A Shifting Landscape for Laboratory Science

Reproducibility and Related Issues

John Maindonald, MSI, Australian National University

10 May 2016

There are increasing calls, in Nature and elsewhere, for reform of major aspects of scientific processes — to address concerns about reproducibility, to improve efficiency, and to leverage current technology more effectively. In some major areas, an unacceptably high proportion of published work is not reproducible. Reasons for this, and initiatives that are designed to address the problem, will be noted. A component of reproducibility is ensuring that data analysis details can be readily reproduced. An ideal is to have as a primary record a document that combines text and code, and which can be processed to produce a final paper that has all tables, graphs and analysis results. Individual memories, bottom drawers, and stacks of files, are less than ideal for this purpose. Data should be accessed from a properly maintained database. The benefits are wide-ranging — for avoiding manual errors, for communication of details of work within and outside the organization, and for preserving data and code for any needed revision or updating or future use.

Creaking Scientific Processes

- 'Obvious' nonsense: the hurricanes saga
  - The *Growth in Time of Debt* saga
- Scientific processes need to leverage new technology
  - Not just the lab equipment, but processes
- The bloggers are coming, already here!
- Complexity is everywhere. Manage it!
  - Transfer complexity to the outside world!
- Demonstrated failures in reproducibility
  - Amgen: 6 'successes' from 53 'landmark' cancer studies.
    NB: Issues with the studies that failed . . .
  - Bayer: Drug studies also come off badly
- Reproducible reporting

US Hurricane Deaths vs $ Damage

Jung, Shavitta, Viswanathana, and Hilbe (2013). "Female hurricanes are deadlier than male hurricanes".
http://www.pnas.org/cgi/doi/10.1073/pnas.1402786111.
PNAS, May 2014

A 'female' name caused less concern than a 'male' name?

Damage data (in 2013 $) are from ICAT: "ICAT . . . was founded in 1998 to provide production, underwriting, and risk management services for insurance companies."

Deaths were related to damage, in 2013 dollars, for a comparable hurricane in 2013 (but my graph will use 2014 data)

