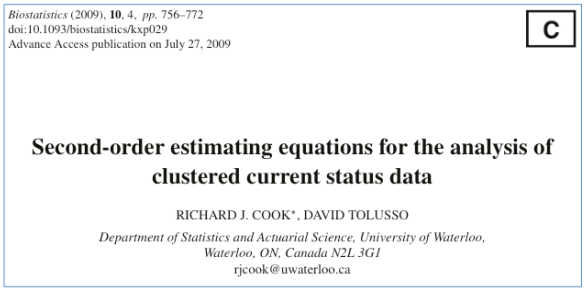
Scientific Dissemination Process



Scientific Dissemination Process



At Biostatistics



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

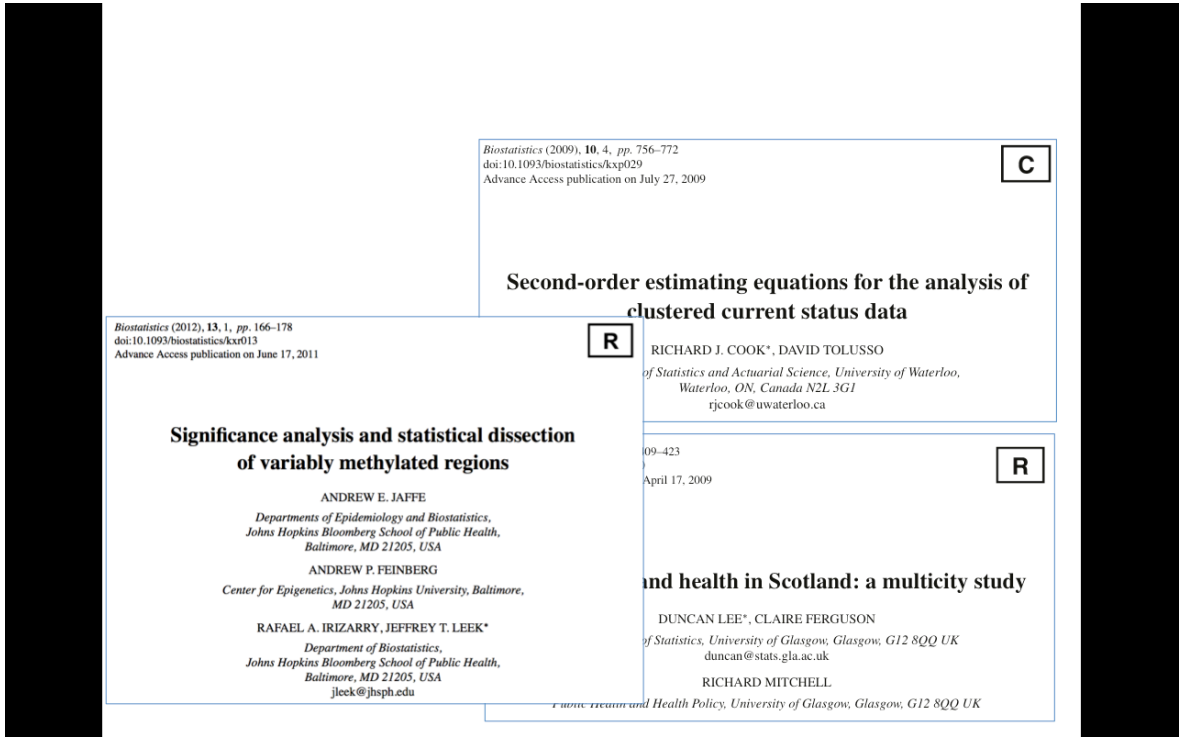
DUNCAN LEE*, CLAIRE FERGUSON

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

At Biostatistics



Biostatistics (2012), **13**, 1, *pp.* 166–178
doi:10.1093/biostatistics/kxr013
Advance Access publication on June 17, 2011

R

Significance analysis and statistical dissection of variably methylated regions

ANDREW E. JAFFE

Departments of Epidemiology and Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

