Week 2: Healthcare Data: Open Data & Reproducibility

Charlotte Hadley

Topics for today

These are the goals for today’s lecture

1. Demonstrate to you that there’s a reproducibility crisis
2. Explain steps you can take to improve the reproducibility of your research
3. Identify the meanings of Open Access and Open Data

> We’re going to spill into the workshop time a little bit today.

“There’s a reproducibility crisis in academia”

Have you ever heard this phrase before?

There is a reproduciblity crisis in academia

Back in 2005 a ground breaking paper by John Ioannidis[1] exposed an unsettling truth

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.

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.

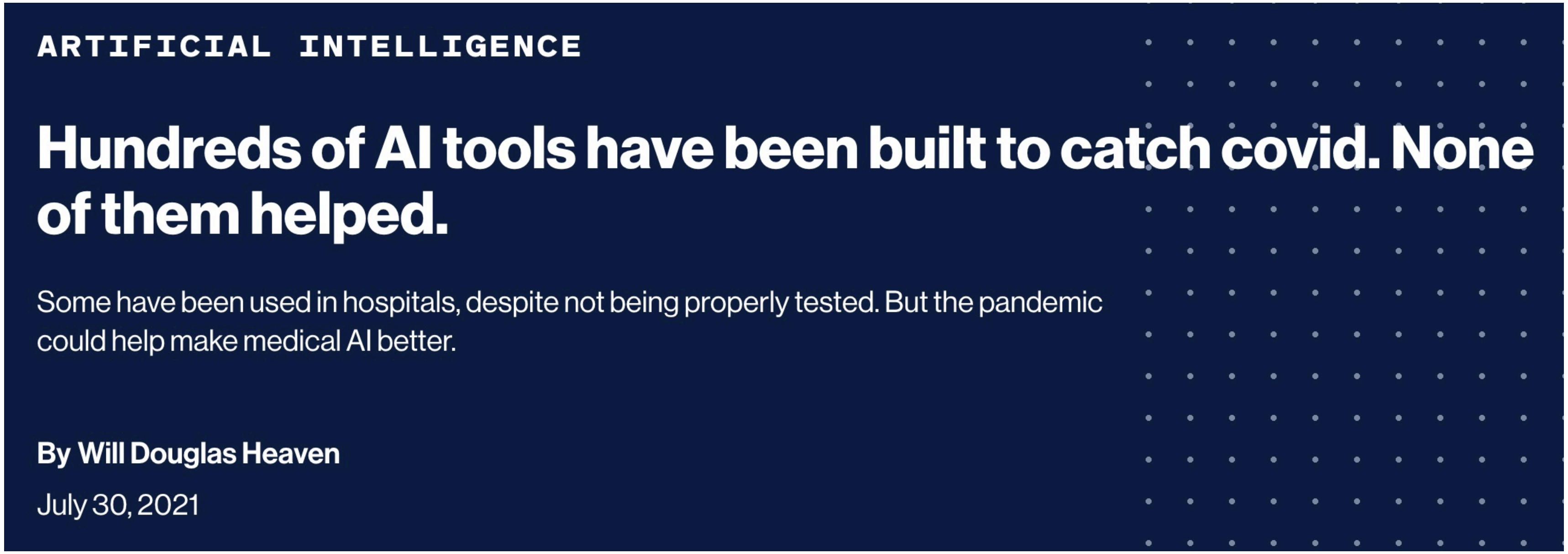
is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed 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

There’s still a crisis almost two decades later

Despite lots being written about the crisis… it’s still here. And with the rise of

> *“machine learning can solve anything!”*

… the crisis is evolving and getting more complicated, as reported by Douglas Heaven[2]



ARTIFICIAL INTELLIGENCE

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven

July 30, 2021

Reproducibility vs Replicability

Quite a lot of the time you'll see these two words used interchangeably - but they have specific meanings. The American Statistical Association (ASA) provides useful advice from Broman et al[3].

> Reproducibility: A study is reproducible if you can take the **original data and the computer code** used to analyze the data and reproduce all of the numerical findings from the study.

- Your module assessment must be reproducible - I must be able to run the code after it's submitted.

Reproducibility

> Reproducibility: A study is reproducible if you can take the **original data and the computer code** used to analyze the data and reproduce all of the numerical findings from the study.

Reproducibility is something to think about from the start of a research project:

- Plan to record and document all *processes* in data collection, wrangling and analysis
- While performing the research keep track of all the things you do - particularly with data!
- When writing up your research ensure that all necessary materials to reproduce your findings are made available.

> Later in this lecture and the workshop we'll look at specific advice for achieving this.

Replicability

> Replicability: This is the act of repeating an entire study, **independently of the original investigator** without the use of original data (but generally using the same methods)

This is a slightly harder topic to conceptualise and much of the lack of replicability comes from what we call “questionable research practices” (QRPs)

> … these misbehaviours lie somewhere on a continuum between scientific fraud, bias, and simple carelessness, so their direct inclusion in the “falsification” category is debatable, although their negative impact on research can be dramatic
>
> Fanelli, 2009[4]

Questionable Research Practices

There are lots of different ways to summarise common QRPs but I quite like this table from 2012 by John et al[5].