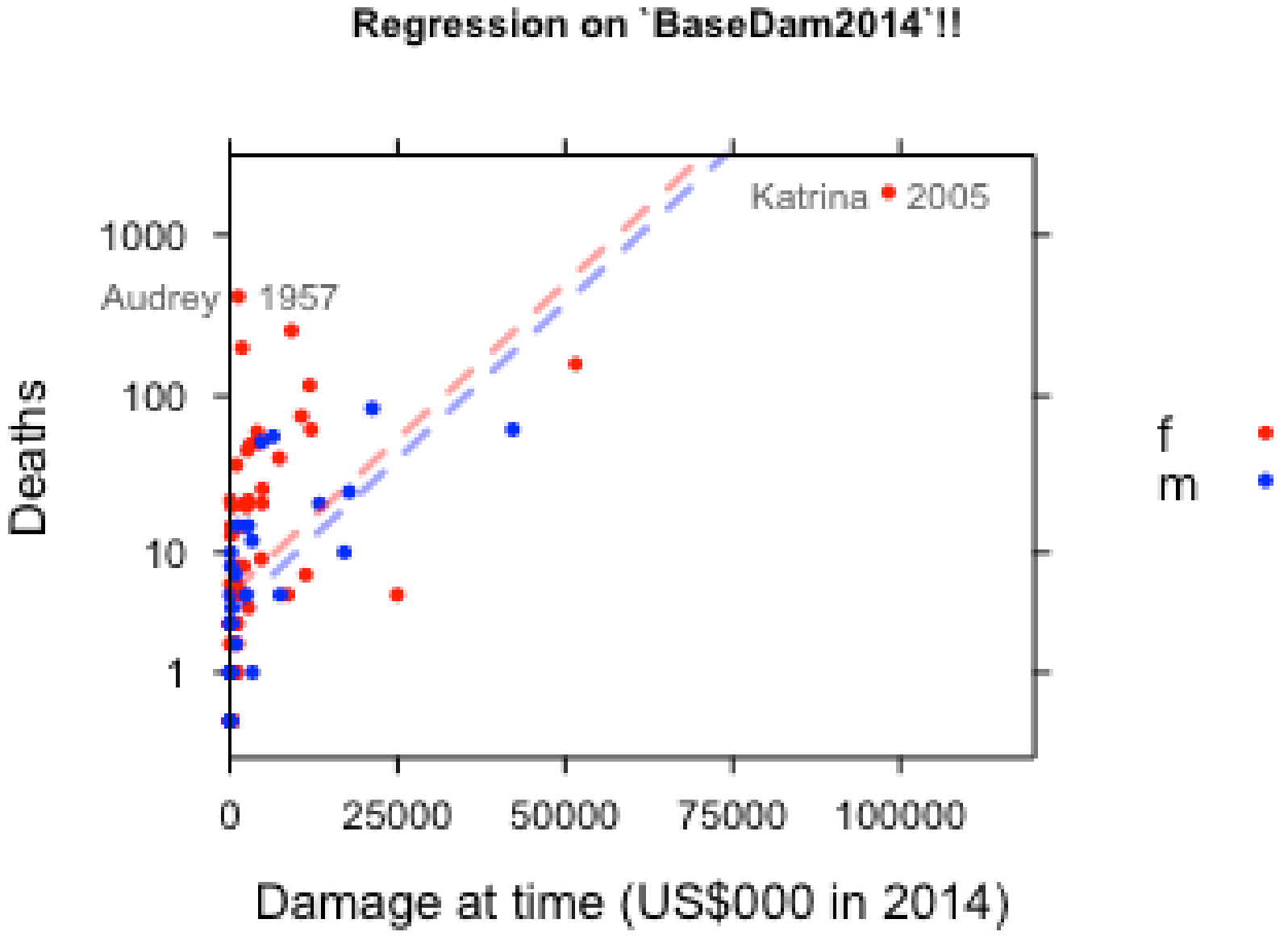
Explanatory Variables

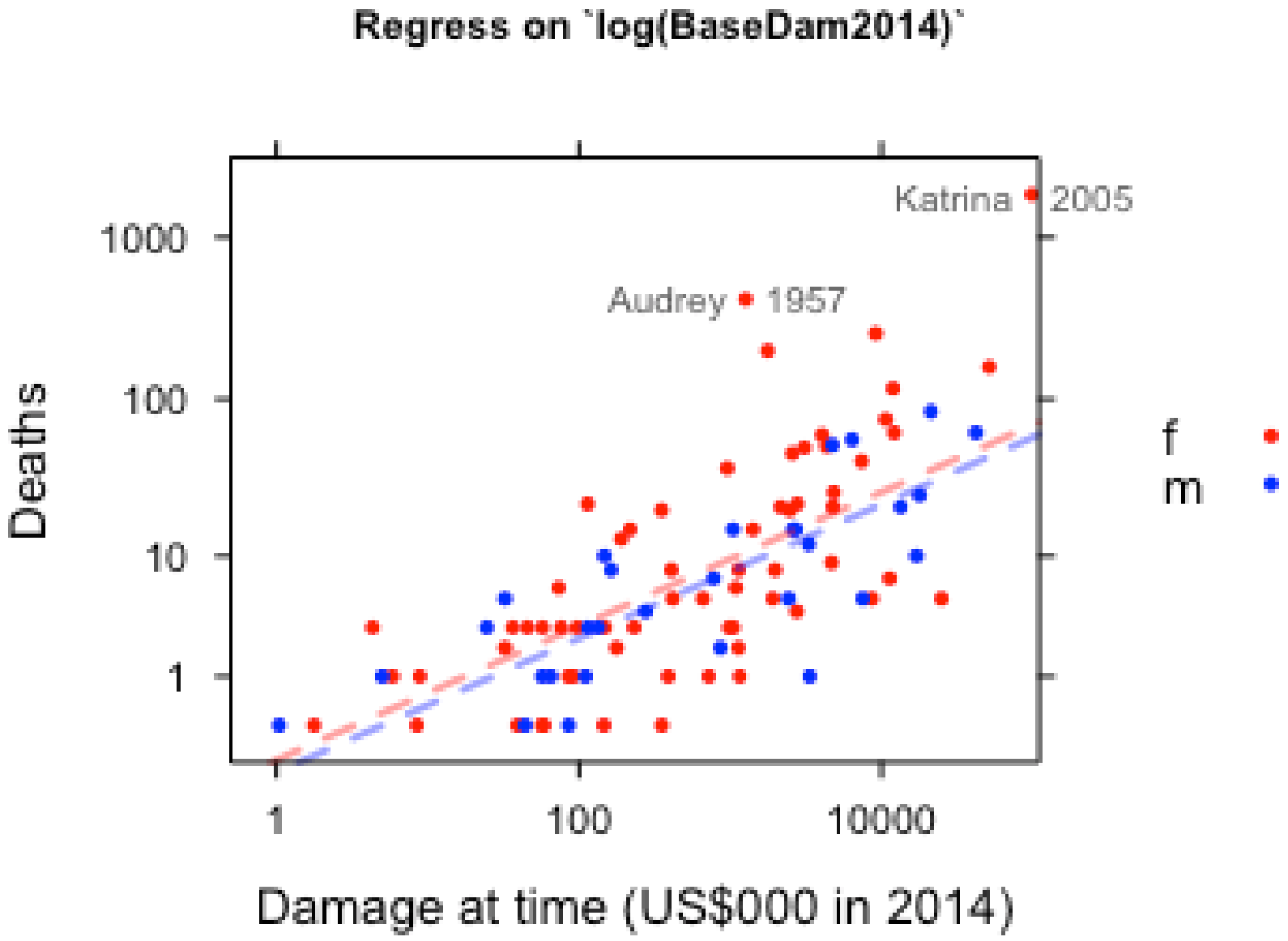
- Damage variables (in either case, in 2013 or 2014 $)
  - Base Damage, i.e., Damage at the time
  - ICAT damage estimate, if storm were in 2013 or 2014
- Gender; factor with levels `f`, `m`
  (Jung et al used a femaleness panel rating, on a 1 to 11 scale.)

[Also available was Barometric pressure at (first) US landfall]

Question: Which makes more sense?

- Use as explanatory variable: Base damage or ICAT damage?
  - Jung et al: ICAT damage; next slide: Base damage
- With deaths on log scale,
  measure damage on untransformed or log scale?


Regression on `BaseDam2014`!!
Katrina 2005
Audrey 1957
1000
100
10
1
Deaths
0
25000
50000
75000
100000
Damage at time (US$000 in 2014)
f
m


Regress on `log(BaseDam2014)`
Katrina 2005
Audrey 1957
1000
100
10
1
Deaths
1
100
10000
Damage at time (US$000 in 2014)
f
m

