

# Open Science and reproducibility

Evan Johnson

2026-01-05

# 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."*

- Jailer seeks to retain full control over precious data
- Jailer 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 little help from ChatGPT!

## What is the problem?

# Open and Reproducible Science

## ELIZABETH HOLMES TRIAL CHARGES



- **GUILTY** Conspiracy to commit wire fraud against Theranos investors
- **NOT GUILTY** Conspiracy to commit wire fraud against Theranos patients
- **DEADLOCKED** Wire fraud against an investor
- **DEADLOCKED** Wire fraud against an investor
- **DEADLOCKED** Wire fraud against an investor
- **GUILTY** Wire fraud against an investor
- **GUILTY** Wire fraud against an investor
- **GUILTY** Wire fraud against an investor
- **NOT GUILTY** Wire fraud against a patient
- **NOT GUILTY** Wire fraud against a patient
- **NOT GUILTY** Wire fraud against Theranos patients



# 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<sup>1,2</sup>; Arjun K. Manrai, PhD<sup>2,3</sup>; Marzyeh Ghassemi, PhD<sup>4,5</sup>

» 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

⇒ The Reproducibility crisis is a Replication crisis

# Reproducibility Crisis

## Mismatch

Slide inspired by Chris Chambers

- Reliable results
- Publication regardless of outcome
- Slow, careful research
- Transparency
- Collaborative research
- Effect on clinical practice

What's good for science

- Surprising results
- Publication of “positive” results
- Fast, incremental progress
- Secrecy
- Competitiveness
- “Impact”, e.g. press attention

What's good for scientists

## 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<sup>\*†</sup>

+ 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

# Reproducibility Crisis

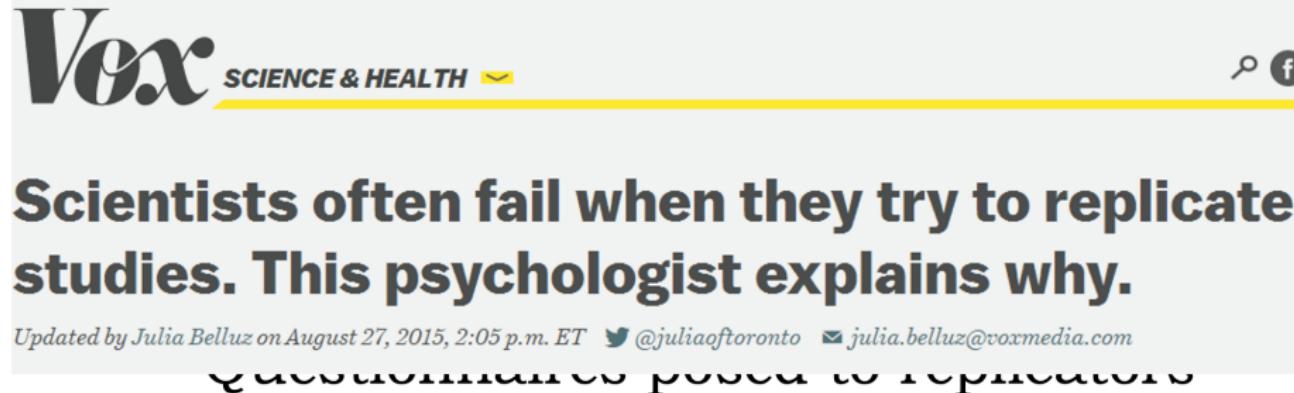
## OSC (Aug 2015)

- 100 pairs of original studies and re
- M. 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
- 

# Reproducibility Crisis

## OSC (Aug 2015)

- 100 pairs of original studies and  
re



A screenshot of a Vox Science & Health article. The title is "Scientists often fail when they try to replicate studies. This psychologist explains why." The article was updated by Julia Belluz on August 27, 2015, at 2:05 p.m. ET. It includes social media sharing icons for Twitter and Facebook.

