Open Science and reproducibility

Evan Johnson

2025-01-15

Problematic phenotypes in science (and business)

(Bild et al., PLOS Biology, 2014)

The Farmer

"Let's harvest our data or invent our tool and then figure out a plan."

- Farmer gathers bushels of data and asks: "What now?"
- May forget to budget for seeds, farmhands or tools needed for harvest
- Develops a new tool for harvesting: "What should we do with it?"

Data or tool might lose usefulness before an appropriate plan is identified.

(Bild et al., PLOS Biology, 2014)

The Miner

"If we keep digging, we will eventually find what we are looking for."

- 'Siren song of a sweet spot'
- Under high pressure: grant due, papers to publish, or complete thesis
- Modifying thresholds, changing parameters, and re-analyzing the data using different algorithms
- Trying multiple datasets until you find one that works!

Gold Miners must beware the allure of fools gold!

(Bild et al., PLOS Biology, 2014)

The Cowboy and the Sheriff

"We don't really understand the data, but we will go ahead and publish!"

- 'Shoot first, ask questions later'
- 'Wrangles' data to support predefined hypothesis
- Fails to account for confounding variables
- Cherry-picks genes without validation
- Inappropriate statistical algorithms or ad hoc data analysis
- Makes inappropriate conclusions

Only a matter of time before the Sheriff hauls him off to the brig!

(Bild et al., PLOS Biology, 2014)

The Hermit

"I don't need help from anyone else."

- Blinded by distrust, over-optimism, or sense of scientific superiority
- Can solve important problems without help from inferior 'applied' scientists
- Technology development, data generation, or analysis are not superior or inferior steps
- Share authorship and write grants to include technology or analytic tools

Need for interdisciplinary research will relegate the Hermit to obscurity.

(Bild et al., PLOS Biology, 2014)

The Master and the Servant

"I saw this in a talk or manuscript. Can you do the same for me?"

- Master needs for new expertise in the lab, so recruits Servants
- Master lacks knowledge to train servants or produce quality research in those fields.
- Slave has incomplete training or lacks commitment to fully understand the task

It's time to educate the Master and free the Slave.

(Bild et al., PLOS Biology, 2014)

The Warden

"We'll keep our data or code, thank you."

- Jailor seeks to retain full control over precious data
- Jailor fails to produce usable software for others to use novel method
- Subjects donated samples and taxpayers provide funding to advance science–not to benefit jailor's career

The data and code didn't do anything wrong, don't keep them locked up!

(Bild et al., PLOS Biology, 2014)

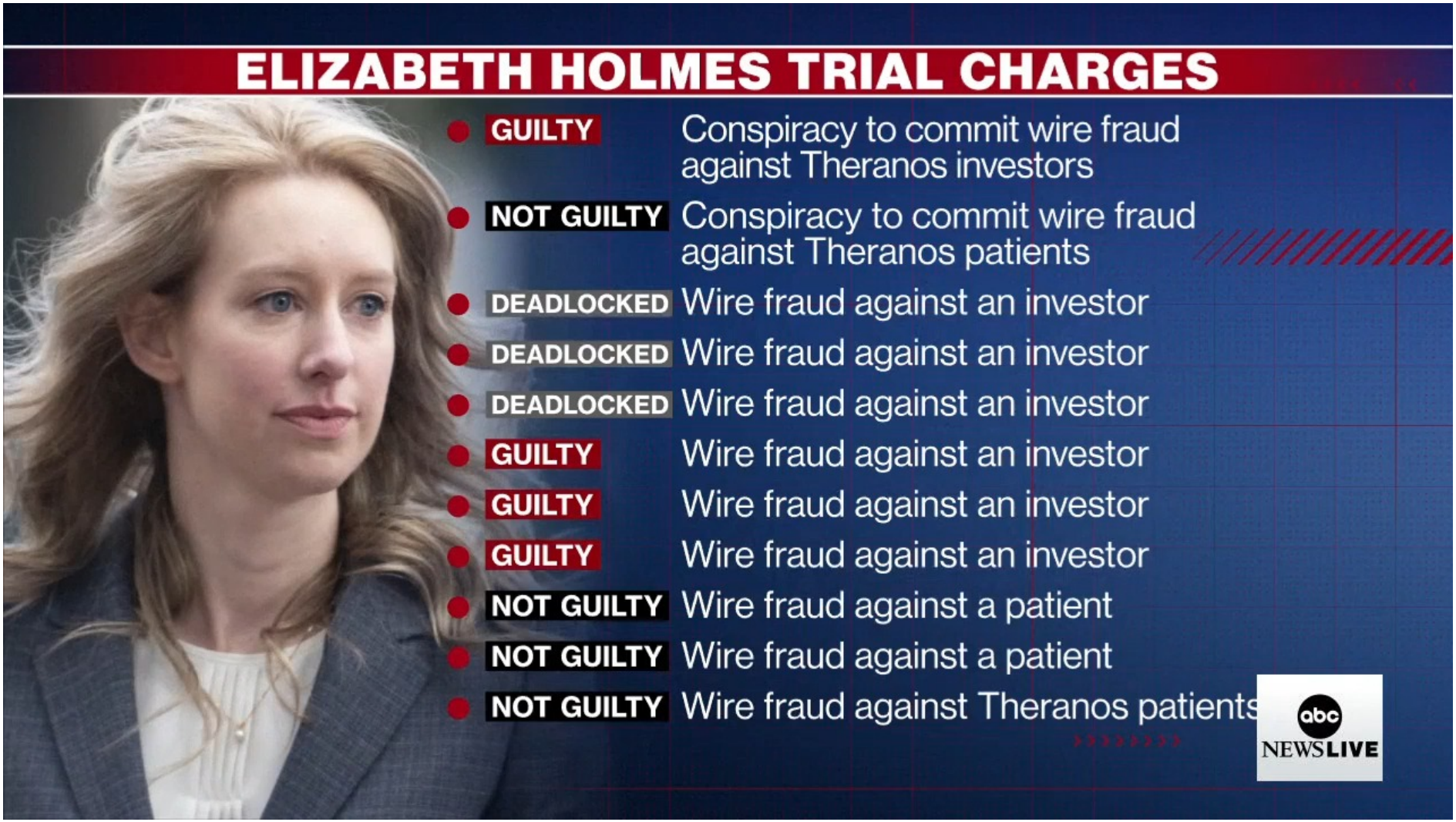
Open and Reproducible Science

Much of the rest of this material is taken from Dr. Verena Heise, her original slides can be found at: https://osf.io/f9xbz/.

I would also like to credit Dr. Prasad Patil at Boston University for his contributions to this side deck, and I would also like to acknowledge a litte help from ChatGPT!

What is the problem?

Open and Reproducible Science



Reproducibility Crisis

Letter | Published: 27 August 2018

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek ✉, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers & Hang Wu

Nature Human Behaviour **2**, 637–644 (2018) | Download Citation ⬇

original studies. We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size. Replicability varies

Reproducibility Crisis

New Online | Views 13,810 | Citations 0 | Altmetric 391 | Comments

Viewpoint

ONLINE FIRST FREE

January 6, 2020

Challenges to the Reproducibility of Machine Learning Models in Health Care

Andrew L. Beam, PhD[1,2]; Arjun K. Manrai, PhD[2,3]; Marzyeh Ghassemi, PhD[4,5]

» Author Affiliations | Article Information

JAMA. Published online January 6, 2020. doi:10.1001/jama.2019.20866

Reproducibility Crisis

Editorial | Open Access | Published: 21 February 2020

No raw data, no science: another possible source of the reproducibility crisis

Tsuyoshi Miyakawa

Molecular Brain **13**, Article number: 24 (2020) | Cite this article

44k Accesses | **35** Citations | **2257** Altmetric | Metrics

Below is an example of the requests I have made to authors:

"Before proceeding, please do the following:

1. Attach raw data (all the images for entire membranes of western blotting with size markers and for staining, quantified numerical data for each sample used for statistical analyses, etc.) as supplementary materials.