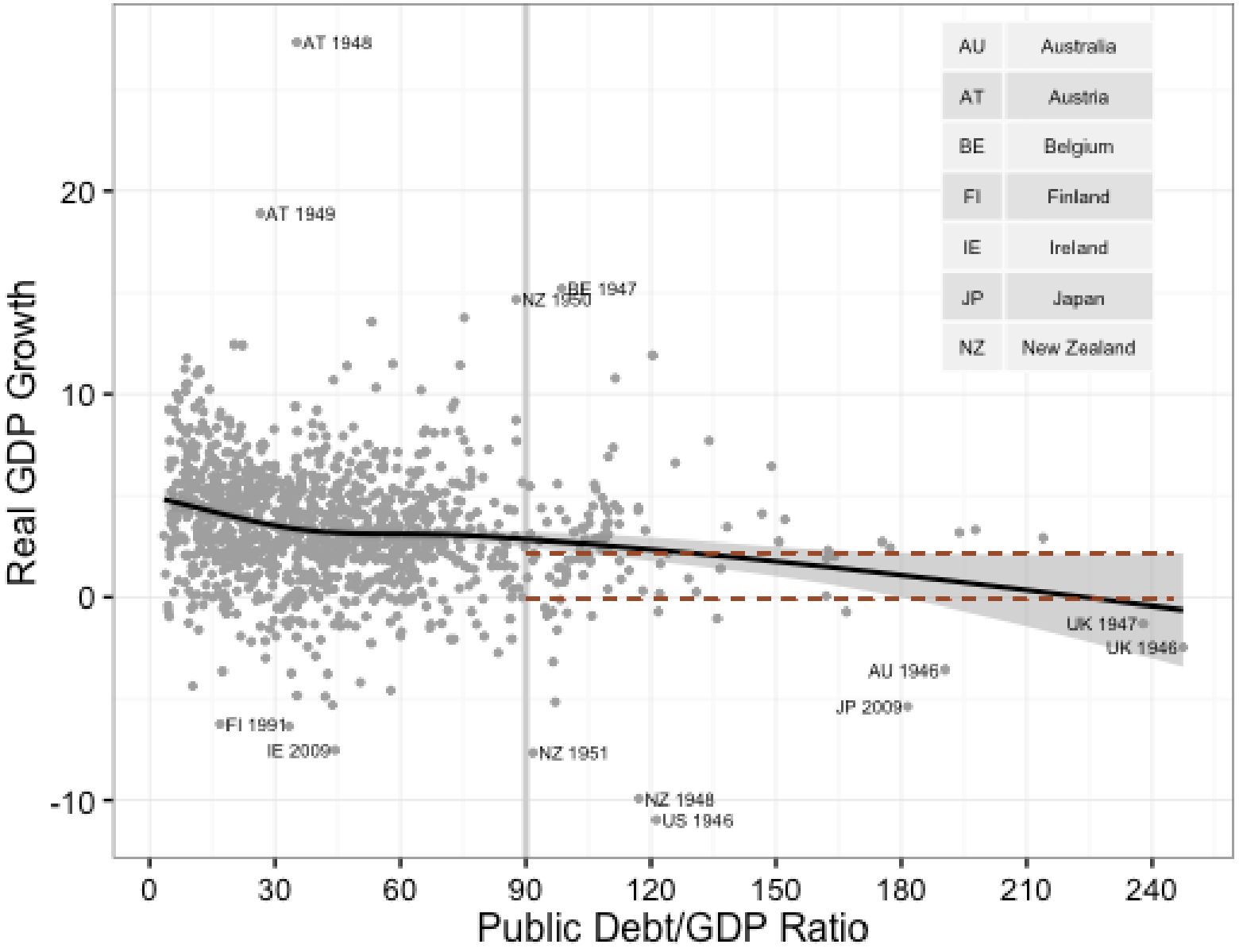
The *Growth in Time of Debt* Saga

The *Herndon et al* debunking of Reinhart and Rogoff: "*Growth in Time of Debt*" (2010) created a huge stir.

"… [Identified errors] include spreadsheet [mistakes] …, omission of available data, weighting, and transcription, … [affecting] countries in the high public debt category.

… [The result is] a false image that high public debt ratios inevitably entail sharp declines in GDP growth. … there is a wide range of GDP growth … at every level of public debt among the 20 advanced economies that RR survey."

http://nymag.com/daily/intelligencer/2013/04/grad-student-who-shook-global-austerity-movement.html


AT 1948
AT 1949
NZ 1950
BE 1947
FI 1991
IE 2009
NZ 1951
NZ 1948
US 1946
AU 1946
JP 2009
UK 1947
UK 1946
AU Australia
AT Austria
BE Belgium
FI Finland
IE Ireland
JP Japan
NZ New Zealand
Real GDP Growth
20
10
0
-10
0
30
60
90
120
150
180
210
240
Public Debt/GDP Ratio

Herndon et al's Conclusions

"… RR's findings have served as an intellectual bulwark in support of austerity politics. The fact that RR's findings are wrong should therefore lead us to reassess the austerity agenda itself in both Europe and the United States."