SCIENCE & HEALTH

## 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](mailto:julia.belluz@voxmedia.com)

# Reproducibility Crisis

## 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



# Reproducibility Crisis

## OSC (Aug 2015)

- 100 pairs of original studies and  
**SCIENTIFIC METHOD / SCIENCE & EXPLORATION**



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

# Reproducibility Crisis

## OSC (Aug 2015)

SCIENCE

### *Many Psychology Findings Not as Strong as Claimed, Study Says*

By BENEDICT CAREY AUG. 27, 2015

Speaking of Science

## Massive International Project Raises Many scientific Questions about the Validity of That's a problem Psychology Research

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

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



By Joel Achenbach August 27 Follow @joelachenbach

### Most Read

# Reproducibility Crisis

## OSC (Aug 2015)

SCIENCE

### *Many Psychology Findings Not as Strong as Claimed, Study Says*

SCIENCE

#### **The Reproducibility Crisis: Cognitive Scientist, Heal Thyself**

**That's a problem** by **Joel Achenbach**

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

#### **International Project Raises Questions about the Validity of Psychology Research**

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



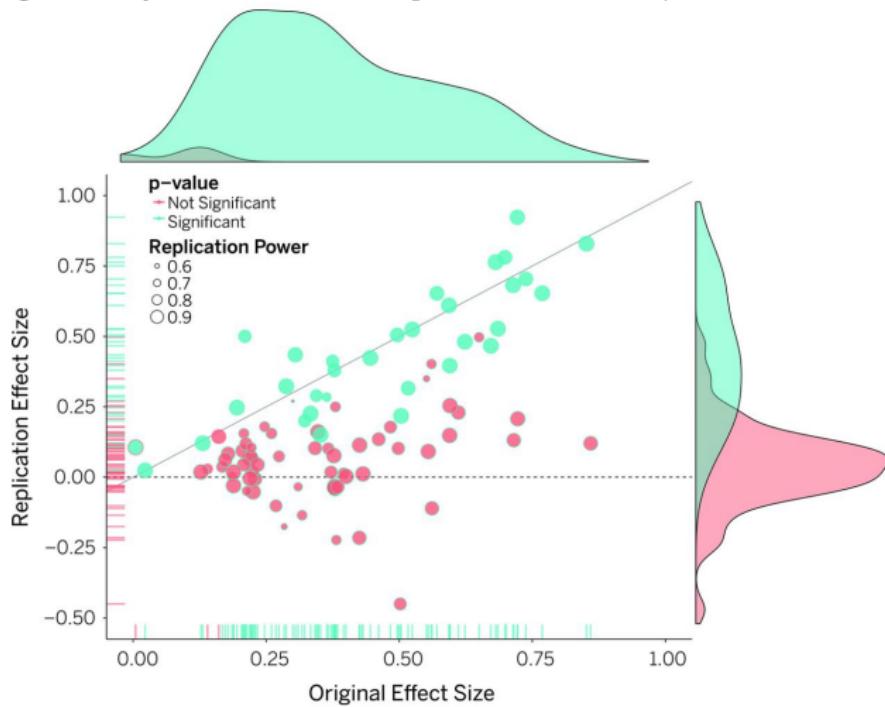
398

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

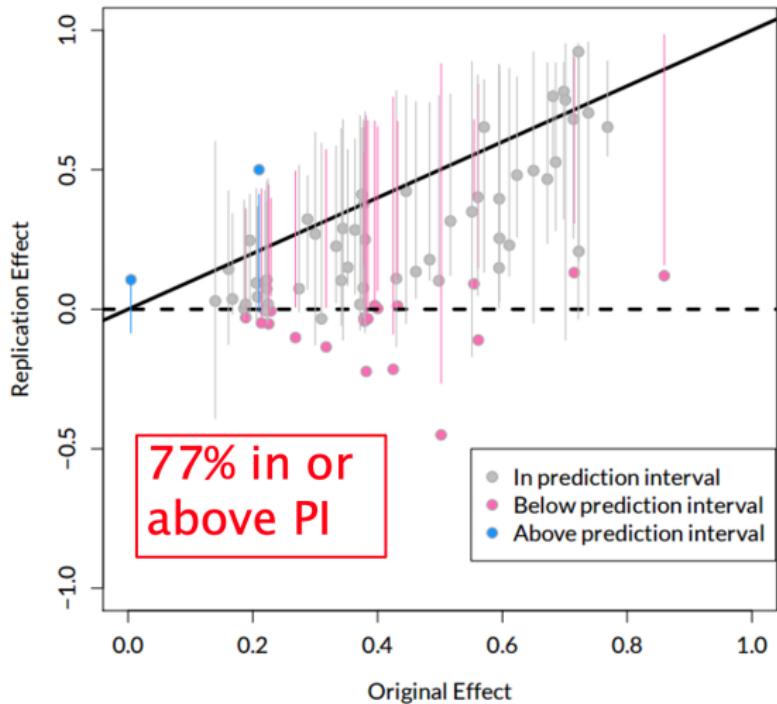
$\hat{\theta}_{orig} \cong \hat{\theta}_{rep} \rightarrow$  variation