ANDREW P. FEINBERG

Center for Epigenetics, Johns Hopkins University, Baltimore, MD 21205, USA

RAFAEL A. IRIZARRY, JEFFREY T. LEEK*

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

jleek@jhsph.edu

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

of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1

rjcook@uwaterloo.ca

09–423

April 17, 2009

R

nd health in Scotland: a multicity study

DUNCAN LEE*, CLAIRE FERGUSON

of Statistics, University of Glasgow, Glasgow, G12 8QQ UK

duncan@stats.gla.ac.uk

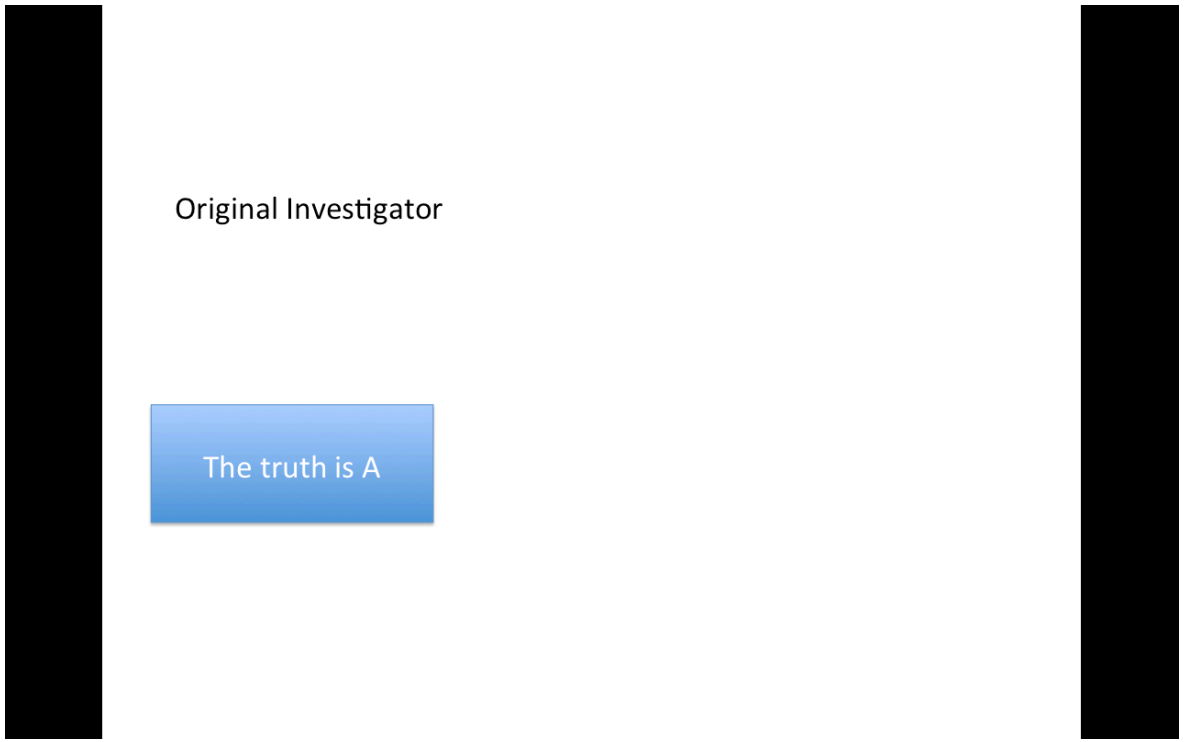
RICHARD MITCHELL

Public Health and Health Policy, University of Glasgow, Glasgow, G12 8QQ UK

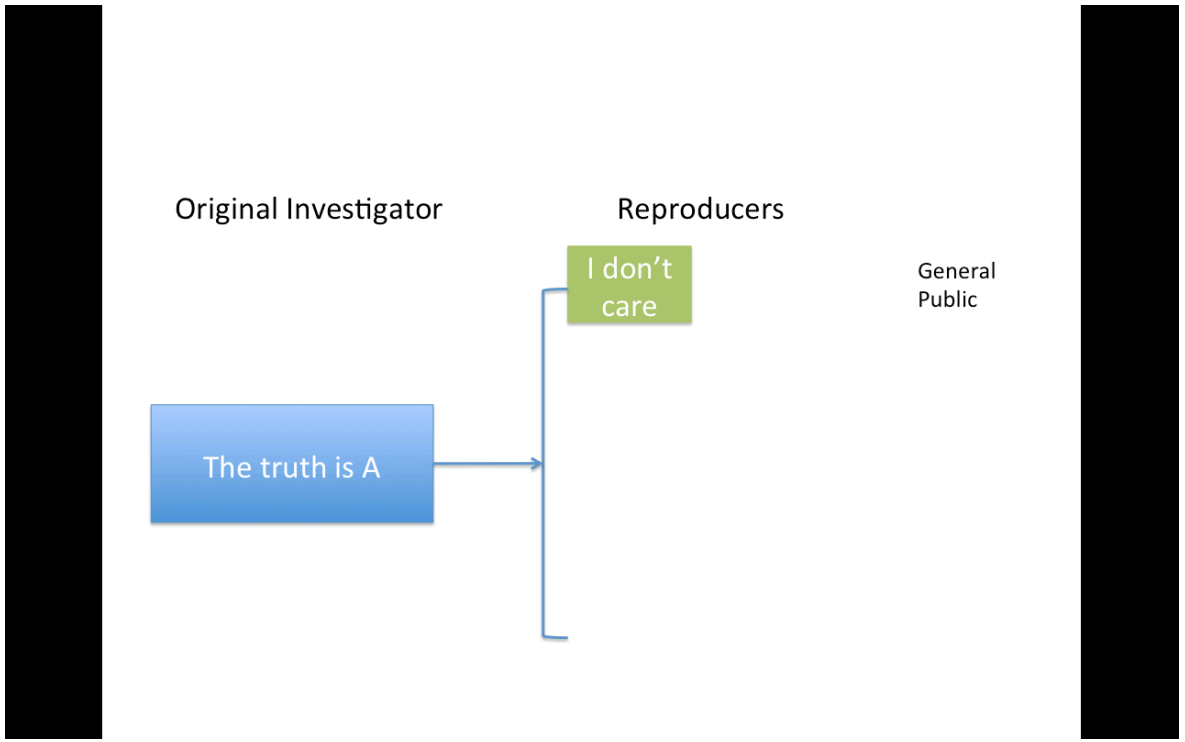
Who Reproduces Research?