2. Provide absolute p-values, instead of expressions like $p < 0.05$, in the results.

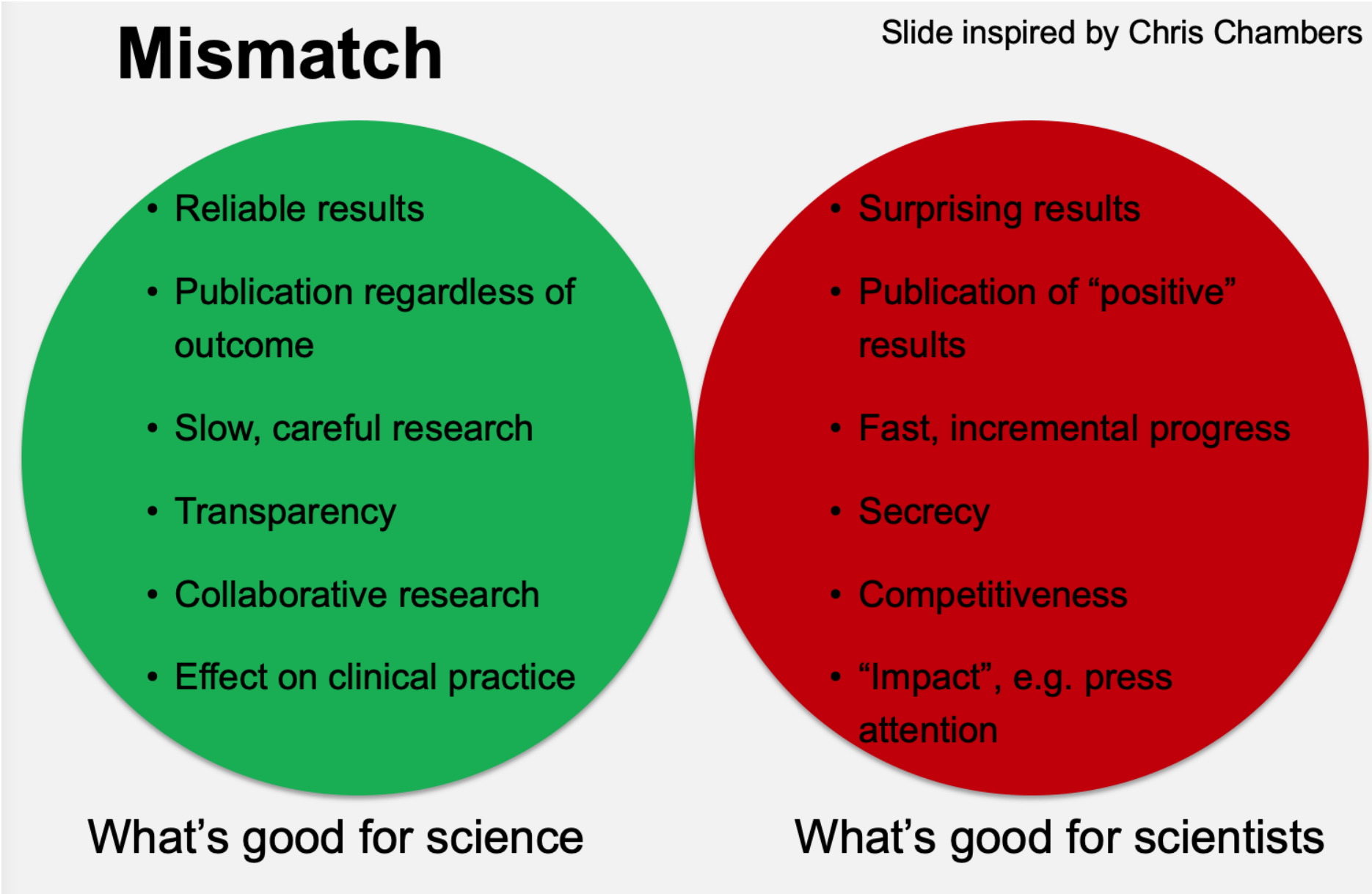
Reproducibility Crisis

Some Terminology:

- Reproducible: a result can be recreated by others using the same data and analysis pipelines
- Replicable: same scientific conclusion reached using independent data and (maybe) independent analysis pipelines

$\Rightarrow$ The Reproducibility crisis is a Replication crisis

Reproducibility Crisis



Reproducibility Crisis

"Many attacks have lately been made on the conduct of various scientific bodies, and of their officers, and severe criticism has been lavished upon some of their productions. Newspapers, Magazines, Reviews, and Pamphlets, have all been put in requisition for the purpose."

Reproducibility Crisis

"Many attacks have lately been made on the conduct of various scientific bodies, and of their officers, and severe criticism has been lavished upon some of their productions. Newspapers, Magazines, Reviews, and Pamphlets, have all been put in requisition for the purpose."

Charles Babbage
Reflections on the Decline of Science in England, and on Some of its Causes
1830

Reproducibility Crisis

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration[*,†]
+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

Article | Figures & Data | Info & Metrics | eLetters | PDF

Empirically analyzing empirical evidence

One of the central goals in any scientific endeavor is to understand causality. Experiments that seek to demonstrate a cause/effect relation most often manipulate the postulated causal factor. Aarts *et al.* describe the replication of 100 experiments reported in papers published in 2008 in three high-ranking psychology journals. Assessing whether the replication and the original experiment yielded the same result according to several criteria, they find that about one-third to one-half of the original findings were also observed in the replication study.

Science, this issue **10.1126/science.aac4716**

OSC (Aug 2015)

- 100 pairs of original studies and replication efforts.
- Many metrics presented, both quantitative and qualitative.
  - P-value comparison
  - 95% CI membership
  - Average effect size
  - Questionnaires posed to replicators

OSC (Aug 2015)

- 100 pairs of original studies and re
- M
  - qu
  -
  -
  -
  - questionnaires posed to replicators

SCIENCE

How Reliable Are Psychology Studies?

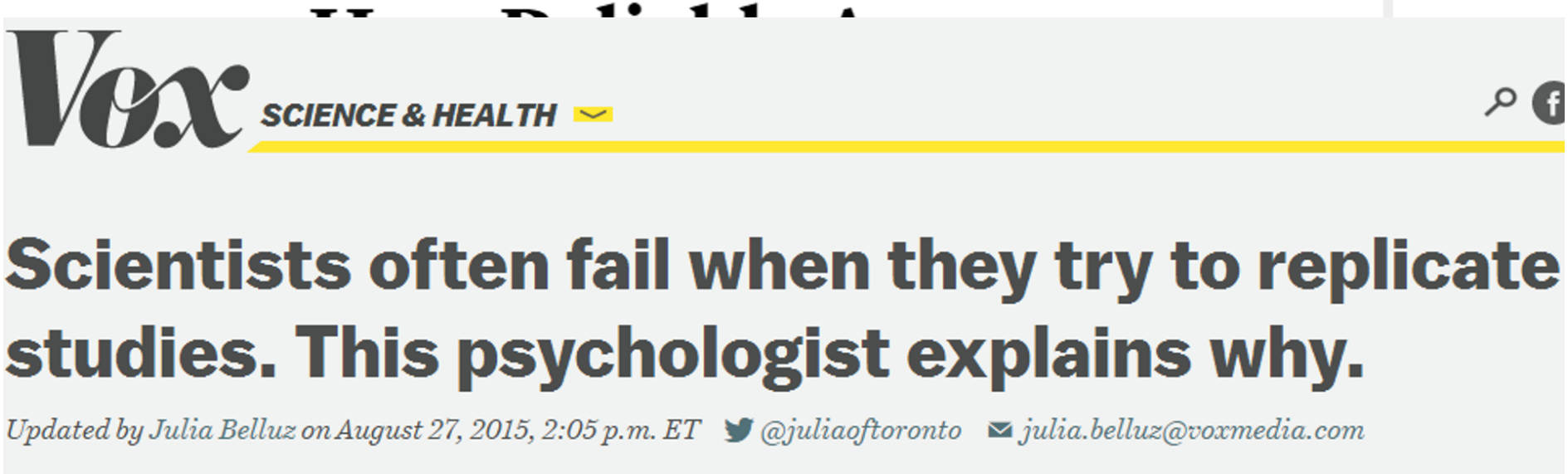
A new study shows that the field suffers from a reproducibility problem, but the extent of the issue is still hard to nail down.