Questionable Research Practice	Self-admission rate (Amongst 2,000 pysychologists)
In a paper, failing to report all of a study’s dependent measures	63.4%
Deciding whether to collect more data after looking to see whether the results were significant	55.9%
In a paper, selectively reporting studies that “worked”	45.8%
Deciding whether to exclude data after looking at the impact of doing so on the results	38.2%
In a paper, failing to report all of a study’s conditions	27.7%
In a paper, reporting an unexpected finding as having been predicted from the start	27.0%
In a paper, “rounding off” a p value (e.g., reporting that a p value of .054 is less than .05)	22.0%
Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6%
In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0%
Falsifying data	0.6%

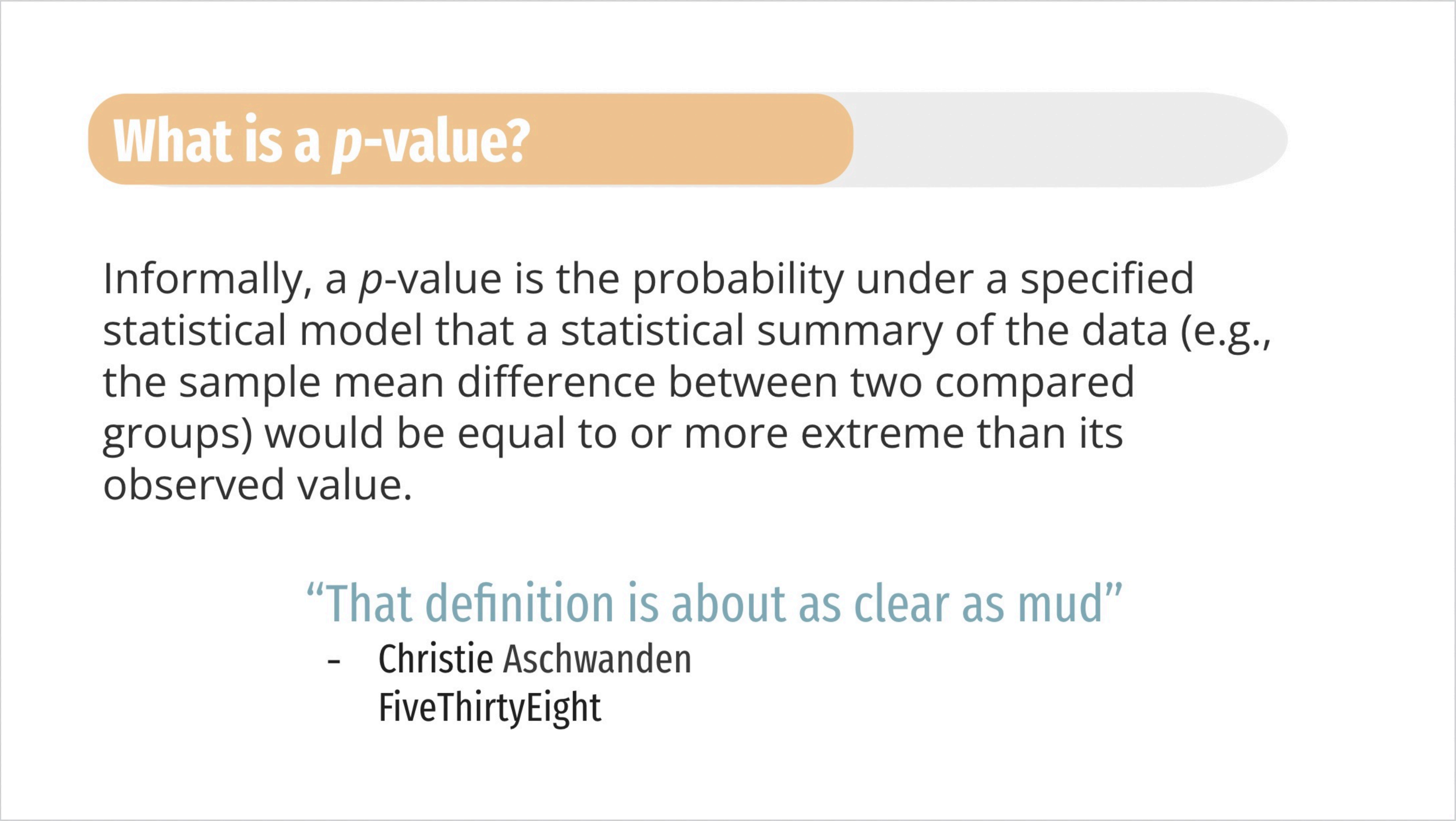
Questionable Research Practices and P-hacking

The majority of these QRPs can be categorised as “P-hacking” or more fully - hacking the P-value.

Questionable Research Practice	Self-admission rate (Amongst 2,000 pysychologists)
In a paper, failing to report all of a study’s dependent measures	63.4%
Deciding whether to collect more data after looking to see whether the results were significant	55.9%
In a paper, selectively reporting studies that “worked”	45.8%
Deciding whether to exclude data after looking at the impact of doing so on the results	38.2%
In a paper, failing to report all of a study’s conditions	27.7%
In a paper, reporting an unexpected finding as having been predicted from the start	27.0%
In a paper, “rounding off” a p value (e.g., reporting that a p value of .054 is less than .05)	22.0%
Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6%
In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0%
Falsifying data	0.6%

What is the p-value?

We’re going to talk about p-values **A LOT** in week 10. For now I wanted to borrow a slide from Lucy D’Agostino McGowan’s talk which is well explained in this Twitter thread.



What is a *p*-value?

Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

> “That definition is about as clear as mud”
>
> \- Christie Aschwanden
> FiveThirtyEight