- For reproducibility to be effective as a means to check validity, someone needs to do something
  - Re-run the analysis; check results match
  - Check the code for bugs/errors
  - Try alternate approaches; check sensitivity
- The need for someone to do something is inherited from traditional notion of replication
- Who is "someone" and what are their goals?

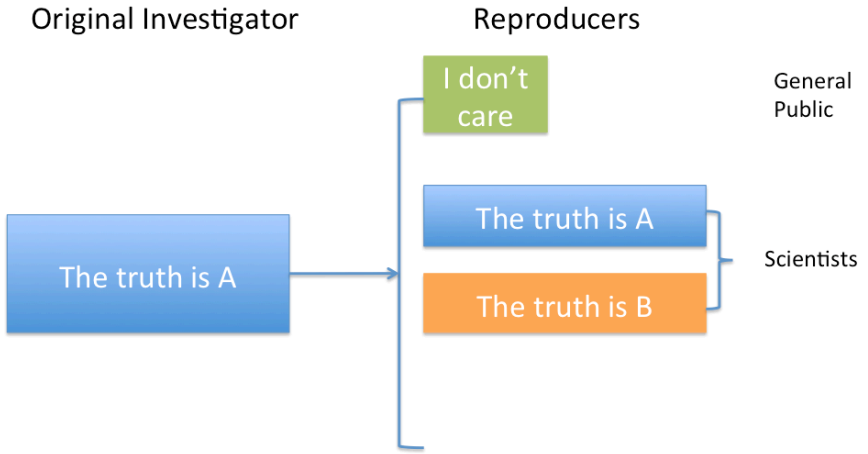
Who Reproduces Research?



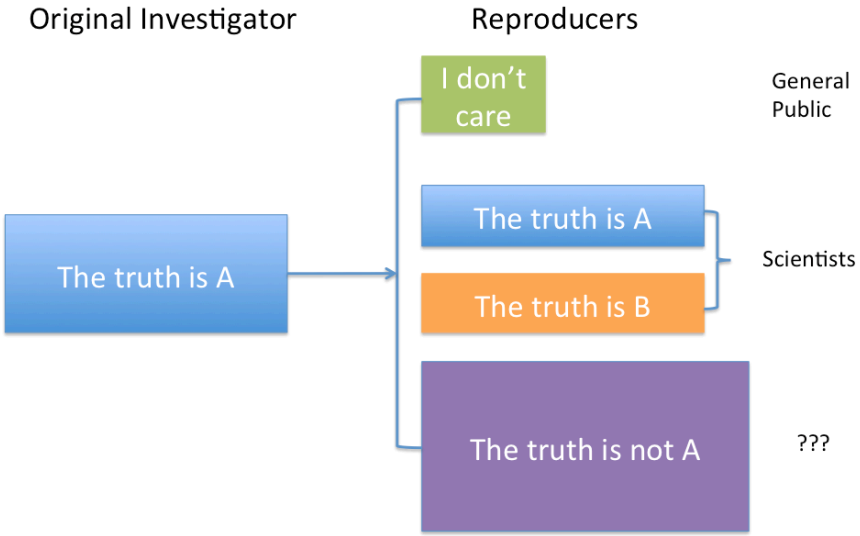
Who Reproduces Research?



Who Reproduces Research?



Who Reproduces Research?



