FAIR Bioinfo 2022

Best practice in your bioinformatic projects [1]

Pierre Marin, Matteo Hiriart & ´ Nadia Goué

Université Clermont Auvergne, AuBi, Mésocentre

25 octobre 2022

1. This work is derived from the IFB and I2BC team members

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and

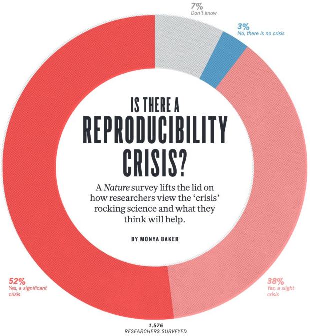
is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets PPV $=(1-\beta)R/(R$

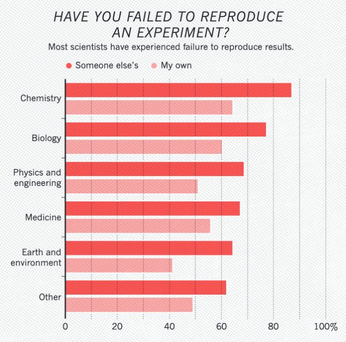
Crisis elements

- Highlighted around 2005
- Since 2010 more articles related to the non reproducibility
- Medecine is one of the most impacted discipline

Reproducibility crisis

2016

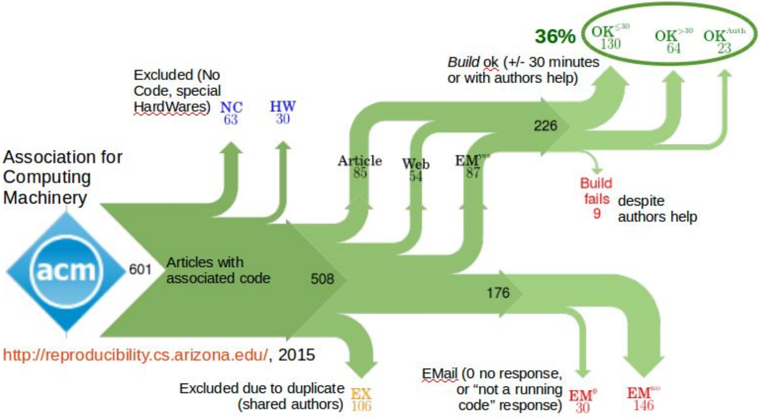




Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533,** 452–454 (**2016**). https://doi.org/10.1038/533452a

4

Also in computer sciences



http://reproducibility.cs.arizona.edu/v2/data/Total.png

5

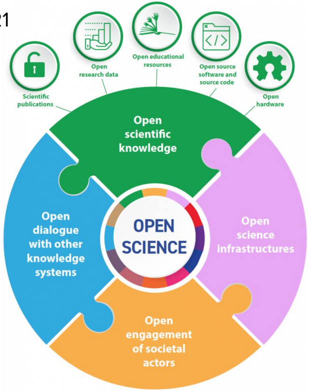
Long term negative impact of retracted papers

Article	Year of retraction	Citing Articles before retraction	Citing Articles after retraction	Total cites (journals indexed by Web of Science)
1. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. N ENGL J MED; APR **2013**. Estruch R, et al.	2018	1919	816	2735
2. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. LANCET; FEB 28 **1998**. Wakefield AJ, et al.	2010	642	867	1509
3. Visfatin: A protein secreted by visceral fat that mimics the effects of insulin. SCIENCE; JAN **2005**. Fukuhara A, et al.	2007	232	1192	1424
4. An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. PLANT J; MAR **2003**. Voinnet O, et al.	2015	896	375	1271
5. Lysyl oxidase is essential for hypoxia-induced metastasis. NATURE; APR **2006**. Erler JT, et al.	2020	977	81	1058

Retraction Watch : Top 10 most highly cited retracted papers
https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/ 6