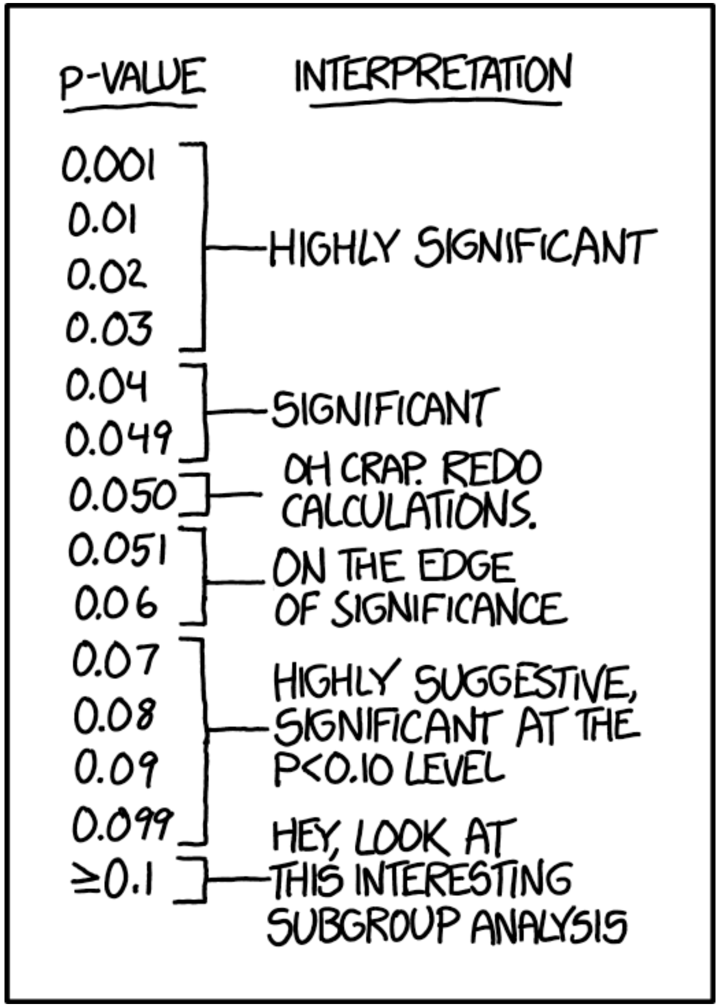
How folks treat the p-value

In some situations “p-values” are considered infallible evidence of an effect or the conclusion of a study.

There are lots of different p-value thresholds, but the most common in healthcare data science is 0.05

Researchers who find their study results in values *just above* 0.05 will explore ways to get the value below 0.05

> That’s p-value hacking.


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

https://xkcd.com/1478/

P-hacking

In 2019 Aschwanden[6] published an article in Wired.com titled “We’re All ‘P-Hacking’ Now” which I highly recommend reading.

It highlights an excellent study by Simmons et al[7] that was able to use p-value hacking to make two increasingly absurd conclusions:

> Study 1: Listening to a children’s song (“Hot Potato” by *The Wiggles*) makes people **feel** older.
>
> Study 2: Listening to a song about old age (“When I’m Sixty-Four” by *The Beatles*) makes people **actually** younger.

… I recommend this paper because it makes very clear recommendations to researchers and reviewers.

Have you ever read an academic paper?

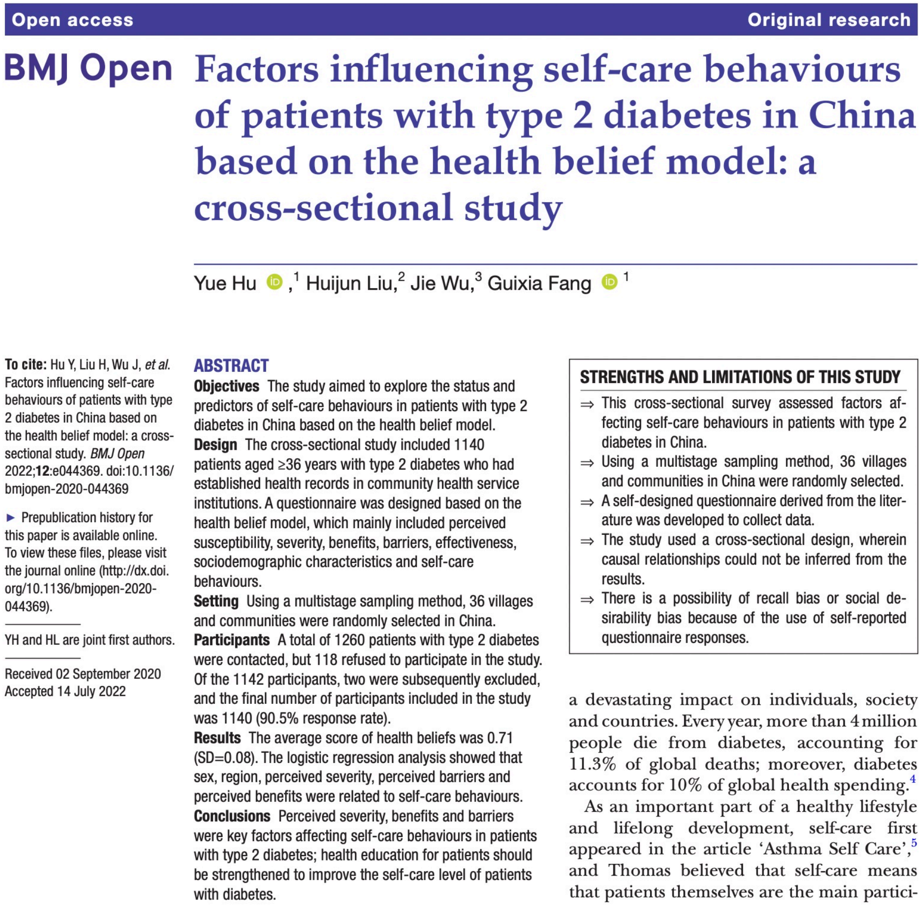
doi.org/10.1177/0956797611417632

Reading papers is a skill (I)

Always expect to read a paper **multiple** times and make notes.

We can kind of neatly split papers into two different types:

Clinical trials like this one from Hu et al[8].



Open access Original research

BMJ Open

Factors influencing self-care behaviours of patients with type 2 diabetes in China based on the health belief model: a cross-sectional study

Yue Hu [1], Huijun Liu,[2] Jie Wu,[3] Guixia Fang [1]

To cite: Hu Y, Liu H, Wu J, *et al*. Factors influencing self-care behaviours of patients with type 2 diabetes in China based on the health belief model: a cross-sectional study. *BMJ Open* 2022;**12**:e044369. doi:10.1136/bmjopen-2020-044369

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-044369).

YH and HL are joint first authors.

Received 02 September 2020
Accepted 14 July 2022

ABSTRACT

Objectives The study aimed to explore the status and predictors of self-care behaviours in patients with type 2 diabetes in China based on the health belief model.