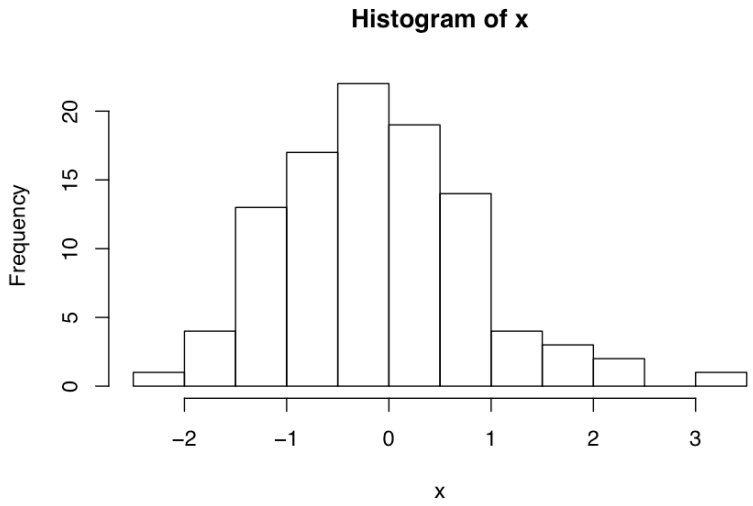
The Story So Far

- Reproducibility brings transparency (wrt code+data) and increased transfer of knowledge
- A lot of discussion about how to get people to share data
- Key question of "can we trust this analysis?" is not addressed by reproducibility
- Reproducibility addresses potential problems long after they've occurred ("downstream")
- Secondary analyses are inevitably coloured by the interests/motivations of others

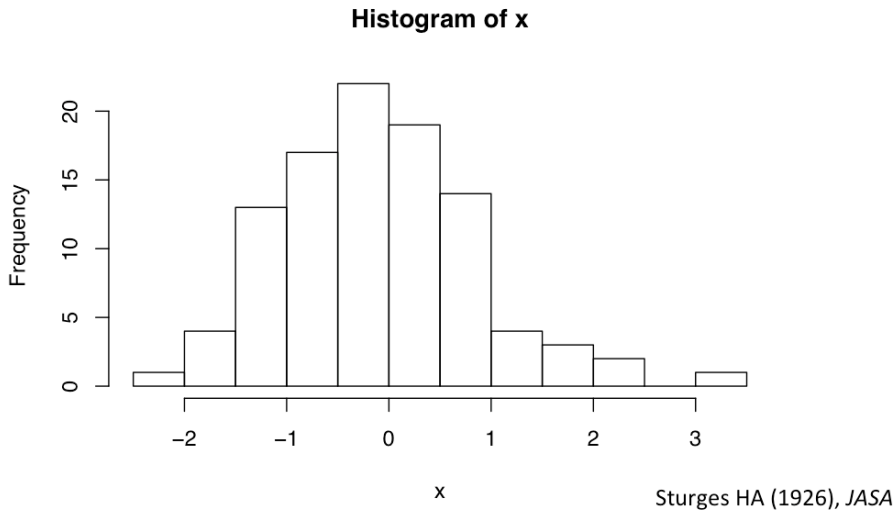
Evidence-based Data Analysis

- Most data analyses involve stringing together many different tools and methods
- Some methods may be standard for a given field, but others are often applied ad hoc
- We should apply thoroughly studied (via statistical research), mutually agreed upon methods to analyze data whenever possible
- There should be evidence to justify the application of a given method