Robustness/Sensitivity

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

## The Lung Metagene Score (LMS)

The NEW ENGLAND JOURNAL OF MEDICINE

ORIGINAL ARTICLE

### A Genomic Strategy to Refine Prognosis in Early-Stage Non-Small-Cell Lung Cancer

Anil Potti, M.D., Sayan Mukherjee, Ph.D., Rebecca Petersen, M.D.,  
Holly K. Dressman, Ph.D., Andrea Bild, Ph.D., Jason Koontz, M.D.,  
Robert Kratzke, M.D., Mark A. Watson, M.D., Ph.D., Michael Kelley, M.D.,  
Geoffrey S. Ginsburg, M.D., Ph.D., Mike West, Ph.D., David H. Harpole, Jr., M.D.,  
and Joseph R. Nevins, Ph.D.

One of “the most significant advances on the front lines of cancer.” – Ozols et al., JCO, 25:146-62, 2007, ASCO survey of 2006.

# 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

# Non-Reproducible Science

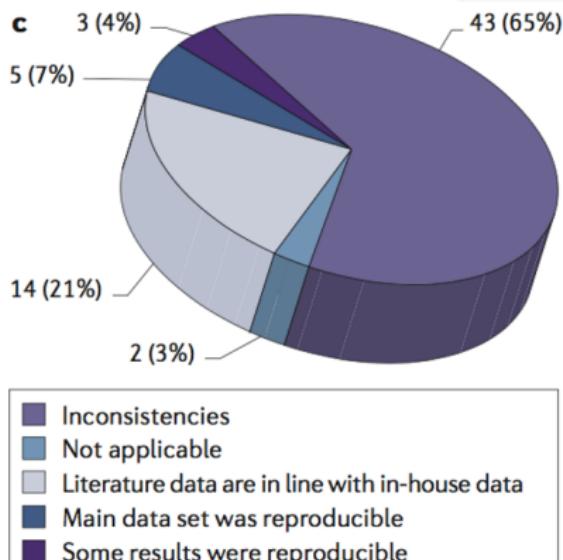
## Correspondence

*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<sup>1</sup>, Thomas Schlange<sup>2</sup> & Khusru Asadullah<sup>3</sup>



## Real-life consequences

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

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

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.B.S., Ph.D.,  
SreyRam Kuy, M.D., M.H.S., Timothy D. Henry, M.D., and Amit N. Patel, M.D.