5.4k 1.5k

ED YONG | AUG 27, 2015

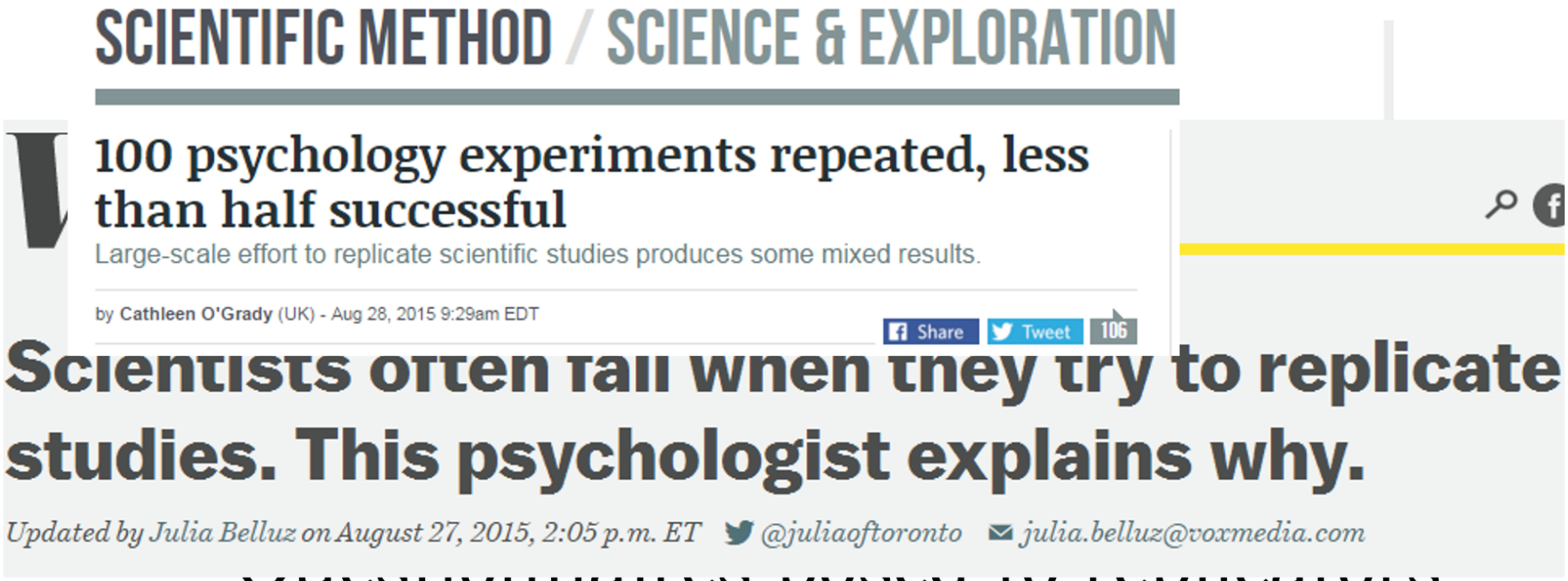
OSC (Aug 2015)

- 100 pairs of original studies and re



OSC (Aug 2015)

- 100 pairs of original studies and



SCIENTIFIC METHOD / SCIENCE & EXPLORATION

100 psychology experiments repeated, less than half successful

Large-scale effort to replicate scientific studies produces some mixed results.

by Cathleen O'Grady (UK) - Aug 28, 2015 9:29am EDT

Share Tweet 106

Scientists often fail when they try to replicate studies. This psychologist explains why.

Updated by Julia Belluz on August 27, 2015, 2:05 p.m. ET @juliaoftoronto julia.belluz@voxmedia.com

questionnaires posed to replicators

Reproducibility Crisis

OSC (Aug 2015)

- 100 pairs of original studies and

SCIENTIFIC METHOD / SCIENCE & EXPLORATION

100 psychology experiments repeated, less

Speaking of Science

Many scientific studies can't be replicated. That's a problem.

A | 398

By Joel Achenbach August 27 Follow @joelachenbach

Most Read

OSC (Aug 2015)

SCIENCE

Many Psychology Findings Not as Strong as Claimed, Study Says

By BENEDICT CAREY AUG. 27, 2015

Speaking of Science

Many scientific
That's a problen

Massive International Project Raises Questions about the Validity of Psychology Research

When 100 past studies were replicated, only 39 percent yielded the same results

By Roni Jacobson | August 27, 2015 | Véalo en español

A 398

By Joel Achenbach August 27 Follow @joelachenbach

Most Read

OSC (Aug 2015)

SCIENCE

Many Psychology Findings Not as Strong as Claimed, Study Says

SCIENCE

The Reproducibility Crisis: Cognitive Scientist, Heal Thyself

nternational Project Raises s about the Validity of gy Research

When 100 past studies were replicated, only 39 percent yielded the same results

By Roni Jacobson | August 27, 2015 | Véalo en español

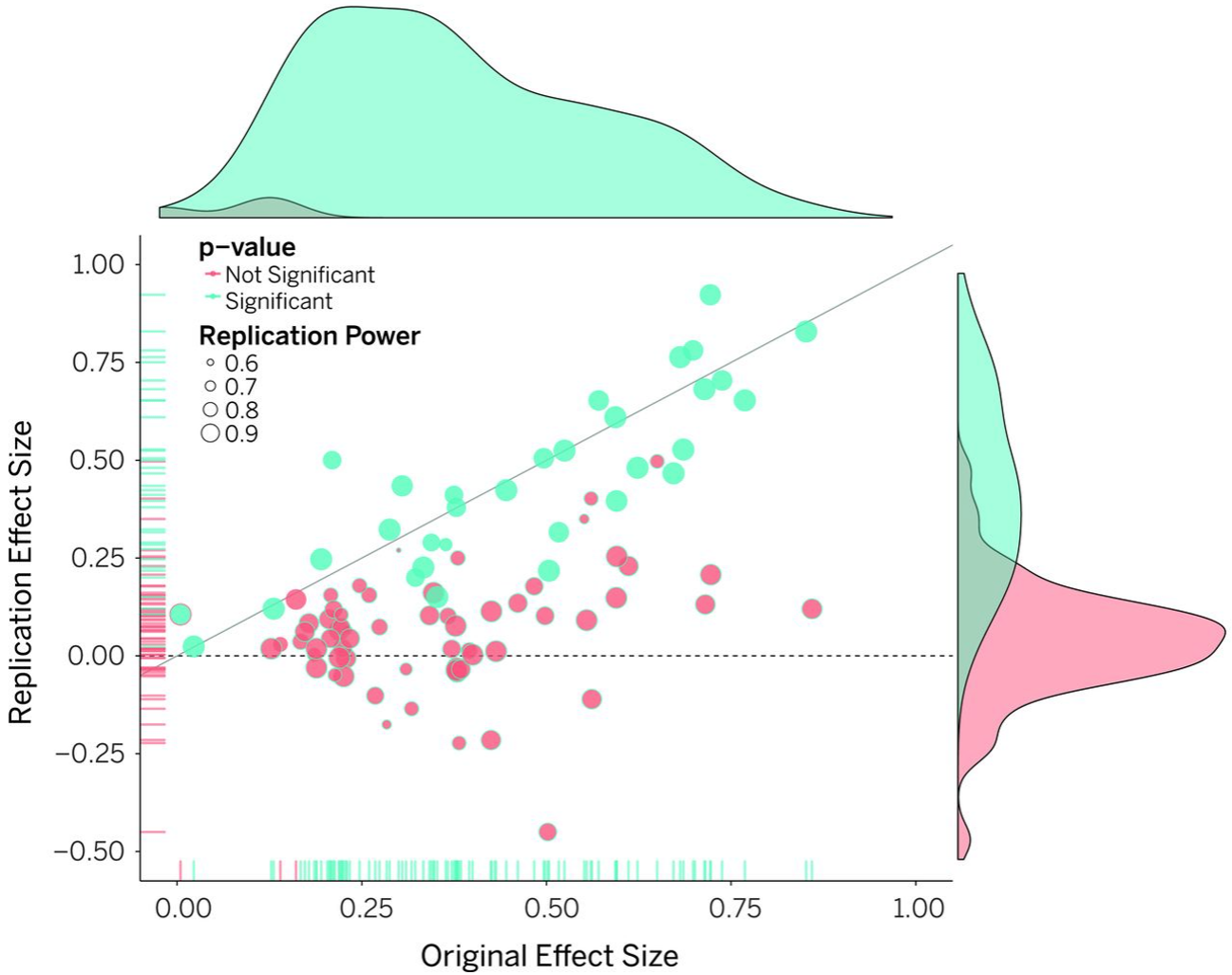
A 398

By Joel Achenbach August 27 Follow @joelachenbach

Most Read

Reproducibility Crisis

Original study effect size versus replication effect size (correlation coefficients)