"For econometricians a lesson from the problems in RR is the advantages of reproducible code relative to working spreadsheets. …

[O]ur simplified version of the spreadsheet and R code that reproduces RR and corrected results … [is] on our website."

Herndon (28yr old Grad Student at UMass) is now famous!

Science With or Versus the Bloggers

- If the media finds work interesting, beware!
  - Deficiencies may get unwelcome scrutiny
- Formalize processes for post-publication review?
  - Make Publish and Perish a reality!
- The blogosphere can be an early warning mechanism
  - It can identify failures of scientific process
  - Would it catch a new Andrew Wakefield event?
- Scientific processes can/must use the blogosphere?
  - Sure, there are dangers, but . . .
  - Good things are happening:
    Stack OverFlow; CrossValidated

Framework Issues — c.f. Disinfestation

- What can insect physiologists tell us about insect death?
- What are the effects of temperature?
  - On insect death?
  - On fruit damage?
- What is an appropriate measure of dose?
  - Does the concentration effect change with time?
  - How far can we trust the C-T product?
- What is the form of the dose-response?
  - Much of the literature has it wrong
- Results must be reproducible across years

Move complexity to the outside world

NB: Daniel Levitin: "The Organized Mind: Thinking Straight in an Age of Information Overload"

- Data — put complexity into well-managed databases
  - GeoNet, with access via QuakeSearch is a good model
  - Use for both internal and external access
- Laboratory processes
  - Monitor automatically wherever possible, . . .
- Complexity in working through the details of analyses
  - Marry together text and code in a "markup" version that can be processed to give the paper or report
  - Serious users of results will often want the filled-out details of the "markup" version.

Reproducibility — Evidence of Failures

- Ioannidis (2005)[1] — '… Most Published … Findings Are False'
  - This paper put reproducibility issues "on the map"
- Direct evidence — results do not reproduce
  - Examples shortly, best evidence is in psychology
  - Most worrying evidence is in cancer studies
- Warning signals, from examination of papers
  - 'lack of +ve & -ve controls', faulty stats, 'inappropriate use of key reagents', 'failure to repeat', …
- Results may be unreplicible
  - Key information may be omitted or wrong

1: Ioannidis (2005), 'Why Most Published Research Findings Are False'

Direct Evidence

- Amgen: Reproduced 6 only of 53 'landmark' cancer studies.[1]
  - Begley (2013) notes issues with the studies that failed
- Bayer: Main results from 19 of 65 'seminal' drug studies
  - NB, journal impact factor was not a good predictor![2]
- fMRI studies: 57 of 134 papers (42%) had >= 1 case lacking check on separate test image. Another 14%, unclear …[4]
- The Reproducibility:Psychology Project (~40% replicated)
  - Summary of results: 28 Aug 2015 issue of Science

1: Begley and Ellis (2012), 'Raise standards …'; NB also Begley (2013)
2: Prinz, Schlange, and Asadullah (2011), 'Believe it or not … drug targets'
3: Kriegeskorte et al. (2009), '… dangers of double dipping'
4: OSC (2015), 'Estimating the reproducibility of psychological science'

Collins & Tabak[1] — Factors include …

- poor training . . . in experimental design
- making provocative statements rather than presenting technical details
- Crucial experimental design elements that are too frequently ignored include blinding, randomization, replication, sample-size calculation and the effect of sex differences
- some scientists reputedly use a 'secret sauce' to make their experiments work — and withhold details . . . or describe them only vaguely . . .

Also: Deviations from stated protocol; errors in data; selective use of data; selection effects

1. Collins and Tabak (2014), '… NIH plans to enhance reproducibility'

What are the issues?