## 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

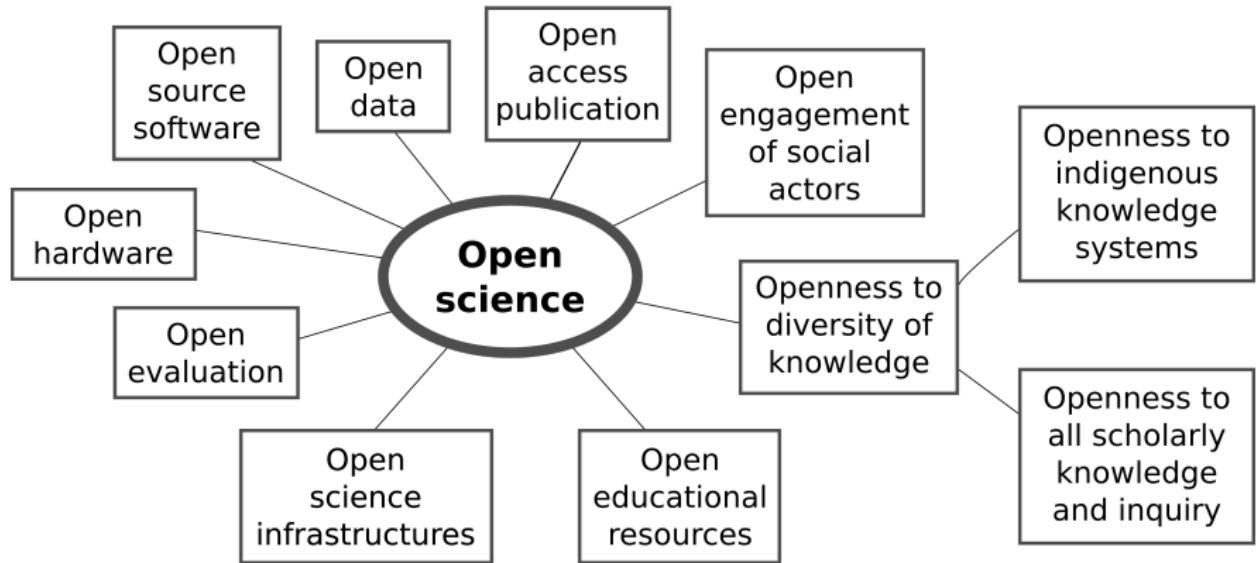


CrossMark

- 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

## Good Research Practice



All images on CC0 license from <https://pixabay.com/>

# Some cool tools for reproducible science

- **RMarkdown/Jupyter Notebooks:** Interactive and dynamic 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.

## *Data, Code and Software*

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

**Code:** Websites, supplementary files, GitHub (preferred)

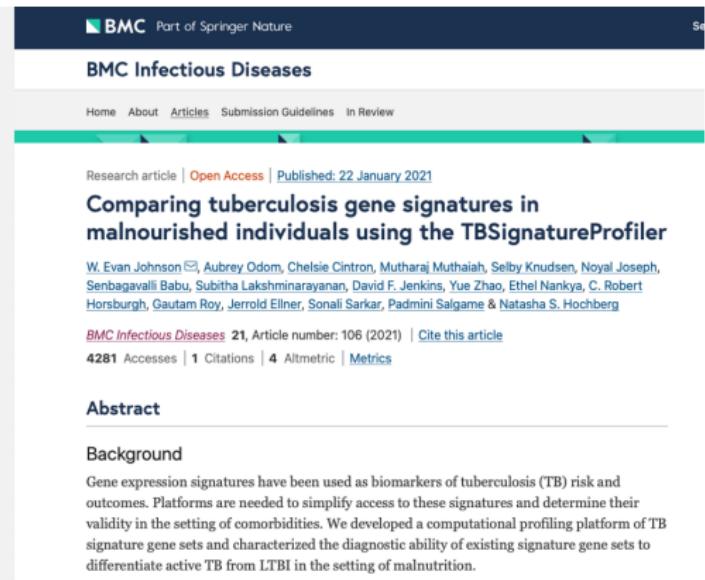
**Software:** Make an R package! (or Python, etc.) Share on GitHub or Bioconductor



# Open Science Practices (Johnson Lab)

## Using R Markdown

R Code: [https://github.com/wevanjohnson/tbsp\\_malnutrition](https://github.com/wevanjohnson/tbsp_malnutrition) Software: <https://github.com/compbioimed/TBSignatureProfiler>  
<https://bioconductor.org/packages/release/bioc/html/TBSignatureProfiler.html>



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 Narkya, C. Robert Horsburgh, Gautam Roy, Jerrold Ellner, Sonali Sarkar, Padmini Saligame & Natasha S. Hochberg

*BMC Infectious Diseases* 21, Article number: 106 (2021) | [Cite this article](#)  
4281 Accesses | 4 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](https://github.com/wevanjohnson/tbsp_malnutrition)

Software: <https://github.com/compbioimed/TBSignatureProfiler>  
<https://bioconductor.org/packages/release/bioc/html/TBSignatureProfiler.html>

# Open Science Practices (Johnson Lab)

## Using R Markdown

### 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](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/compbioimed/TBSignatureProfiler>).