Design The cross-sectional study included 1140 patients aged ≥36 years with type 2 diabetes who had established health records in community health service institutions. A questionnaire was designed based on the health belief model, which mainly included perceived susceptibility, severity, benefits, barriers, effectiveness, sociodemographic characteristics and self-care behaviours.

Setting Using a multistage sampling method, 36 villages and communities were randomly selected in China.

Participants A total of 1260 patients with type 2 diabetes were contacted, but 118 refused to participate in the study. Of the 1142 participants, two were subsequently excluded, and the final number of participants included in the study was 1140 (90.5% response rate).

Results The average score of health beliefs was 0.71 (SD=0.08). The logistic regression analysis showed that sex, region, perceived severity, perceived barriers and perceived benefits were related to self-care behaviours.

Conclusions Perceived severity, benefits and barriers were key factors affecting self-care behaviours in patients with type 2 diabetes; health education for patients should be strengthened to improve the self-care level of patients with diabetes.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ This cross-sectional survey assessed factors affecting self-care behaviours in patients with type 2 diabetes in China.
- ⇒ Using a multistage sampling method, 36 villages and communities in China were randomly selected.
- ⇒ A self-designed questionnaire derived from the literature was developed to collect data.
- ⇒ The study used a cross-sectional design, wherein causal relationships could not be inferred from the results.
- ⇒ There is a possibility of recall bias or social desirability bias because of the use of self-reported questionnaire responses.

a devastating impact on individuals, society and countries. Every year, more than 4 million people die from diabetes, accounting for 11.3% of global deaths; moreover, diabetes accounts for 10% of global health spending.[4]

As an important part of a healthy lifestyle and lifelong development, self-care first appeared in the article 'Asthma Self Care',[5] and Thomas believed that self-care means that patients themselves are the main partici-

In some clinical trials the abstract includes **a lot of structured information** - it depends on the publisher.

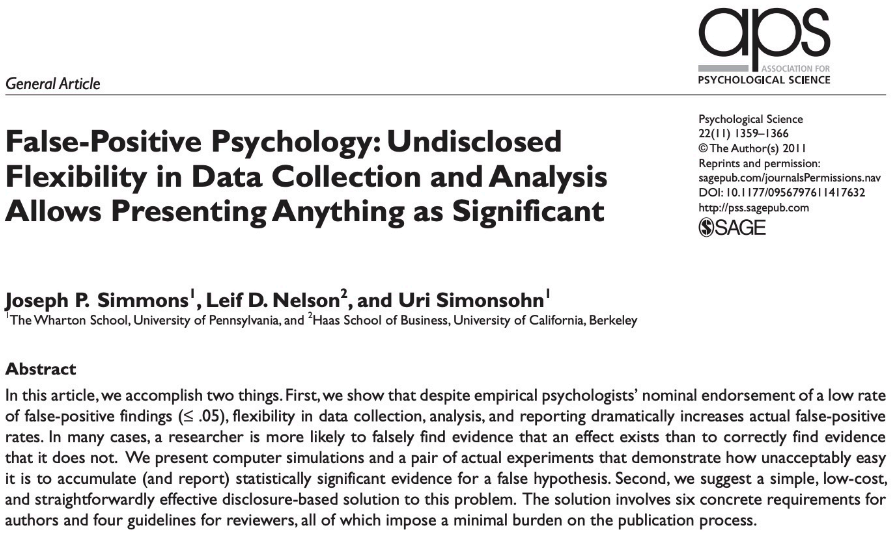
Some journals like BMJ Open even include study strengths and weaknesses.

Read these papers in this order:

- Abstract
- Tables and/or figures
- Conclusions
- Introduction

Reading papers is a skill (II)

But **most** papers look this one from Simmons et al[simmons_false?].



General Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
22(11) 1359–1366
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Abstract

In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.

Unlike in some medical journals, the abstract is unstructured and might not contain much quantitative information.

But the reading order for the paper is the same:

- Abstract
- Tables and/or figures
- Conclusions
- Introduction

Reading papers is a skill (III)

- On your first pass of a paper you are trying to understand if the paper is relevant and provides substantial information and/or evidence for your needs.
- Often it will take another 1 or 2 passes to understand the **results** of the paper.
- It usually takes even more effort to understand the **methods** of the paper
  - Methodology information is often provided in “supplementary materials”
  - But unfortunately a good portion of the time you won’t find sufficient information to fully understand the methodology because of **poor reproducibility**.

> Learning about replicability and reproducibility now will help you in understanding the existing literature **and** prepare you to succeed in a research career later on.

Reading papers is a skill (IV)

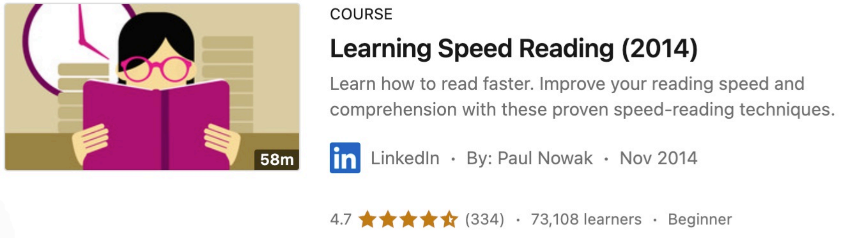
If you decide to go into a research career you’ll likely be reading 10+ papers a week.

I’d highly recommend investing in learning speed reading early in your career.

There’s lots of [very interesting] eye tracking and neurological research into how we read and evidenced methodologies for speed reading that are nicely summarised by Clifton et al[9].

BCU gives you free access to LinkedIn Learning.

Go to linkedin.com/learning-login/ and login with your BCU email address.



COURSE

Learning Speed Reading (2014)

Learn how to read faster. Improve your reading speed and comprehension with these proven speed-reading techniques.