- Faulty design/reporting/execution (Begley, Collins & Tabak)
  - Repeating the same mistakes will not help
  - Some results are in principle unreplicable
- Selection effects, where mostly there is no effect
  - More enlightened use of p-values will help
  - More crucially, the total process is not transparent

We do not know, certainly not with any certainty, where the balance lies between these two different types of issue. In this area, science lacks a scientific understanding of its own processes.

Aside: Fisher on *p*-values

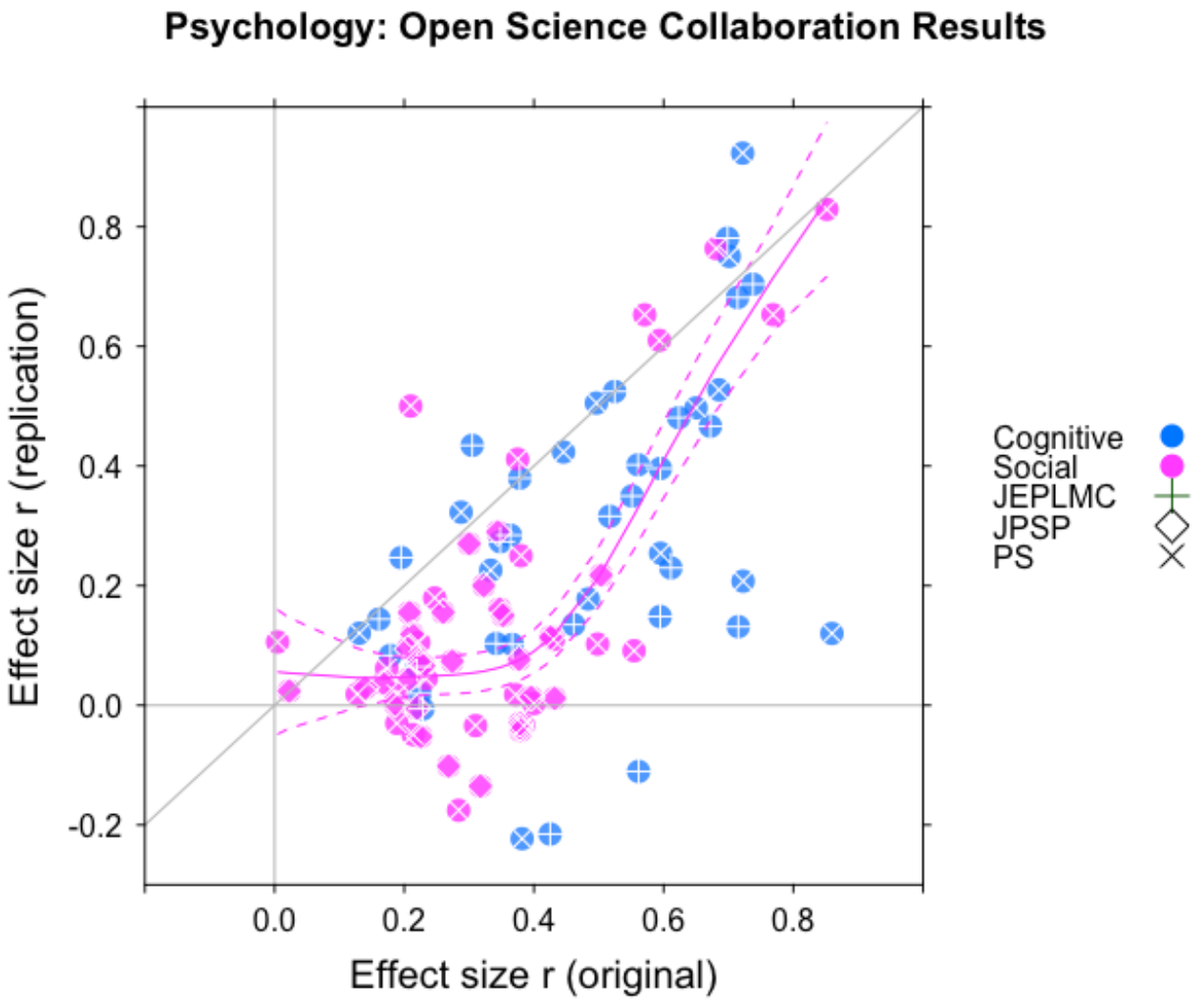
If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). … **A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level (0.05 or 0.02) of significance.**[1]