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 Jarkar, Padmuni Salgane & Natasha S. Hochberg

16 (2021) | Cite this article

Metrics

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](https://github.com/wevanjohnson/tbsp_malnutrition)

Software: <https://github.com/compbioimed/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](https://github.com/wevanjohnson/tbsp_malnutrition)

# Open Science Practices (Johnson Lab)

## Using R Shiny

OXFORD  
ACADEMIC

The American Journal of  
**CLINICAL NUTRITION**

Issues More Content ▾ Submit ▾ About ▾ Purchase Advertise ▾ All The American Jour



Volume 112, Issue 6  
December 2020

**Article Contents**

ABSTRACT  
Introduction  
Methods  
Results  
Discussion  
Acknowledgments  
Notes  
References

Exploring changes in the human gut microbiota and microbial-derived metabolites in response to diets enriched in simple, refined, or unrefined carbohydrate-containing foods: a post hoc analysis of a randomized clinical trial 

Tyler Faits, Maura E Walker, Jose Rodriguez-Morato, Huicui Meng, Julie E Gervis, Jean M Galluccio, Alice H Lichtenstein, W Evan Johnson, Nirupa R Matthan 

*The American Journal of Clinical Nutrition*, Volume 112, Issue 6, December 2020, Pages 1631–1641, <https://doi.org/10.1093/ajcn/nqaa254>

Published: 16 September 2020 Article history ▾

 PDF  Split View  Cite  Permissions  Share ▾

**ABSTRACT**

**Background**

Dietary carbohydrate type may influence cardiometabolic risk through alterations in the gut microbiome and microbial-derived metabolites, but evidence is limited.

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

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

# Open Science Practices (Johnson Lab)

*Docker and  
Docker Hub*

The image shows a composite screenshot of two web pages. On the left is a large black circular graphic containing the text "Docker and Docker Hub". To its right is a screenshot of a GitHub repository page for "schifferl/tuberculosis". The GitHub page includes a list of commits, a sidebar with repository details like "Tuberculosis Gene Expression Data for Machine Learning", and a "Docker Hub" tab. Below the GitHub page is a screenshot of the Docker Hub page for "schifferl/tuberculosis.pipeline", which shows a blue cube icon, a brief description, and a "Docker Pull Command" section with the command "docker pull schifferl/tuberculosis".

[GitHub: https://github.com/schifferl/tuberculosis](https://github.com/schifferl/tuberculosis)

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

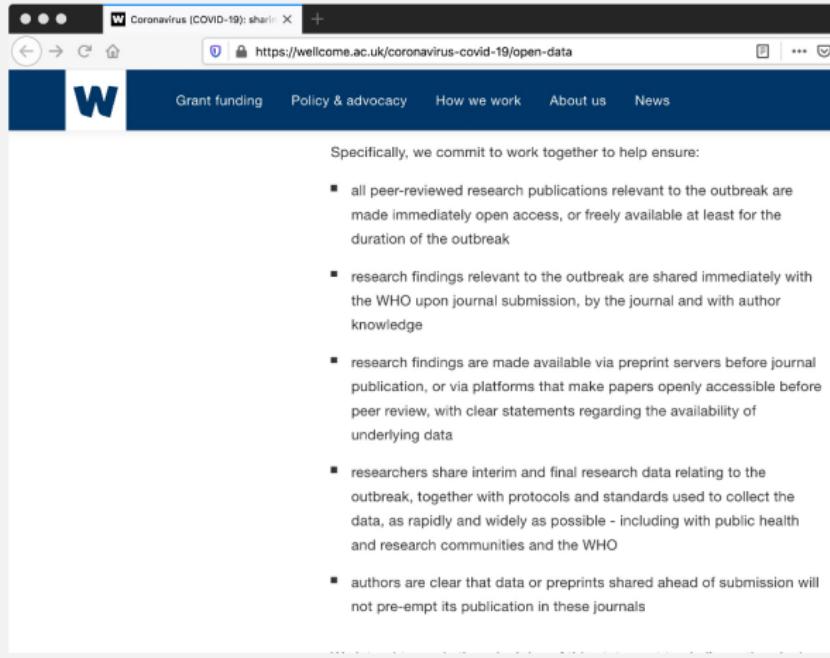
- ① **Reproducibility:** Ensure pipelines run the same everywhere.
- ② **Scalability:** Efficiently scale to cloud or HPC environments.
- ③ **Portability:** Share workflows and containers effortlessly.