LinkedIn · By: Paul Nowak · Nov 2014

4.7 ★★★★☆ (334) · 73,108 learners · Beginner

… let’s get back to p-hacking

Fighting p-hacking with pre-registration

So far I’ve been speaking about the academic literature at large.

But let’s look specifically at clinical trials.

> All clinical trials began after July 1st 2005 are explicitly required to be **registered**[1] in order to be published in **all biomedical journals** overseen by the International Committee of Medical Journal Editors (ICMJE)[10].
>
> This is independent of the country in which the trial took place.

[1] Frustratingly in the clinical trials community we use the phrase “registration” but everyone else says “pre-registration”.

National Institutes of Health (NIH) and registration

From a cursory search for clinical trial registration information you might come to the conclusion that the advice is only appropriate for studies in the US.

> That is not true.

In 2008 the World Medical Association[11] updated the *Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects* to include a paragraph about registration.

> 19. Every clinical trial must be registered in a publicly accessible database before recruitment of the first subject.

International advice on clinical trials

The NIH provides a really useful tool for comparing the clinical research regulations from several countries:

clinregs.niaid.nih.gov/country/united-kingdom

UK trial registration

Despite these “requirements”… violations are still common as detailed in Bradley et al[12].

However, the UK has announced new infrastructure requiring 100% clinical trial registration via a collaboration between the Human Research Authority and the ISRCTN registry. See Bruckner[13] for a thorough overview of what’s changing.

Registration to “Reduce waste”

For unexplainable reasons there is a lot written about registration of trials “reducing research waste” as it helps reduce duplication of studies.

While elements of this are true… the big takeaway is pre-registration helps prevent Questionable Research Practices.

> In fact it’s the reason why GSK reached a $2.5million settlement in 2004

GlaxoSmithKline & Clinical Trial Registration

GSK chose to settle a civil case instead of engaging in an expensive legal battle over “repeated and persistent fraud” concerning the use of paroxetine in treating depression in adolescents.

As detailed in 2004 by Dyer[14]:

- Two studies showed no benefit to using paroxetine when compared with placebo.
- Three studies found evidence for an increase in suicidal thoughts and behaviour.

Internal company documents confirmed the suppression of these results:

> “it would be commercially unacceptable to include a statement that efficacy had not been demonstrated, as this would undermine the profile of paroxetine.”
>
> Spurgeon 2004[15]

Pre-registration for all studies

There's a growing body of researchers both actually doing pre-registration and calling for it in all disciplines - particularly those that intersect with healthcare.

- Simmons et al[7] have been running aspredicted.org since 2015 to help authors to create pre-registration reports.
- The Centre for Open Science encourages pre-registration on OSF and has in the past run a Preregistration Challenge with a monetary prize.
- I highly recommend reading Nosek et al[16] which gives a great overview of pre-registration and walk through multiple examples of how it works in practice.

> At the moment there's nothing binding you to pre-register non-clinical trials but this topic is simmering away in the background

Prediction markets and replication studies

Prediction markets and replication studies

> In prediction markets, investors make predictions of future events by buying shares in the outcome of the event and the market price indicates what the crowd thinks the probability of the event is.
>
> Harvard Press Release

This was first applied to predicting replicability of research results in 2015 by Dreber et al[17].

There is now evidence for these markets being a reliable estimate of predictability[18].

> Prof. Anna Dreber gives an excellent 40minute overview of replication prediction markets here - https://youtu.be/a5rFDKB1aZc?t=1036

A final word on replicability studies

Most (if not all) of studies about replicability are deeply technical and rely on statistical methods we don't have time to cover in this course.

This includes the foundational paper "Why Most Published Research Findings Are False" by Ioannidis[1].

> Your assessment does not require you to understand or replicate any of the methodologies behind replicability studies.

Reproducibility

Reproducibility

Let’s get back to the ASA reproducibility recommendations from Broman et al[3].

> Reproducibility: A study is reproducible if you can take the **original data and the computer code** used to analyze the data and reproduce all of the numerical findings from the study.

In order for reproducibility to be *possible* we need the original data to be accessible

> We need **Open Data**.

Open Data

Most papers don’t provide their data

More often than papers simply **do not provide** the data that they:

- Use to create charts and tables
- Use to perform statistical tests
- Use to generate their conclusions

> These papers are the antithesis of reproducible.

“Data available on request”

It’s common to see the phrase “data available on request” but that’s frequently meaningless:

> “Data requests to authors are successful in 27–59% of cases, whereas the request is ignored in 14–41% cases”
>
> Tedersoo et al 2021[19]

Even in cases where data is returned it’s often insufficient for reproducibility, Roche et al[20]:

- “Data” might actually be screenshots of charts stored in an Excel workbook
- “Data” might be stored in other “non-machine-readable” formats like PDF or images
- “Data” might only be provided in the post analysis form

Researchers have confusing opinions

In one of the earliest and most cited studies in data sharing in 2011 by Tenopir et al[21] surveyed 1,329 scientists and found:

> “Most respondents (at least 60% across disciplines) agree that lack of access to data generated by other researchers or institutions is a major impediment to progress in science.”

… but

> “A majority of all respondents indicate they are not willing to place all of their data in central repositories with no restrictions”

These findings have been replicated again, and again.

Open Data helps everyone

Researchers

- Open Data helps reproduce previous studies
- Open Data means researchers can do **new** studies, including meta analyses.
- Open Data gives an additional way researchers can be cited
- There is clear evidence Open Data is linked with higher citation rates, eg Colavizza et al 2020[22].