Open Science Collaboration Science 2015;349:aac4716

Reproducibility Crisis

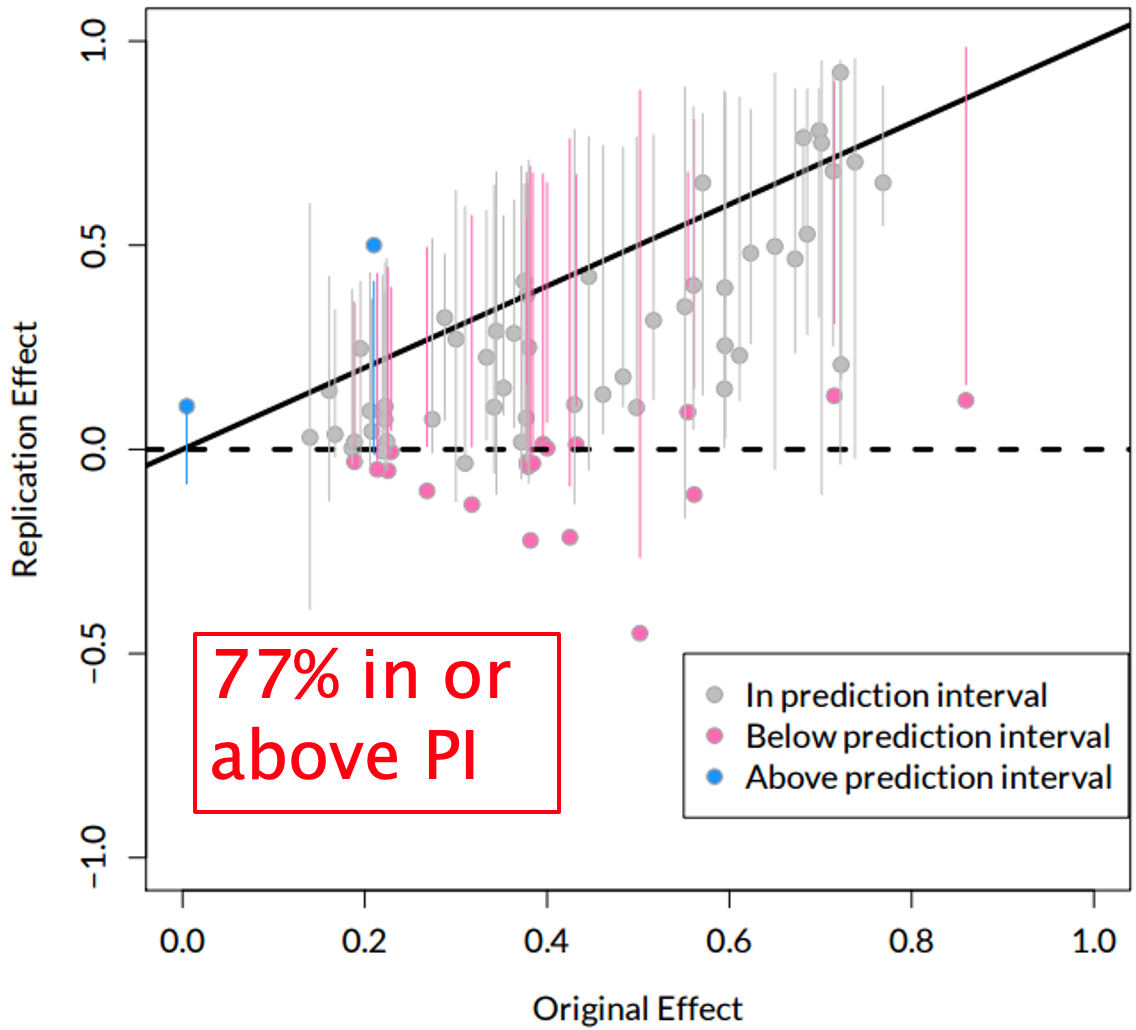
$$\hat{\theta}_{orig} \cong \hat{\theta}_{rep} \Rightarrow \text{variation}$$

Robustness/Sensitivity

$$\hat{\theta}_{orig} \cong \hat{\theta}_{rep} \mid \text{variation}$$

Replicability

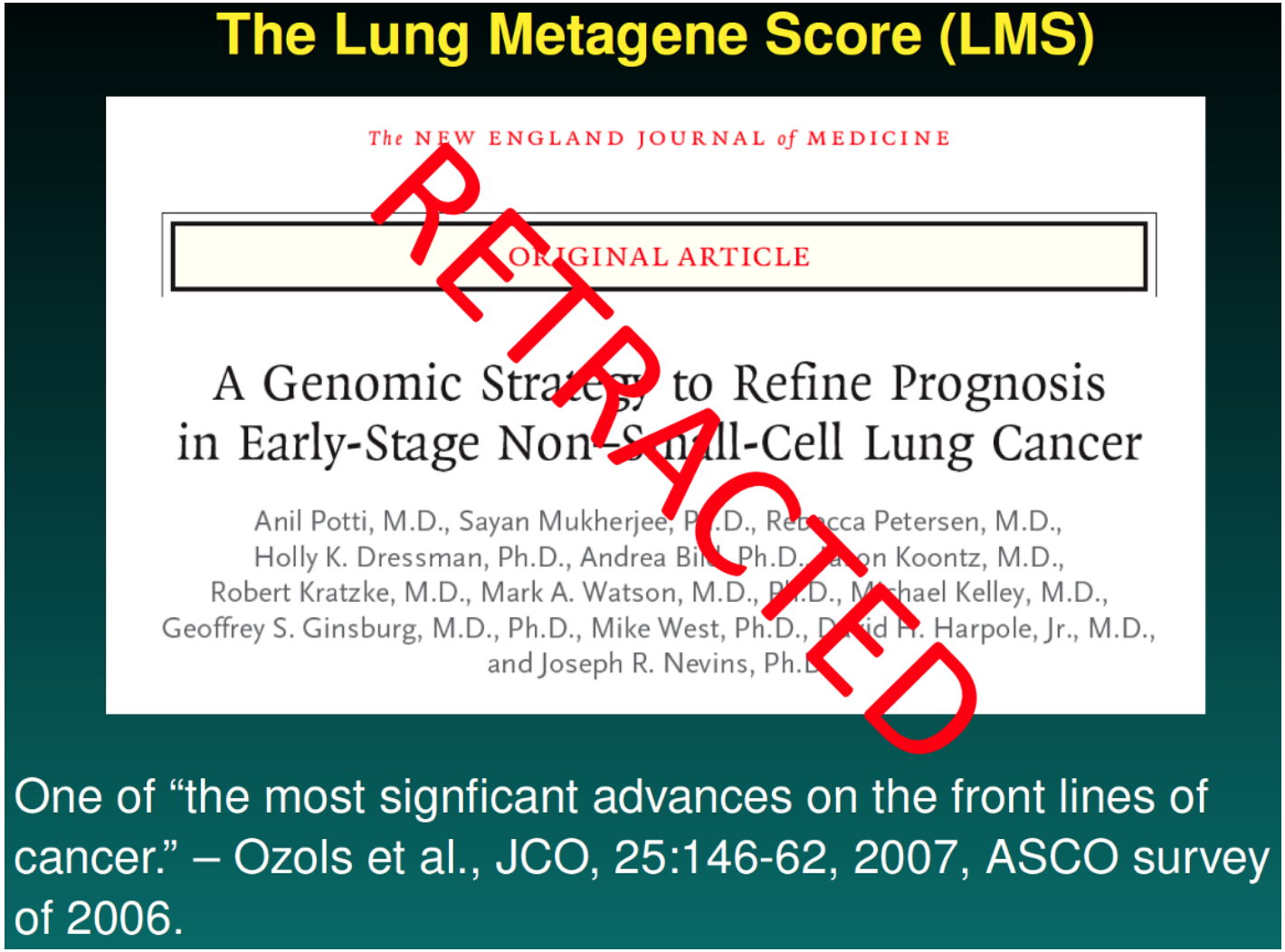
Reproducibility Crisis



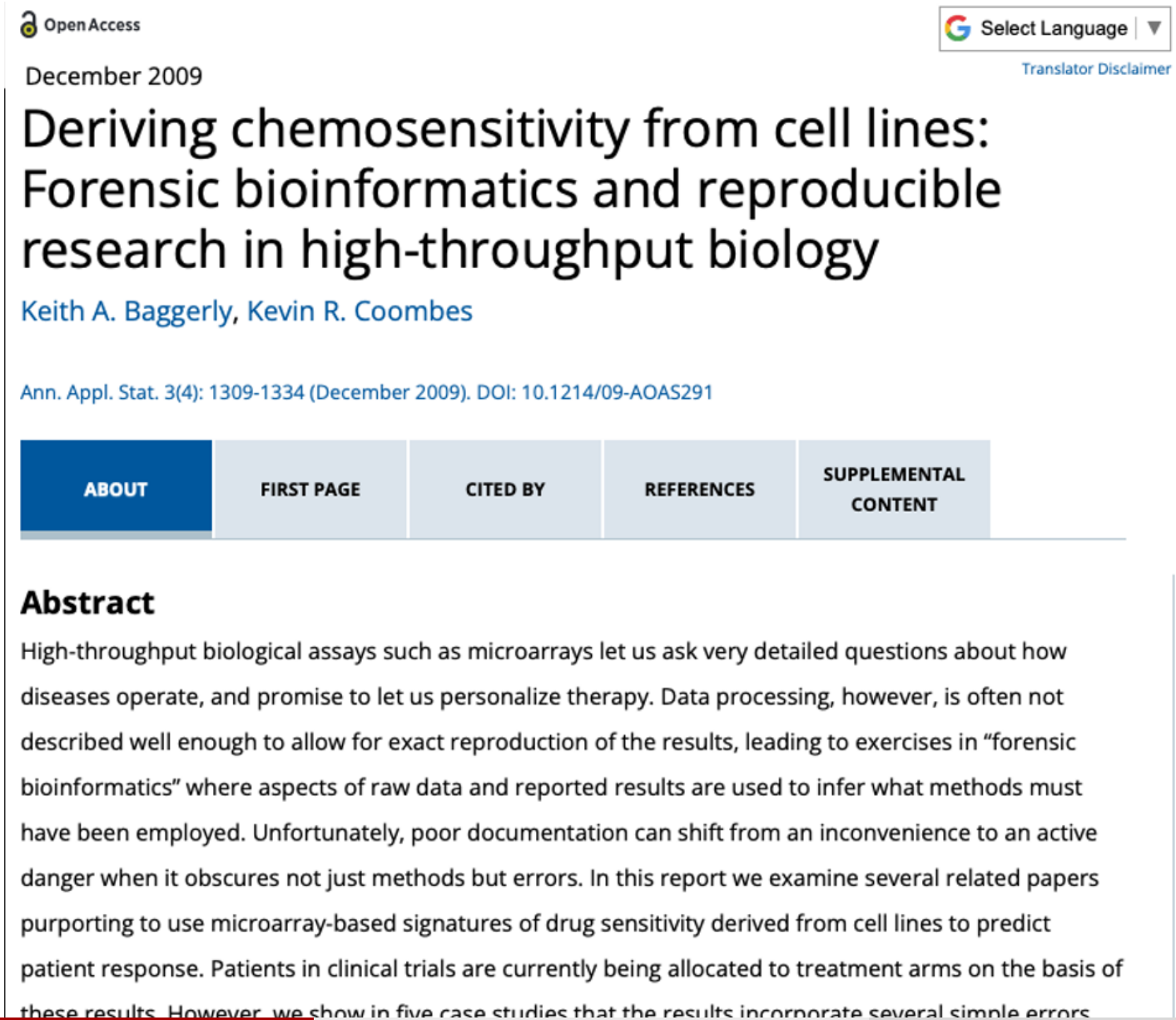
Patil, P., Peng, R. D., & Leek, J. T. (2016). *Perspectives on Psychological Science*, *11*(4), 539–544.