… we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.[2]

1: Fisher (1926)
2: Fisher (1937)

Science Community Responses

- The most effective responses have come from psychology.
- The next slide will summarize one major study.
- Studies in other areas are under way
  - Wet lab studies are harder to do
  - They interfere with regular work programs
  - Greater defensiveness than among psychologists.


Psychology: Open Science Collaboration Results
Effect size r (replication)
0.8
0.6
0.4
0.2
0.0
-0.2
0.0
0.2
0.4
0.6
0.8
Effect size r (original)
Cognitive
Social
JEPLMC
JPSP
PS

Is the criticism overblown?

- *The scientific process does finally identify the chaff*
  This contrasts with, e.g., alternative medicine.

Sure, but the process is far too protracted & tortuous.
Too many methodological failures go undetected.
Rewards systems encourage work that is poor quality.

- *One can never achieve 100% certainty*

Sure, but we can do a lot better than at present. Science should not be ignorant of its own processes.

- *Present processes are pretty much OK* (implied, not said)

We have the technology needed to do a vastly better job, but are not using it!

Journal & refereeing failures

- Referees & readers do not have the information needed
  - Know exactly what was done; check data, code
- Referees are at or beyond the limits of their expertise
  - Statistical analysis is an especial difficulty
- The system is not making good use of modern technology
  - plus, it interacts with rewards systems in malign ways
- Savvy critics are a huge untapped resource that is wasted
  - Other experts, in the area or in relevant areas
- Put papers out for comment pre-publication.

Open Science's response: make all processes transparent

Commentary in *Science* (June 26 2015)

1. Self-correction in Science at work[1]

- Publish replications (PPS now has a section for this)
- Highlight & reward completeness of information
- Encourage publishing well (not often), …
- Create a culture that is willing to admit mistakes, …
- School scientists in research ethics

Communication and Storage

- Spoken word
- Writing
- Printing (as it relates to reporting & publishing)
  - For scientific work, a fairly complete record
  - Nowadays, typically, a very incomplete record
- Computer-based systems and the world wide web
  - Rethink (NOT adapt) paper-based systems
  - Old excuses for the deficiencies of paper based sytems no longer wash.

Note the importance of organizational memory.

Promiscuous Publication?

"Now we are witnessing the transition to yet another scholarly communication system — one that will harness the technology of the Web to vastly improve dissemination. … The Web opens the workshop windows to disseminate scholarship as it happens, erasing the artificial distinction between process and product. …

Today's publication silos will be replaced by a set of decentralized, interoperable services that are built on a core infrastructure of open data and evolving standards — like the Web itself …. This 'decoupled journal' publishes promiscuously, then subjects products to rigorous review through the aggregated judgements of expert communities, supporting both rapid, fine-grained filtering and consistent, meaningful evaluation."

Jason Priem: Nature 495, 437–440 (28 March 2013) doi:10.1038/495437a

An Open Source Model for Science

- Open Source Malaria — think “Linux for Malaria Research"[1]
  This follows a successful Schistosomiasis project.
- The Validation Science Exchange's Reproducibility Initiative[2]
- Cancer Studies — 50 "most impactful" from 2010-2012[3]

1: Todd and others (2015) (Matt Todd & others), OSBR (2015)
2: Iorns and others (2015) (Iorns & others)
3: Errington et al. (2014); Kaiser (2015), in June 26 2015 *Science*

Hard to Reproduce Reporting

Steps in preparing a report or paper include:

- Microsoft Word or LaTeX or . . .
- Code files: one for each table, graph, . . .
- Data files
- Requires work to adapt code to new data, or . . .
- Manual steps required are prone to induce errors
- Difficult to retrace steps a year or more later.