Society

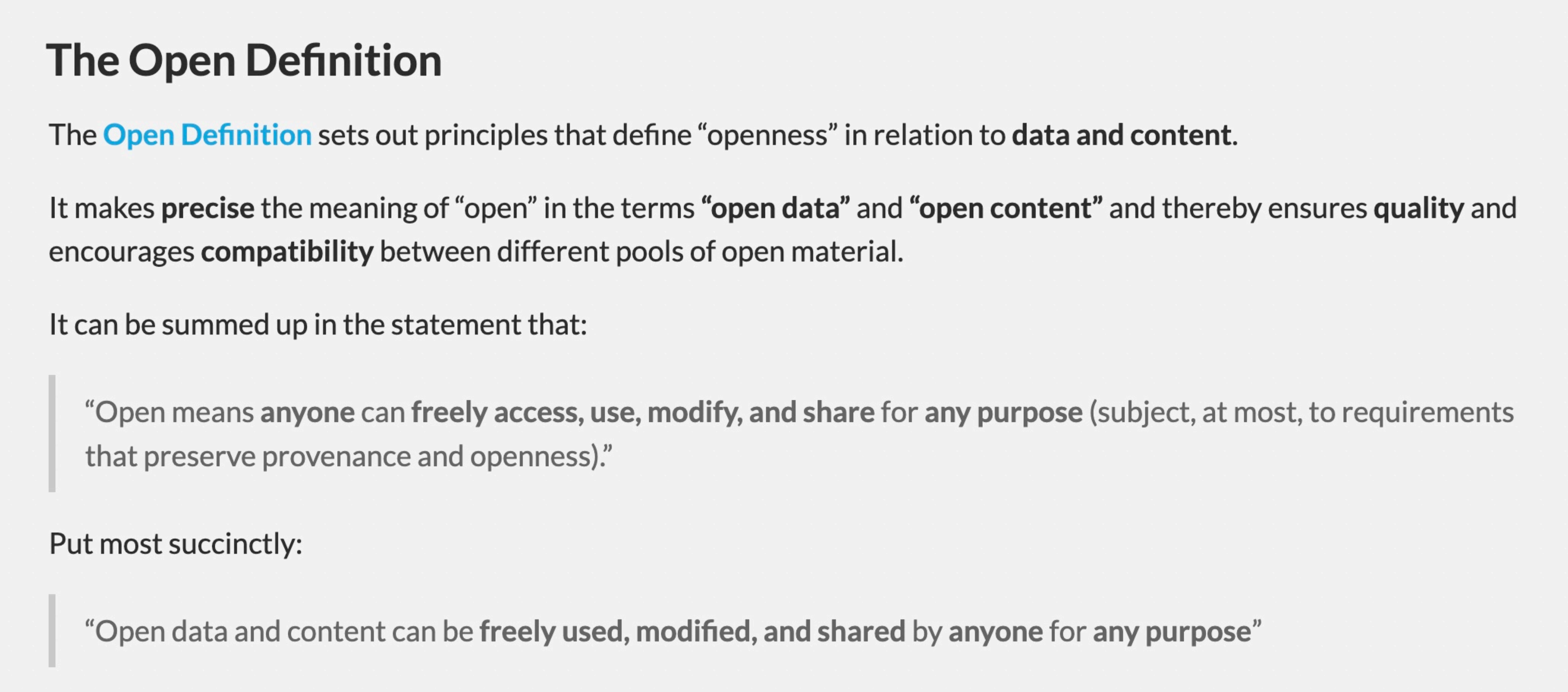
- The coordinated global COVID-19 response benefited significantly from Open Data portals[23].
- Reducing energy consumption of buildings based in occupancy from public datasets[24].
- Emergency planning via use of the Behavioral Risk Factor Surveillance System (BRFSS)[25]

> You can find *some* healthcare specific examples on the course website eng7218.netlify.app/resources/open-data. Please do research your own - and consider sharing them with the group.

Open Data vs Big Data

Big Data is great. It’s the driver behind the Internet of Things and much of modern healthcare technology.

But in most circumstances it is **not** Open Data.



The Open Definition

The Open Definition sets out principles that define “openness” in relation to **data and content**.

It makes **precise** the meaning of “open” in the terms **“open data”** and **“open content”** and thereby ensures **quality** and encourages **compatibility** between different pools of open material.

It can be summed up in the statement that:

> “Open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness).”

Put most succinctly:

> “Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**”

Screenshot of opendefinition.org[26]

Open Data short definition

Let’s minimise these definitions:

> Open data must be..

- Legally open. The data must be subject to an open data license
- Technically open.
  - Data files must be machine-readable and non-proprietary, which often means *plain text*.
  - Accessible from a public server without password protection

Or even more succinctly:

> Open data must be open to humans and computers.

Please note this definition of open data for your assessment.

Open Data licenses

In most cases the Creative Commons' "Choose a license tool" is the best and easiest choice if you have a **dataset** you want to make into Open Data.