Why do we care?

Retracted Science



Retracted Science



Open Access

Select Language

Translator Disclaimer

December 2009

Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology

Keith A. Baggerly, Kevin R. Coombes

Ann. Appl. Stat. 3(4): 1309-1334 (December 2009). DOI: 10.1214/09-AOAS291

ABOUT | FIRST PAGE | CITED BY | REFERENCES | SUPPLEMENTAL CONTENT

Abstract

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors

Non-Reproducible Science

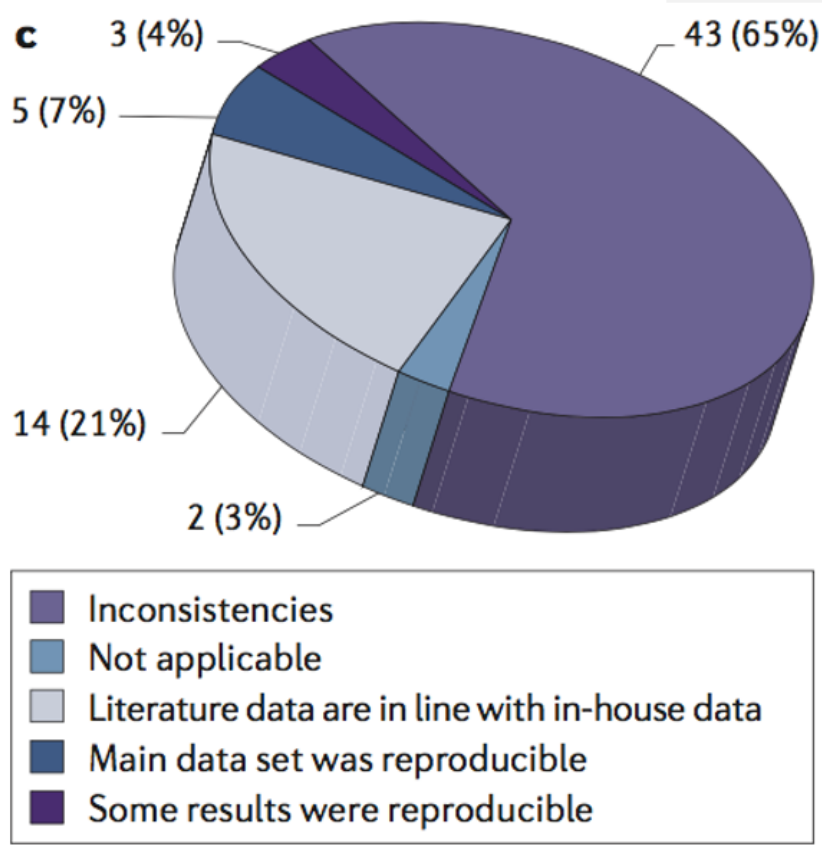
Correspondence

Nature Reviews Drug Discovery **10**, 712 (September 2011) | doi:10.1038/nrd3439-c1

Believe it or not: how much can we rely on published data on potential drug targets?

See also: News and Analysis by Arrowsmith

Florian Prinz[1], Thomas Schlange[2] & Khusru Asadullah[3]



Retracted Science

Real-life consequences

Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis

Mandeep R Mehra, Sapan S Desai, Frank Ruschitzka, Amit N Patel

The Lancet

RETRACTED

New England Journal of Medicine

ORIGINAL ARTICLE

Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19

Mandeep R. Mehra, M.D., Sapan S. Desai, M.B.A., Ph.D., SreyRam Kuy, M.D., M.H.S., Timothy D. Henry, M.D., and Amit N. Patel, M.D.

Retracted Science

Real-life consequences

The Guardian

WHO halts hydroxychloroquine trial for coronavirus amid safety fears

Malaria drug taken by Trump could raise risk of death and heart problems, study shows

77 clinical trials were launched testing hydroxychloroquine, and all failed, many terminated early

Reproducible research

Markowetz *Genome Biology* (2015) 16:274
DOI 10.1186/s13059-015-0850-7

COMMENT Open Access

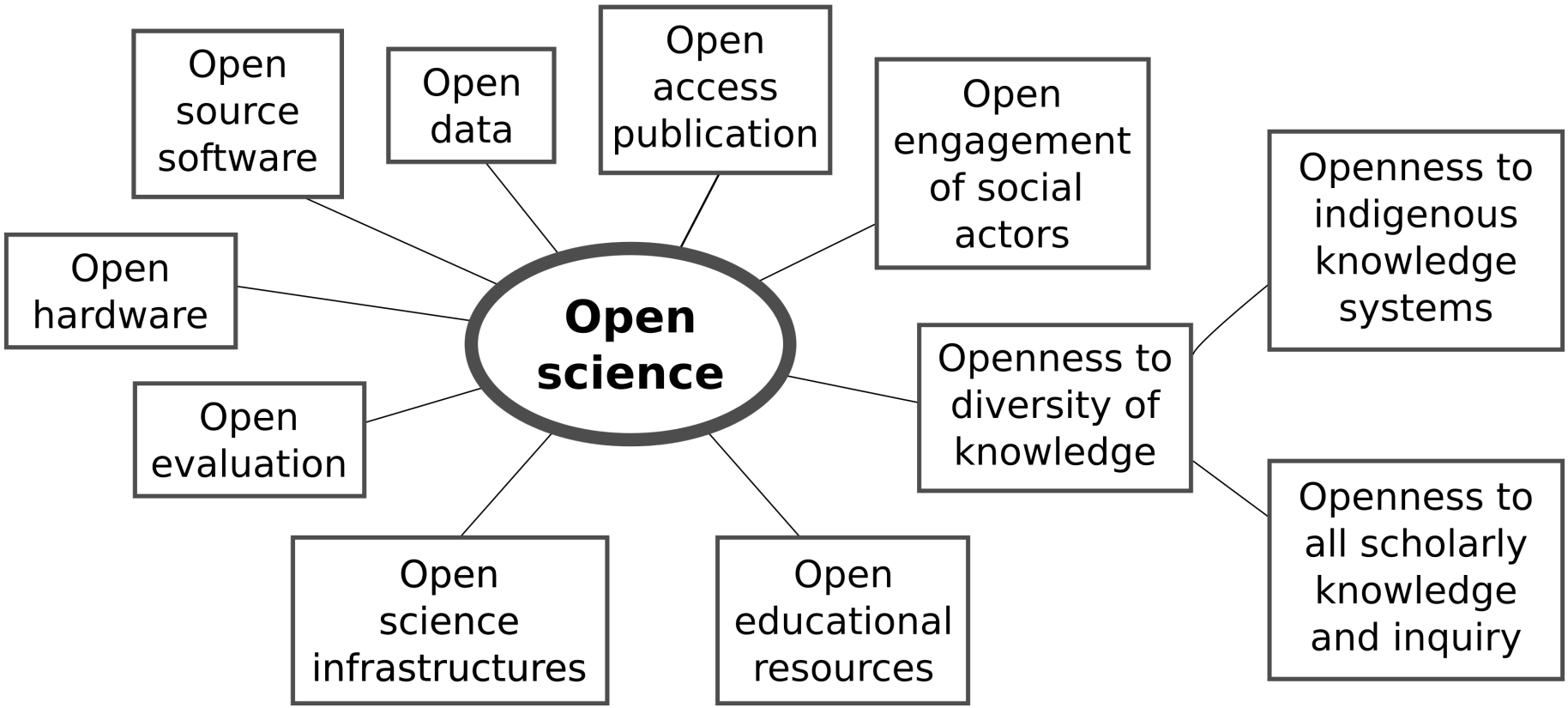
Five selfish reasons to work reproducibly