Evidence-based Data Analysis



Evidence-based Data Analysis



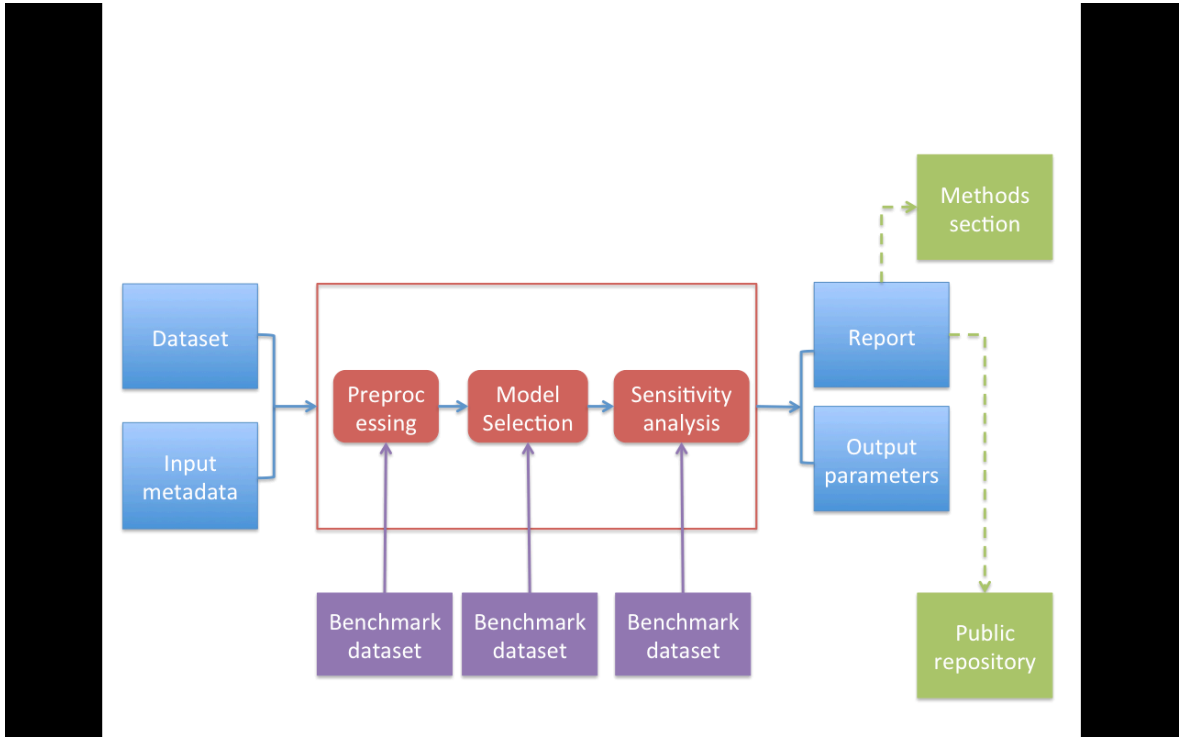
Sturges HA (1926), *JASA*

Scott DW (1979), *Biometrika*

Evidence-based Data Analysis

- Create analytic pipelines from evidence-based components – standardize it
- A Deterministic Statistical Machine http://goo.gl/Qvlhuv
- Once an evidence-based analytic pipeline is established, we shouldn't mess with it
- Analysis with a "transparent box"
- Reduce the "researcher degrees of freedom"
- Analogous to a pre-specified clinical trial protocol

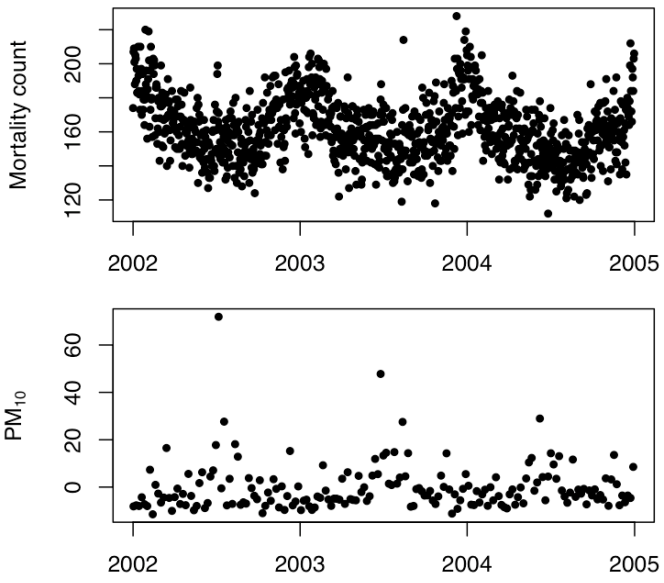
Deterministic Statistical Machine



Case Study: Estimating Acute Effects of Ambient Air Pollution Exposure

- Acute/short-term effects typically estimated via panel studies or time series studies
- Work originated in late 1970s early 1980s
- Key question: "Are short-term changes in pollution associated with short-term changes in a population health outcome?"
- Studies usually conducted at community level
- Long history of statistical research investigating proper methods of analysis

Data from New York City



Case Study: Estimating Acute Effects of Ambient Air Pollution Exposure

- Can we encode everything that we have found in statistical/epidemiological research into a single package?
- Time series studies do not have a huge range of variation; typically involves similar types of data and similar questions
- We can create a deterministic statistical machine for this area?