Reproducible Reporting

- One file for text, code and data or data access
- To make revisions, just revise and re-process
- Easy for a later worker to check code & re-use
- Allows a consultant to know just what was done
- Ideal setup:
  - Code in the file accesses packaged code
  - Check help files to know what code does
  - Data is likewise from a 'packaged' source (Such sources are called 'databases')
- Both R and Python can be used, others?

Reporting with R Markdown

- Very simple to use with RStudio
    - 10 minutes tuition can get one started
- For finer control, can use HTML markup
- Can output to HTML or PDF or Word or . . .

Slides

- Slides for this talk (pdf + R Markdown sources) will be posted at: http://maths-people.anu.edu.au/~johnm/stats-issues/
- The technology use for the R Markdown sources for this talk can also be used for reports and papers
  - Simple text markup: use R Markdown
  - Sophisticated: Use Sweave = LaTeX with markup
  - Further possibility: HTML with markup
  - At least one CRI is using Sweave for reports

The R Markdown file is http://maths-people.anu.edu.au/~johnm/stats-issues/shiftingLandscape.Rmd Also from the directory http://maths-people.anu.edu.au/~johnm/stats-issues/ put the files RR.RData (has the Reinhart and Rogoff data) and osc1.bib in the same directory. It is easiest to process the file shiftingLandscape.Rmd from an RStudio session; click on the 'File' dropdown menu, then on 'Open File'. Packages that will need to be installed are latticeExtra, ggplot2, and DAAG. Once they are in place, click on 'Knit HTML'. Or click on the down arrow to the right if you wish to output to Word. With slides, the list of references gets truncated after one slide, and I had to take separate steps to put them all back into the pdf that I created. (For this, click on 'Open in Browser'. Then, from the browser, print to a PDF.

References

Begley, C. Glenn. 2013. “Reproducibility: Six Red Flags for Suspect Work.” *Nature* 497 (7450): 433–34.

Begley, C. Glenn, and Lee M. Ellis. 2012. “Drug Development: Raise Standards for Preclinical Cancer Research.” *Nature* 483 (7391): 531–33.

Collins, Francis S., and Lawrence A. Tabak. 2014. “Policy: NIH Plans to Enhance Reproducibility.” *Nature* 505 (7485): 612–13.

Errington, Timothy M, Elizabeth Iorns, William Gunn, Fraser Elisabeth Tan, Joelle Lomax, and Brian A Nosek. 2014. “An Open Investigation of the Reproducibility of Cancer Biology Research.” *ELife* 3.

Fisher, Ronald Aylmer. 1926. “The Arrangement of Field Experiments.” *Journal of the Ministry of Agriculture GB* 33: 503–13.

———. 1937. *The Design of Experiments*. 2nd ed. Oliver; Boyd.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLoS Med* 2 (8): e124.

Iorns, Elizabeth, and others. 2015. “Validation by Science Exchange - Identifying and Rewarding High-Quality Research.” validation.scienceexchange.com.

Kaiser, Jocelyn. 2015. “The Cancer Test.” *Science* 348 (6242): 1411–3.

Kriegeskorte, Nikolaus, W Kyle Simmons, Patrick S F Bellgowan, and Chris I Baker. 2009. “Circular Analysis in Systems Neuroscience: The Dangers of Double Dipping.” *Nature Neuroscience* 12 (5): 535–40.

OSBR. 2015. “The Synaptic Leap: Open Source Biomedical Research.”

OSC. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): ’aac4716–1’–’aac4716–7’.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. “Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?” *Nature Reviews Drug Discovery* 10 (9): 712–12.

Todd, Mat, and others. 2015. “OSM - Open Source Malaria.” *Opensourcemalaria.org*.