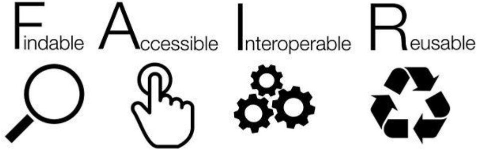
A way out: Open science and FAIR principles

2021



Graphic on page 11. UNESCO Recommendation on Open Science. CC BY IGO 3.0 C. Green

2016



Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

By SangyaPundir - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=53414062

7

FAIR history

- Born in 2016 with *The FAIR Guiding Principles for scientific data management and stewardship*

. https://doi.org/10.1038/sdata.2016.18

www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.#

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

FAIR principles

Findable

By 維基小霸王 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=88894774

PID
Repository

[8] https://doi.org/10.1038/sdata.2016.18

FAIR principles

Findable


By 維基小霸王 - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=88894774

PID
Repository

Accessible

https://nilsfirstworldproblems.tumblr.com/post/147555650875/i-cant-reach-the-top-shelves-of-the-kitchen

Protocols
(free, open, auth.)

9
https://doi.org/10.1038/sdata.2016.18

FAIR principles

Findable

By 維基小霸王 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=88894774

PID
Repository

Accessible

https://nilsfirstworldproblems.tumblr.com/post/147555650875/i-cant-reach-the-top-shelves-of-the-kitchen

Protocols
(free, open, auth.)

Interoperable

By Unknown author - Popular Science Monthly Volume 88, Public Domain, https://commons.wikimedia.org/w/index.php?curid=22614407

Standards
(format, vocabulary)

10
https://doi.org/10.1038/sdata.2016.18

FAIR principles

Findable

By 維基小霸王 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=88894774

PID
Repository

Accessible

https://nilsfirstworldproblems.tumblr.com/post/147555650875/i-cant-reach-the-top-shelves-of-the-kitchen

Protocols
(free, open, auth.)

Interoperable

By Unknown author - Popular Science Monthly Volume 88, Public Domain, https://commons.wikimedia.org/w/index.php?curid=22614407

Standards
(format, vocabulary)

Reusable

By Sun Ladder - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=5746428

Metadata
License
Origin

11 https://doi.org/10.1038/sdata.2016.18

FAIR tools

Findable  Accessible  Interoperable  Reusable 

?

Data

ENA European Nucleotide Archive

FAIRsharing.org standards, databases, policies

LibreOffice The Document Foundation

DMP OPIDoR

cc

doi

HAL open science

jupyter

eLabFTW

Software and analyses

GitLab

CONDA

docker

snakemake

CeCILL

12

Tools & use cases

Several tools but which ones to use and how? do some of them interact with each other?

3 use cases based on the previous sessions:

- E-labbook
- Reproducibility of running code
- Reproducibility in HPC

https://nbis-reproducible-research.readthedocs.io/en/course_2104/introduction/

13

FAIR session with AuBi

Objectives

- Discover FAIR practices
- Discover tools for best practices
- Use tool and best practices in practice sessions
- 5 sessions for courses and practices
  - Day 1 : Introduction to FAIR training and Git
  - Day 2 : Git practice
  - Day 3 : Encapsulation course
  - Day 4 : Encapsulation training
  - Day 5 : Documentation course and training

Contents

Contents

Contents

- Introduction to FAIR practices
- Code control using Git
  - Git environment
  - Gitlab and Github
- Encapsulation process
  - Conda environment and packages use CONDA
  - Containers as docker & singularity
  - Reproducible workflow using snakemake

Contents

- Introduction to FAIR practices
- Code control using Git
  - Git environment
  - Gitlab and Github
- Encapsulation process
  - Conda environment and packages use CONDA
  - Containers as docker & singularity
  - Reproducible workflow using snakemake
- Literate programming and documentation
  - Markdown syntax
  - Rmarkdown for R
  - Jupyterlab for Python