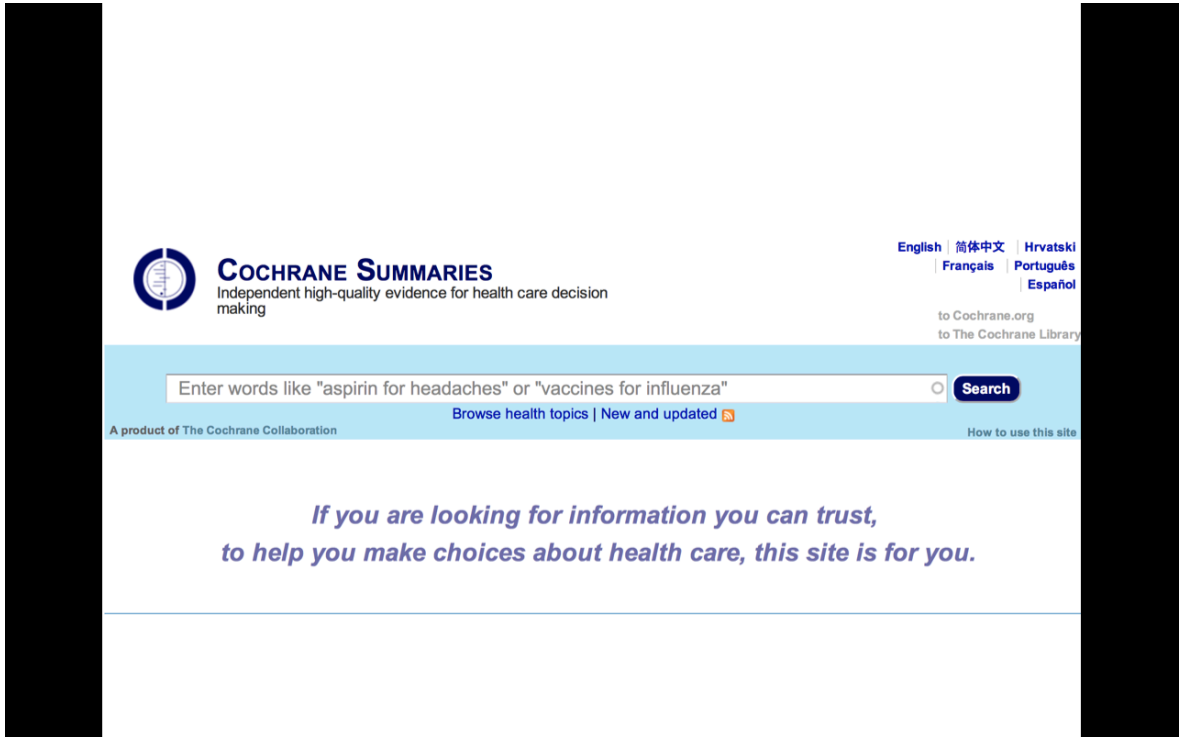
DSM Modules for Time Series Studies of Air Pollution and Health

1. Check for outliers, high leverage, overdispersion
2. Fill in missing data? NO!
3. Model selection: Estimate degrees of freedom to adjust for unmeasured confounders
    - Other aspects of model not as critical
4. Multiple lag analysis
5. Sensitivity analysis wrt
    - Unmeasured confounder adjustment
    - Influential points

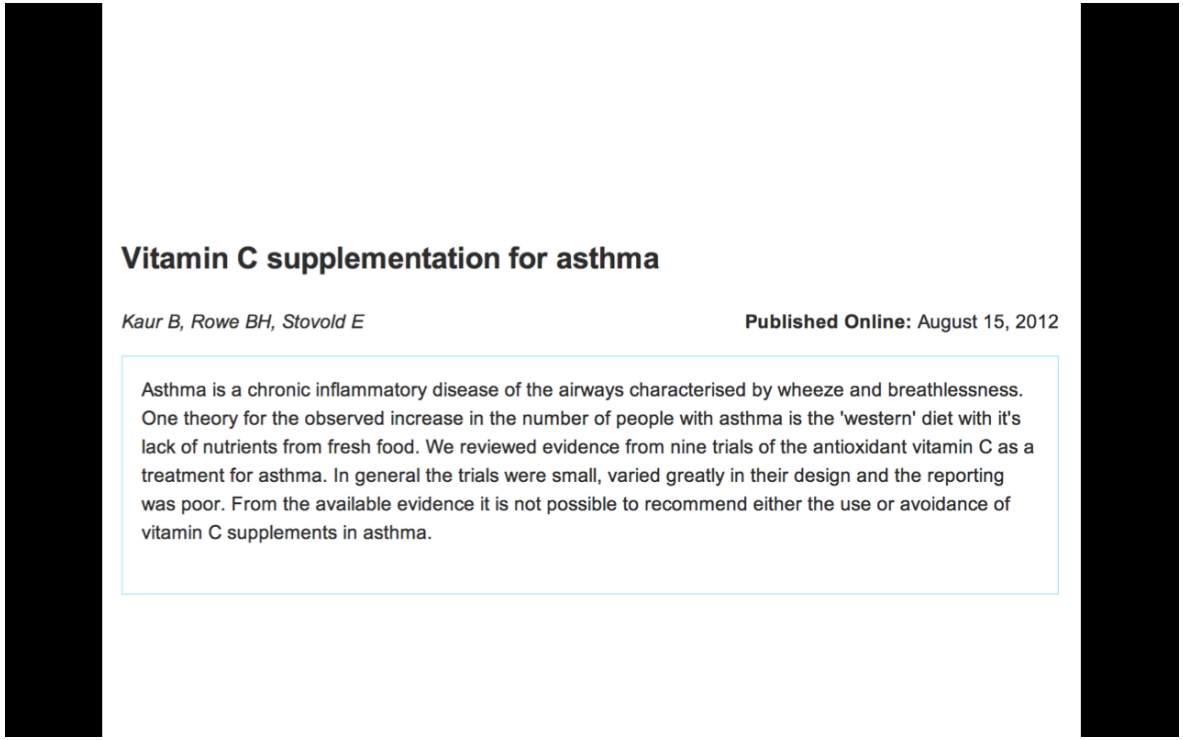
Where to Go From Here?