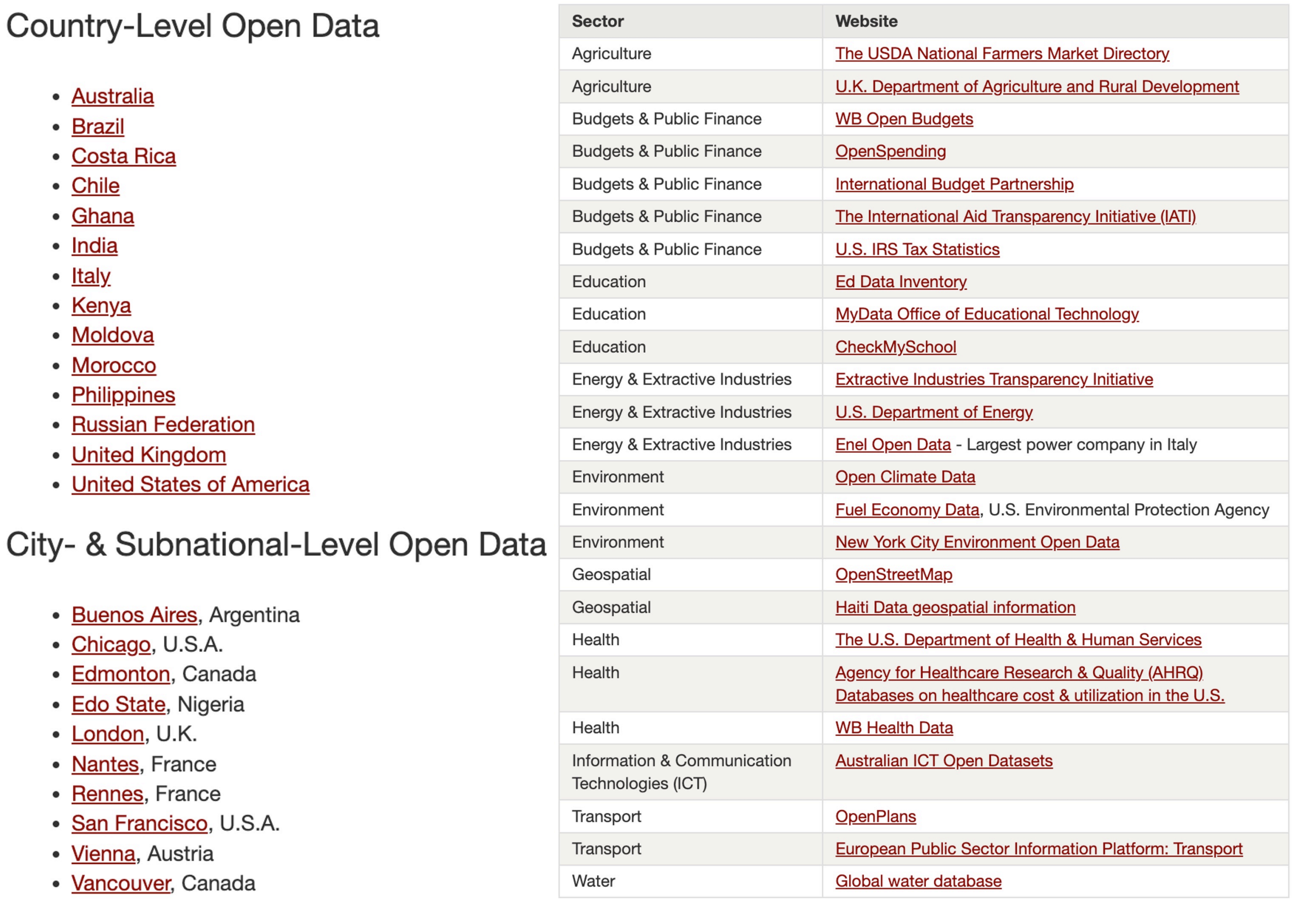
- The CC0 license is the most permissive license.

There are special Open Data licenses used by Governmental/Charity organisations designed to waive liability for use, eg the Open Government License from the UK Government[27]

> In general it's best to use data licenses for data and software licenses for software.

Open Data examples

There are **SO MANY** different sources (or publishers) of Open Data, for a good sample checkout the Open Data Essentials page from the World Bank[28].



Country-Level Open Data

- Australia
- Brazil
- Costa Rica
- Chile
- Ghana
- India
- Italy
- Kenya
- Moldova
- Morocco
- Philippines
- Russian Federation
- United Kingdom
- United States of America

City- & Subnational-Level Open Data

- Buenos Aires, Argentina
- Chicago, U.S.A.
- Edmonton, Canada
- Edo State, Nigeria
- London, U.K.
- Nantes, France
- Rennes, France
- San Francisco, U.S.A.
- Vienna, Austria
- Vancouver, Canada

Sector	Website
Agriculture	The USDA National Farmers Market Directory
Agriculture	U.K. Department of Agriculture and Rural Development
Budgets & Public Finance	WB Open Budgets
Budgets & Public Finance	OpenSpending
Budgets & Public Finance	International Budget Partnership
Budgets & Public Finance	The International Aid Transparency Initiative (IATI)
Budgets & Public Finance	U.S. IRS Tax Statistics
Education	Ed Data Inventory
Education	MyData Office of Educational Technology
Education	CheckMySchool
Energy & Extractive Industries	Extractive Industries Transparency Initiative
Energy & Extractive Industries	U.S. Department of Energy
Energy & Extractive Industries	Enel Open Data - Largest power company in Italy
Environment	Open Climate Data
Environment	Fuel Economy Data, U.S. Environmental Protection Agency
Environment	New York City Environment Open Data
Geospatial	OpenStreetMap
Geospatial	Haiti Data geospatial information
Health	The U.S. Department of Health & Human Services
Health	Agency for Healthcare Research & Quality (AHRQ) Databases on healthcare cost & utilization in the U.S.
Health	WB Health Data
Information & Communication Technologies (ICT)	Australian ICT Open Datasets
Transport	OpenPlans
Transport	European Public Sector Information Platform: Transport
Water	Global water database

Open Data and Health Data

Open Data is awesome. But if we're responsible for data that can identify individuals or groups we have a [legal] duty of care to protect that data.

In the UK we have the Data Protection Act[29] which is the UK's implementation of the General Data Protection Regulation (GDPR).

> In week 6 we will discuss GDPR and the Data Protection Act in the context of anonymising data.
>
> This isn't a course in the law school so we won't go *hard* into the definitions. But there are some things we need to discuss.

Open Data and Health Data

In the DPA[29] there are 6 different types of sensitive data defined in section 86

> I'm highlighting the ones that cover data that you might reasonably collect and consider "health data".

- (a) the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs or trade union membership;
- (b) the processing of genetic data for the purpose of uniquely identifying an individual;
- (c) the processing of biometric data for the purpose of uniquely identifying an individual;
- (d) the processing of data concerning health;
- (e) the processing of data concerning an individual's sex life or sexual orientation;
- (f) the processing of personal data as to [commission or alleged commission of an offence]

> There's some recursion, so let's pull out the definitions of "health data" from Part 7 Section 205 and list everything together

Open Data and Health Data

Part 7 Section 205 defines the following

- **“biometric data”** means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of an individual, which allows or confirms the unique identification of that individual, such as facial images or dactyloscopic data;
- **“data concerning health”** means personal data relating to the physical or mental health of an individual, including the provision of health care services, which reveals information about his or her health status;
- **“genetic data”** means personal data relating to the inherited or acquired genetic characteristics of an individual which gives unique information about the physiology or the health of that individual and which results, in particular, from an analysis of a biological sample from the individual in question;

But we should also still include these sections from section 86:

- (a) the processing of personal data revealing racial or ethnic origin,
- (e) the processing of data concerning an individual’s sex life or sexual orientation;

Open Data and Health Data

There's more reader friendly documentation about health data and DPA 2018 from the Information Commissioner's Office[30].