Florian Markowetz

1. Reproducibility helps to avoid disaster
2. Reproducibility makes it easier to write papers
3. Reproducibility helps reviewers see it your way
4. Reproducibility enables continuity of your work
5. Reproducibility helps to build your reputation

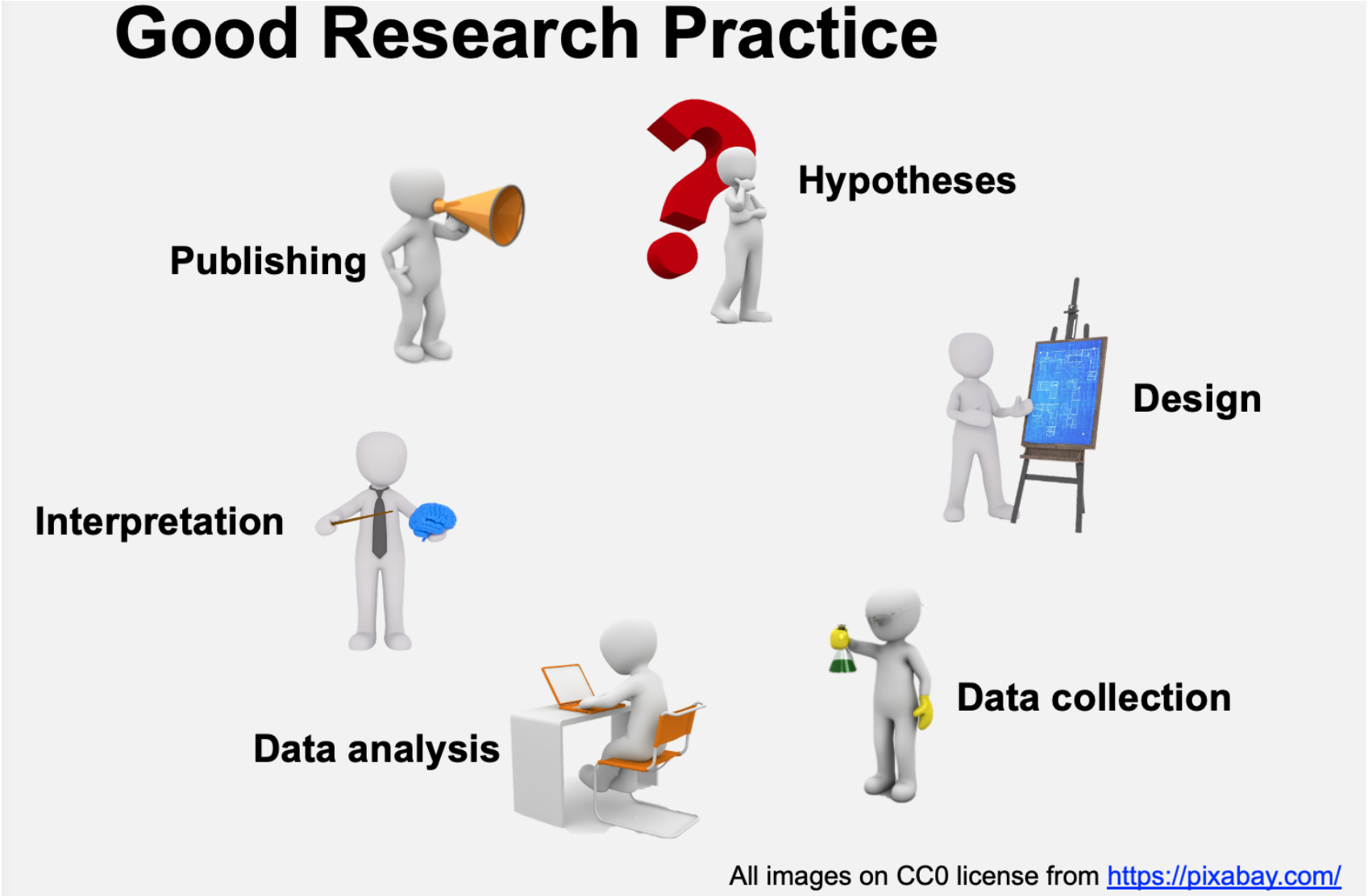
Potential solutions?

Open Science



(Open science elements based on UNESCO presentation of 17 February 2021)

Good research practice



Some cool tools for reproducible science

- **RMarkdown/Jupyter Notebooks**: Interactive abd dynampic documents that combine code (R or Python), visualizations, narrative text, and output for fully reproducible reports.
- **R/Shiny**: Build interactive web applications for data analysis and visualization directly in R. Allows for building real-time dashboards for data exploration.
- **GitHub**: Version control and collaboration for code, documents, and workflows. Share and track changes to your research scripts.
- **Nextflow/Snakemake**: Workflow management for scalable and portable pipelines and automate multi-step computational analyses across platforms.
- **Docker/Singularity**: Containerization for consistent computational environments to ensure your code runs anywhere.

Open Science Practices (Johnson Lab)

Data, Code and Software

Data: Shared in website links, public repositories (preferred)

Code: Websites, supplementary files, GitHub (preferred)

Software: Make an R package! (or Python, etc.) Share on GitHub or Bioconductor

Open Science Practices (Johnson Lab)

Using R Markdown

BMC Part of Springer Nature

BMC Infectious Diseases

Home About Articles Submission Guidelines In Review

Research article | Open Access | Published: 22 January 2021

Comparing tuberculosis gene signatures in malnourished individuals using the TBSignatureProfiler

W. Evan Johnson, Aubrey Odom, Chelsie Cintron, Mutharaj Muthaiah, Selby Knudsen, Noyal Joseph, Senbagavalli Babu, Subitha Lakshminarayanan, David F. Jenkins, Yue Zhao, Ethel Nankya, C. Robert Horsburgh, Gautam Roy, Jerrold Ellner, Sonali Sarkar, Padmini Salgame & Natasha S. Hochberg

BMC Infectious Diseases **21**, Article number: 106 (2021) | Cite this article

4281 Accesses | **1** Citations | **4** Altmetric | Metrics

Abstract

Background

Gene expression signatures have been used as biomarkers of tuberculosis (TB) risk and outcomes. Platforms are needed to simplify access to these signatures and determine their validity in the setting of comorbidities. We developed a computational profiling platform of TB signature gene sets and characterized the diagnostic ability of existing signature gene sets to differentiate active TB from LTBI in the setting of malnutrition.

R Code: https://github.com/wevanjohnson/tbsp_malnutrition

Software: https://github.com/compbiomed/TBSignatureProfiler
https://bioconductor.org/packages/release/bioc/html/TBSignatureProfiler.html

R Code: https://github.com/wevanjohnson/tbsp_malnutrition Software: https://github.com/compbiomed/TBSignatureProfiler
https://bioconductor.org/packages/release/bioc/html/TBSignatureProfiler.html

Open Science Practices (Johnson Lab)

Using R Markdow

BMC Part of Springer Nature

BMC Infectious Diseases

Home About Articles Submission Guidelines In Review

Research article | Open Access | Published: 22 January 2021

Comparing tuberculosis gene signatures in malnourished individuals using the TBSignatureProfiler

W. Evan Johnson, Aubrey Odom, Chelsie Cintron, Mutharaj Muthaiah, Selby Knudsen, Noyal Joseph, David F. Jenkins, Yue Zhao, Ethel Nankya, C. Robert Sarkar, Padmini Salgame & Natasha S. Hochberg

06 (2021) | Cite this article

Metrics

Availability of data and materials

The raw and processed sequencing data from this study are available in the GEO repository, under accession numbers GSE101705 and GSE152218. Furthermore, processed sequencing data and R code used for analysis and figure generation is available in the following GitHub repository: https://github.com/wevanjohnson/tbsp_malnutrition. The TBSignatureProfiler software is available through Bioconductor (https://bioconductor.org/packages/release/bioc/html/TBSignatureProfiler.html) and GitHub (https://github.com/compbiomed/TBSignatureProfiler).

as biomarkers of tuberculosis (TB) risk and outcomes. Platforms are needed to simplify access to these signatures and determine their validity in the setting of comorbidities. We developed a computational profiling platform of TB signature gene sets and characterized the diagnostic ability of existing signature gene sets to differentiate active TB from LTBI in the setting of malnutrition.