# Resources

- Nextflow Documentation
- Docker Documentation

# Open Science Practices

## OS and the pandemic



The screenshot shows a browser window with the URL <https://wellcome.ac.uk/coronavirus-covid-19/open-data>. The page features the Wellcome Trust logo and navigation links for Grant funding, Policy & advocacy, How we work, About us, and News. Below the navigation, a statement reads: "Specifically, we commit to work together to help ensure:" followed by a bulleted list of nine points detailing their commitment to open access and data sharing.

- all peer-reviewed research publications relevant to the outbreak are made immediately open access, or freely available at least for the duration of the outbreak
- research findings relevant to the outbreak are shared immediately with the WHO upon journal submission, by the journal and with author knowledge
- research findings are made available via preprint servers before journal publication, or via platforms that make papers openly accessible before peer review, with clear statements regarding the availability of underlying data
- researchers share interim and final research data relating to the outbreak, together with protocols and standards used to collect the data, as rapidly and widely as possible - including with public health and research communities and the WHO
- authors are clear that data or preprints shared ahead of submission will not pre-empt its publication in these journals



## Publishing



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



- Preprint of finished manuscript before peer review

## 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

The screenshot shows a web browser displaying the COVID-19 Data Portal. In the top left corner, there is a logo for "OPEN DATA" featuring a blue hexagon with a white bar chart icon inside. The main header of the website is "COVID-19 Data Portal". Below the header, there is a navigation bar with links to "About", "News", "Partners", "Related resources", "FAQ", "Bulk downloads", and "Submit data". The main content area features a large blue banner with the text "Accelerating research through data sharing" and a link "Read and sign our letter in support of open COVID-19 data >". To the right of the banner, there is a close-up image of a COVID-19 virus particle.

# Open Science Methods

## Methods sharing

The screenshot shows a GitHub repository page for 'mrc-ide/covid-sim'. At the top, there's a large orange hexagonal icon with a white box inside, labeled 'OPEN MATERIALS'. Below it is a browser window showing the URL <https://github.com/mrc-ide/covid-sim/tree/7282c948b940c8bd90d6afaa1575efb3848aa8b5>. The GitHub header includes links for Why GitHub?, Team, Enterprise, Explore, Marketplace, Pricing, a search bar, and sign-in options.

The main content area shows the repository details: 'mrc-ide / covid-sim' (Public), 33 issues, 16 pull requests, 16 actions, 1 commit, 92 branches, 9 tags, and a 'Code' button. Below this is a list of files: '.github/workflows', 'Rscripts', 'ci', and 'data', all listed as 'Initial import' 2 years ago.

To the right, there's an 'About' section with the following text:

This is the COVID-19 CovidSim microsimulation model developed by the MRC Centre for Global Infectious Disease Analysis hosted at Imperial College, London.

Below this are links for 'Readme' and 'GPL-3.0 License'.

# 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 4.5.1 (2025-06-13)
## Platform: aarch64-apple-darwin20
## Running under: macOS Tahoe 26.1
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.5-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_4.5.1    fastmap_1.2.0    cli_3.6.5      tools_4.5.1
## [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10    rmarkdown_2.29
## [9] knitr_1.50        xfun_0.52      digest_0.6.37  rlang_1.1.6
## [13] evaluate_1.0.4
```