- One DSM is not enough, we need many!
- Different problems warrant different approaches and expertise
- A curated library of machines providing state-of-the art analysis pipelines
- A CRAN/CPAN/CTAN/… for data analysis
- Or a “Cochrane Collaboration” for data analysis

A Model: Cochrane Collaboration



A Model: Cochrane Collaboration

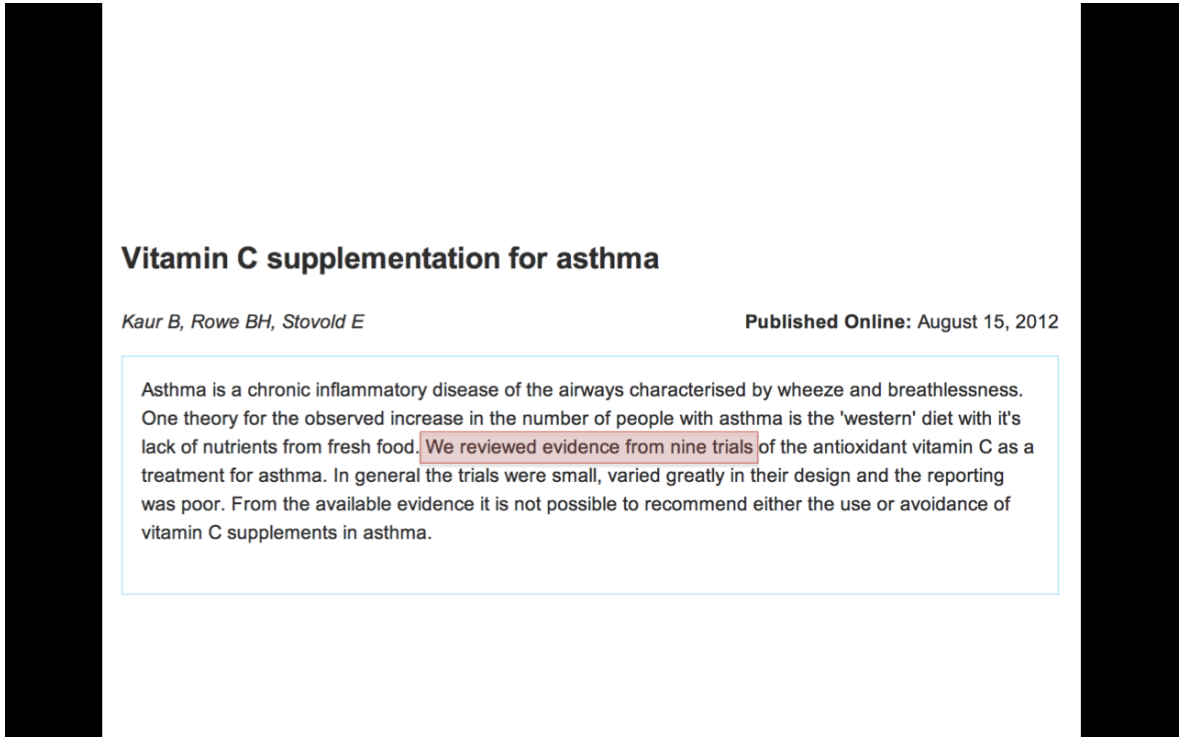


Vitamin C supplementation for asthma

Kaur B, Rowe BH, Stovold E **Published Online:** August 15, 2012

Asthma is a chronic inflammatory disease of the airways characterised by wheeze and breathlessness. One theory for the observed increase in the number of people with asthma is the 'western' diet with it's lack of nutrients from fresh food. We reviewed evidence from nine trials of the antioxidant vitamin C as a treatment for asthma. In general the trials were small, varied greatly in their design and the reporting was poor. From the available evidence it is not possible to recommend either the use or avoidance of vitamin C supplements in asthma.