R Code: https://github.com/wevanjohnson/tbsp_malnutrition

Software: https://github.com/compbiomed/TBSignatureProfiler
https://bioconductor.org/packages/release/bioc/html/TBSignatureProfiler.html

Data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE101705 and
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4609421
RMarkdown: https://github.com/wevanjohnson/tbsp_malnutrition

Open Science Practices (Johnson Lab)

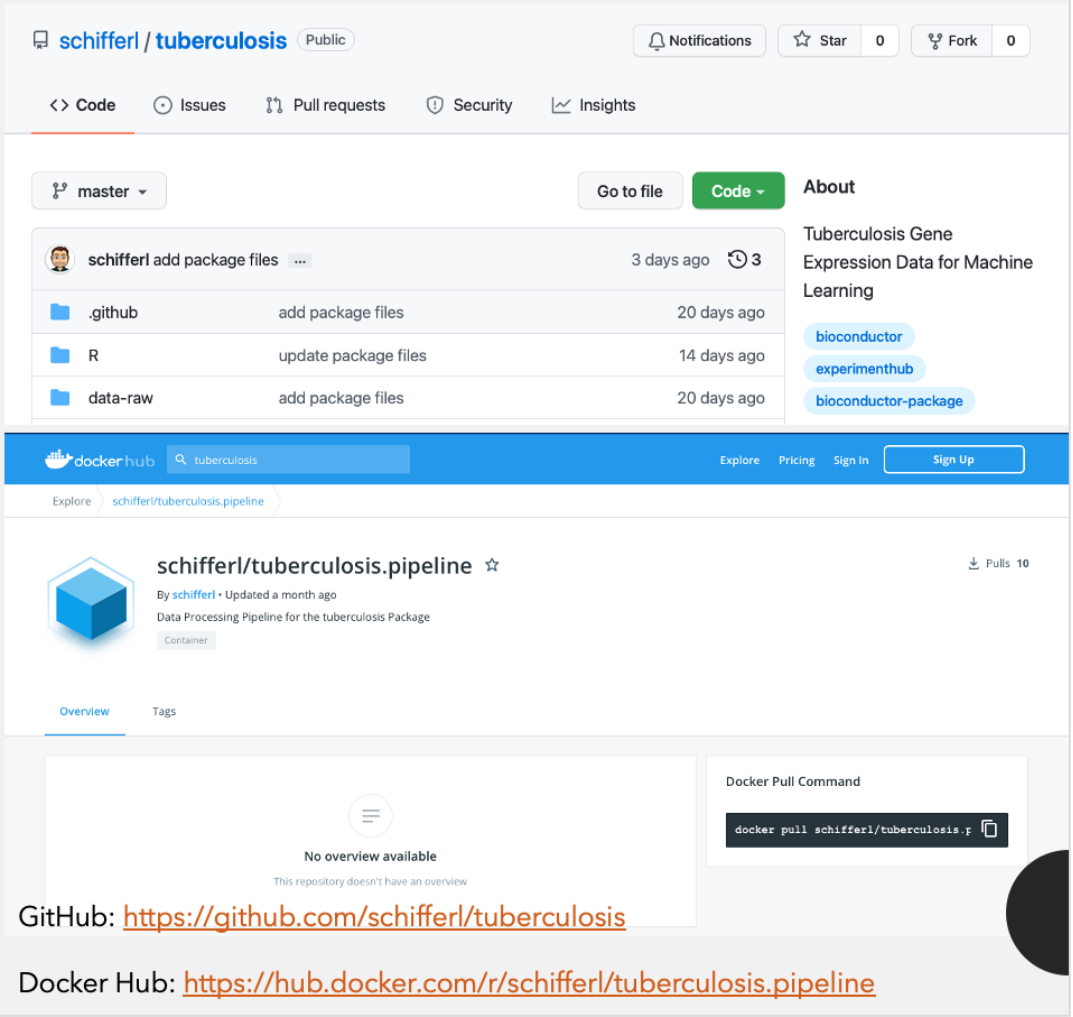
Using R Shiny

https://carbohydratequalitymicrobiome.shinyapps.io/supplementalapp/

Shiny app: https://carbohydratequalitymicrobiome.shinyapps.io/supplementalapp/

Open Science Practices (Johnson Lab)

Docker and Docker Hub



GitHub: https://github.com/schifferl/tuberculosis

Docker Hub: https://hub.docker.com/r/schifferl/tuberculosis.pipeline

GitHub: https://github.com/schifferl/tuberculosis
Docker Hub: https://hub.docker.com/r/schifferl/tuberculosis.pipeline

What is Nextflow?

- A framework for writing and executing computational pipelines.
- Key features:
  - Supports **parallel** and **distributed** computing.
  - Works with multiple execution platforms (e.g., local, cloud, and HPC).
  - Script-based: Pipelines written in a domain-specific language (DSL) extending **Groovy**.

```
process example_process {
    input:
    path input_file
    output:
    path output_file
    """
    cat $input_file > $output_file
    """
}
workflow {
    example_process(input_file: "data.txt", output_file: "result.txt")
}
```

What is Docker?

- A platform for creating, deploying, and managing lightweight, portable containers.
- Why Docker?
  - Consistent environments across development, testing, and production.
  - Isolation of dependencies and libraries.
- Key terms:
  - Image: A snapshot of a container.
  - Container: A running instance of an image.

What is Docker?

```
# Example Dockerfile
FROM ubuntu:20.04
RUN apt-get update && apt-get install -y python3
CMD ["python3", "--version"]
```

```
# Building and running a container
docker build -t my-python .
docker run my-python
```

Integrating Nextflow and Docker

- Nextflow supports Docker natively for containerizing processes.
- Benefits:
  - Ensures pipeline portability.
  - Simplifies dependency management.

```
process container_example {
    container 'ubuntu:20.04'
    script:
    """
    echo "Running in a container!"
    """
}
```

Advantages of Using Nextflow and Docker

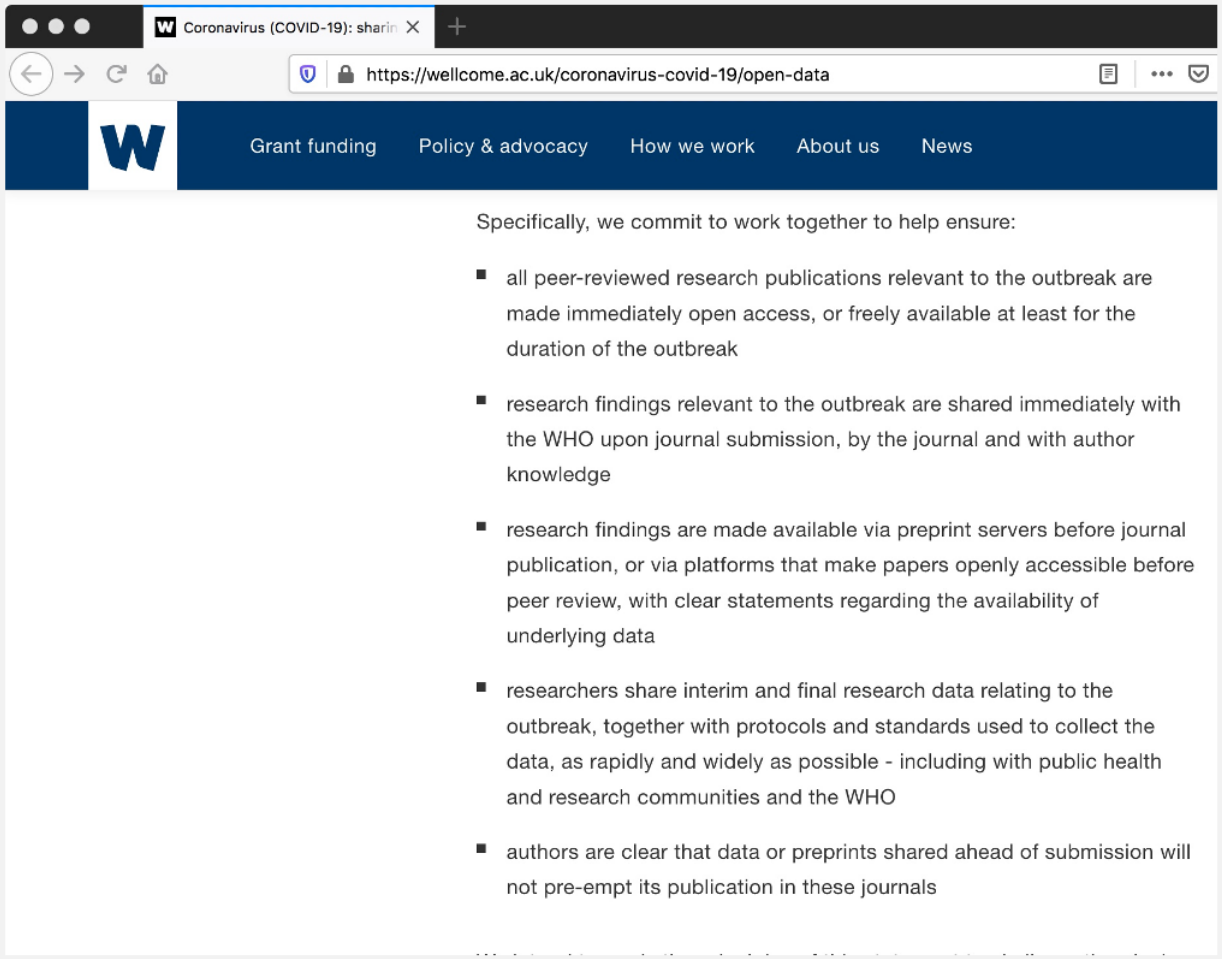
1. **Reproducibility:** Ensure pipelines run the same everywhere.
2. **Scalability:** Efficiently scale to cloud or HPC environments.
3. **Portability:** Share workflows and containers effortlessly.

Resources

- Nextflow Documentation
- Docker Documentation

Open Science Practices

OS and the pandemic



Open Science Publishing

Publishing

- Peer-reviewed publications online, freely available to read
- Authors keep copyright to their own work

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

- Preprint of finished manuscript before peer review

Open Science Publishing

Open Source “hardware”

PLOS BIOLOGY

OPEN ACCESS

COMMUNITY PAGE

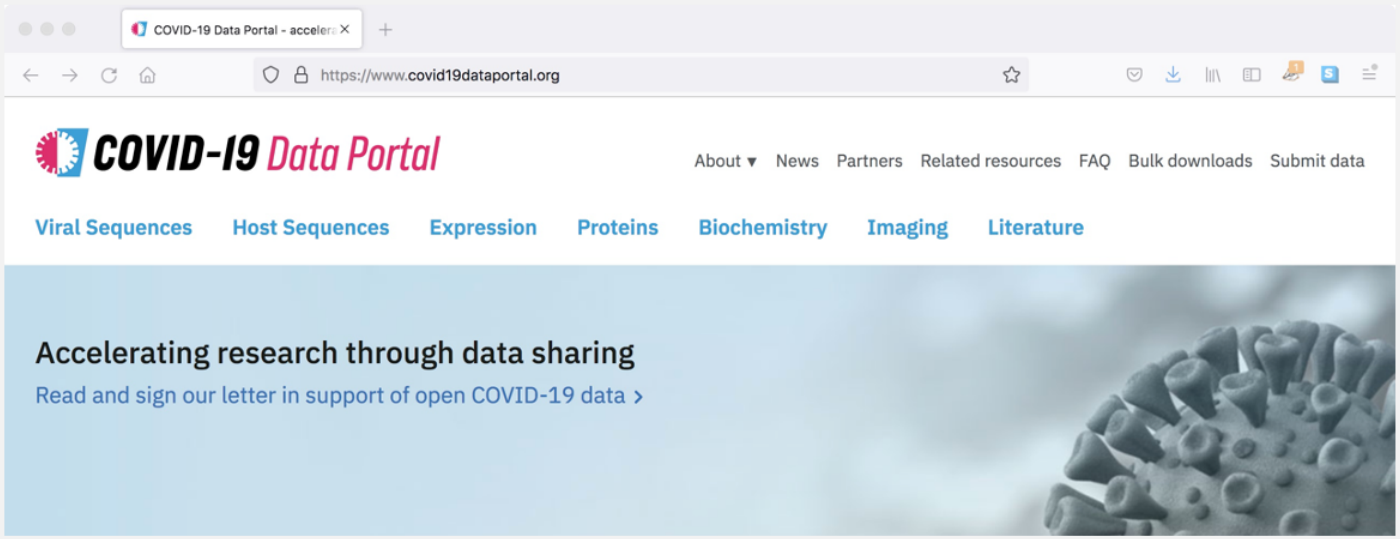
Leveraging open hardware to alleviate the burden of COVID-19 on global health systems

Andre Maia Chagas, Jennifer C. Molloy, Lucia L. Prieto-Godino, Tom Baden

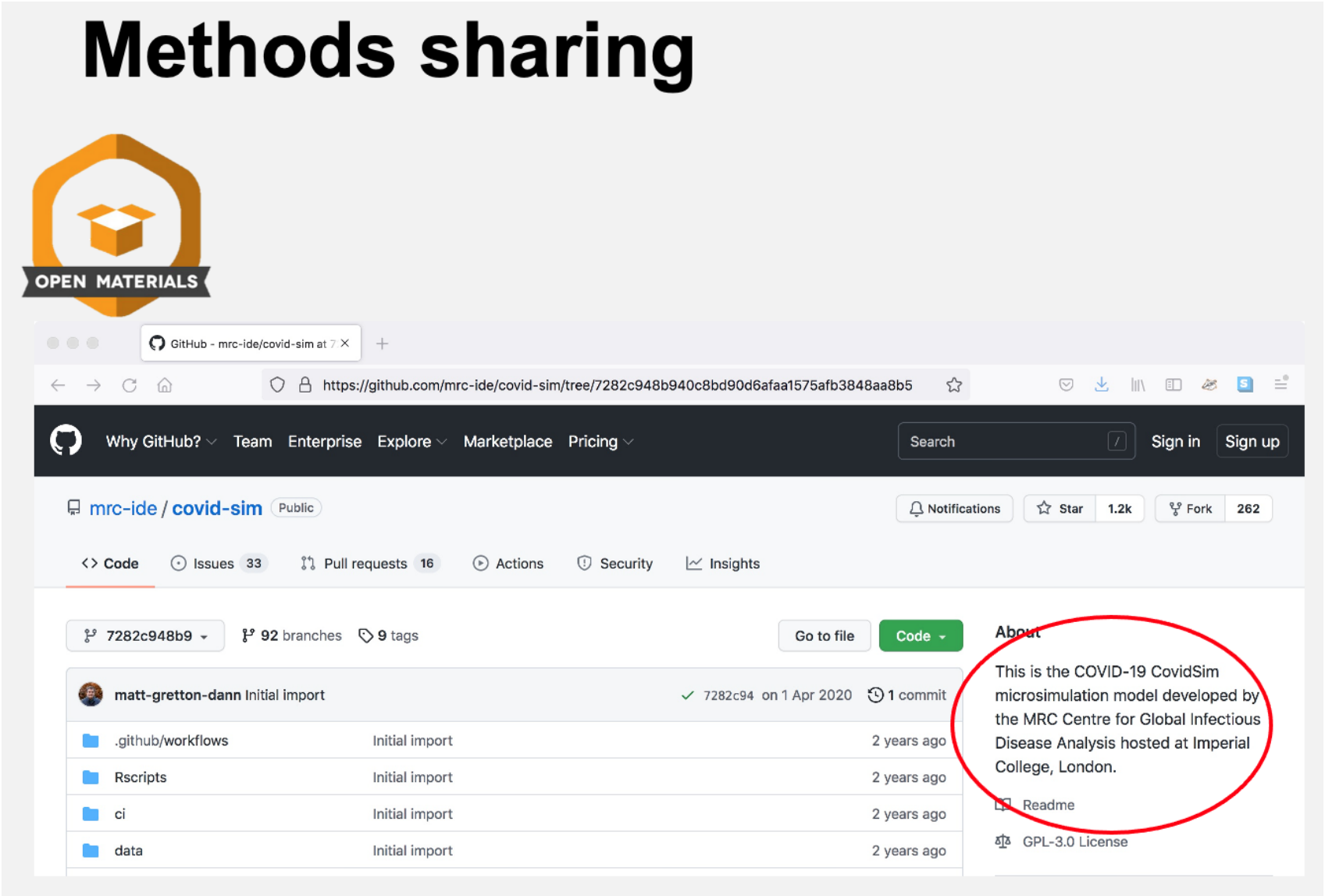
Published: April 24, 2020 • https://doi.org/10.1371/journal.pbio.3000730

Open Science Data

Data sharing



Open Science Methods



Open Science reporting

Open reporting

- Avoid publication bias, publish ALL results
- Be honest about biases and conflicts of interest
- Publish according to best practice guidelines
- Publish preprints

Attention: Terminal

Note: For the next lecture, you need access to a Terminal!

For Mac Users, the work is already done for you, just:

- Open the **Terminal** App on your Mac, or
- Use the terminal directly from RStudio

Attention: Windows Users

- **Git Bash** is a terminal for Windows that provides a Git command-line experience.
  - It allows users to interact with their repositories and run shell commands.
- Visit the official Git website: https://git-scm.com/
  - Click on the "Download for Windows" button and install the program.
- Once installed, launch Git Bash by searching for it in the Windows Start Menu.

Session Info

```
## R version 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.2.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.4.2    fastmap_1.2.0     cli_3.6.3         tools_4.4.2
##  [5] htmltools_0.5.8.1 rstudioapi_0.16.0 yaml_2.3.10       rmarkdown_2.28
##  [9] knitr_1.48        xfun_0.47         digest_0.6.37     rlang_1.1.4
## [13] evaluate_1.0.0
```