A Model: Cochrane Collaboration



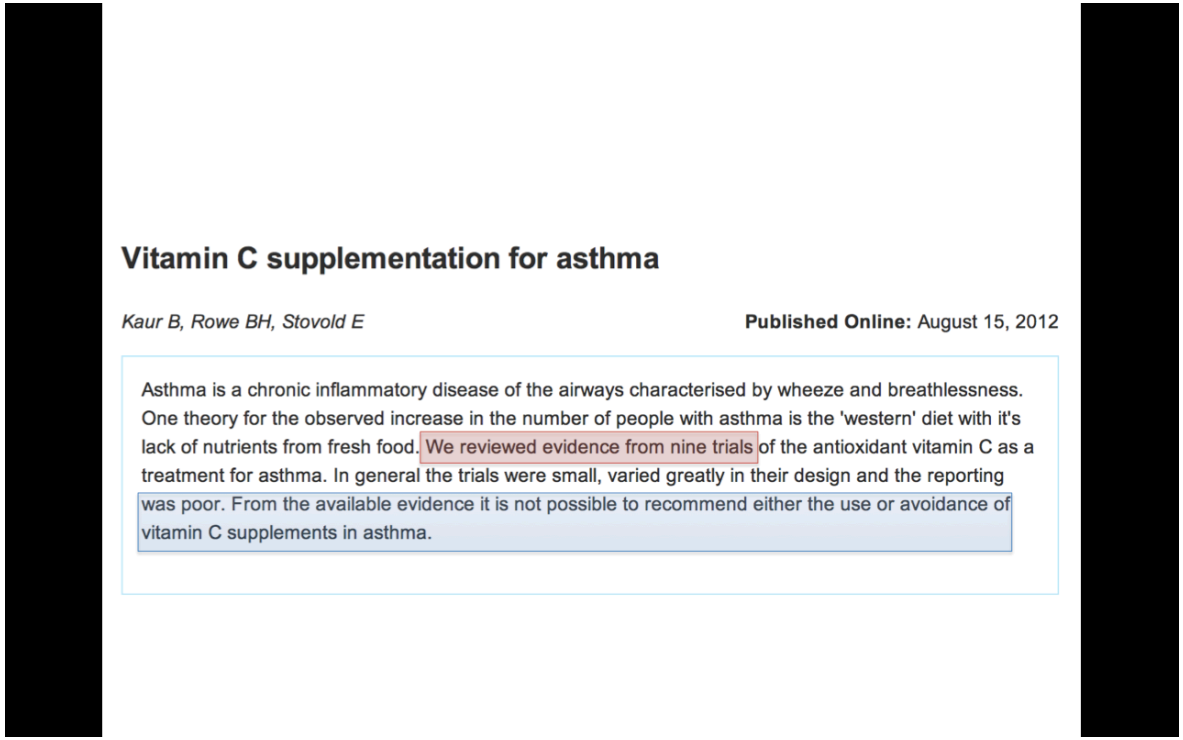
Vitamin C supplementation for asthma

Kaur B, Rowe BH, Stovold E

Published Online: August 15, 2012

Asthma is a chronic inflammatory disease of the airways characterised by wheeze and breathlessness. One theory for the observed increase in the number of people with asthma is the 'western' diet with it's lack of nutrients from fresh food. We reviewed evidence from nine trials of the antioxidant vitamin C as a treatment for asthma. In general the trials were small, varied greatly in their design and the reporting was poor. From the available evidence it is not possible to recommend either the use or avoidance of vitamin C supplements in asthma.

A Model: Cochrane Collaboration



Vitamin C supplementation for asthma

Kaur B, Rowe BH, Stovold E **Published Online:** August 15, 2012

Asthma is a chronic inflammatory disease of the airways characterised by wheeze and breathlessness. One theory for the observed increase in the number of people with asthma is the 'western' diet with it's lack of nutrients from fresh food. We reviewed evidence from nine trials of the antioxidant vitamin C as a treatment for asthma. In general the trials were small, varied greatly in their design and the reporting was poor. From the available evidence it is not possible to recommend either the use or avoidance of vitamin C supplements in asthma.

A Curated Library of Data Analysis

- Provide packages that encode data analysis pipelines for given problems, technologies, questions
- Curated by experts knowledgeable in the field
- Documentation/references given supporting each module in the pipeline
- Changes introduced after passing relevant benchmarks/unit tests

Summary

- Reproducible research is important, but does not necessarily solve the critical question of whether a data analysis is trustworthy
- Reproducible research focuses on the most "downstream" aspect of research dissemination
- Evidence-based data analysis would provide standardized, best practices for given scientific areas and questions
- Gives reviewers an important tool without dramatically increasing the burden on them
- More effort should be put into improving the quality of "upstream" aspects of scientific research