> We'll talk about this more later.

Open Data mandates and UKRI

Open Data mandates and UKRI

UKRI is responsible for the 6 UK research councils who fund most university-based research in the UK.

Some research councils have their own “data sharing policies”, but others depend on the “common principles on on research data”[31]:

> Publicly funded research data are a public good and produced in the public interest. They should be made openly available with as few restrictions as possible in a timely and responsible manner.

Open Data mandates and UKRI

- Arts and Humanities Research Council (AHRC[32]: Relies on the UKRI common principles.
- Biotechnology and Biological Sciences Research Council (BBSRC[33]): “BBSRC expects that all data (with accompanying metadata) should be shared in a timely fashion as soon as it is verified”.
- Engineering and Physical Sciences Research Council (EPSRC[34]): Has the most explicit data policy, including a requirement for DOI.
- Economic and Social Research Council (ESRC[35]): Explicit requirement that “data **will** be made available […] as Open Data”.
- Medical Research Council (MRC): There’s an “expectation” from MRC[36] that data must be made open, they helpfully provide lots of advice about patient and population data.
- Natural Environment Research Council (NERC[37]): Flubs it by saying all research “must include a statement on how the supporting data and any other relevant research materials can be accessed”
- Science and Technologies Facilities Council (STFC[38])“:”STFC expects that published data should be made publicly available within six months of publication unless justified otherwise”

Understanding the whole process

1. Submit a Data Management Plan (DMP) to a funder, including a data sharing plan
2. Pre-register your research (often simultaneously with step 1)
3. Reserve a DOI to store your research data at a data repository
4. Do the research

- Keep raw data pristine. Do not modify your raw data.
- Keep track of how you wrangle the data (much easier when you write R code!)
- Craft an anonymised, shareable version of your datasets described in your DMP.

5. Write up the study
6. Choose a journal to submit to

> That’s an entire process in of itself!

7. Make the data deposit public when your research is made public

- Applying an embargo can help automate this process

You might want to create follow up studies or new studies where you add to the existing data deposit.

This works nicely thanks to DOI versioning!

Understanding the whole process (more!)

… what about code?

> So far I’ve only spoken about the data element of reproducibility, we’ll get onto the code in the workshop.

Digital Object Identifiers (DOI)

DOI are extremely important to ensure research availability into the future.

Academic journal links are fragile and could change at any time:

> sciencedirect.com/science/article/pii/S2665927122000879

DOI are **persistent and extremely long-term** identifiers that look like this:

> 10.1016/j.crfs.2022.05.015

The publisher and the DOI Foundation are then responsible for directing you to the resource by constructing a URL like this:

> doi.org/10.1016/j.crfs.2022.05.015

DOI for more than just publications

Initially DOI were only issued by academic publishers to resolve journal articles.

Data Repositories started issuing DOI so we could resolve links to data, code and more.

Specialist Data Repositories

Sometimes you **need** a repository with specialist features, eg:

- Genome sequences
- Protein sequences
- Climate data

Nature Publishing Group[39] provides an excellent overview of these tools.

General Purpose Repositories

These tools all have slighty different advantages and disadvantages

- Figshare
- Zenodo
- Open Science Framework

Identifiers for researchers?

DOI are great for resolving our research outputs, what about uniquely identifying researchers?

Researchers often share the same names or change their names throughout their career.

The **only** open researcher identifier is ORCID.

Thankfully - it also works really well!

It keeps track of all publications and deposits on data repositories.

Here's mine: orcid.org/0000-0002-3039-6849.

ORCID

stands for

Open Researcher and Contributor ID

> I want you to register for an ORCID now so you can use it for everything - including in your CV

General purpose repositories

📝 Task: Setup a Collection on Figshare

SLIDE 1 OF 1

Everything we’ve spoken about has been very theoretical. I want you to go through the steps of creating a collection on Figshare.

> We’re creating a Collection because it can contain multiple Figshare items. At the beginning of your research you likely don’t know exactly how many data files you’ll end up with.

- Sign up for Figshare with your ORCID
- Go to “My data” in Figshare
- Go to “Collections”
- Create a Collection
- Reserve a DOI
- Go to “My data”
- Add a new item
- Reserve a DOI
- Go to your collection and add this data item.

Figshare for talks and more

Figshare is very much a general purpose repository.

If you create something you want to make available for the future the best thing you can do is get a DOI.

- Consider using Figshare for presentations
- Consider using Figshare for posters

We need to talk about Open Access

We need to talk about Open Access

It used to be extremely hard to access the research that UKRI funds - despite it being funded by public money.

The Open Access movement began in the 90s and is ever growing.

It’s now a requirement of UKRI[31] funding that

> the final Version of Record or the Author’s Accepted Manuscript must be free to view and download via an online publication platform, publishers’ website, or institutional or subject repository within a maximum of 12 months of publication

… however, this often means that **someone** is paying an Article Processing Charge (APC).

Different routes to Open Access

APCs can be split between publishers and authors in different ways.

UKRI is usually responsible for paying the author’s portion of the bill.

- Gold Open Access: All articles in a journal are Open Access.
- Hyrbrid Open Access: Specific articles in a journal are made Open Access through APCs. Journals receive money through both subscriptions and APCs.

No APC is paid

The most important category here is **Green Open Access**.

In Green Open Access the author self-archives their article in a publicly available repository.

This gets complicated.

- Some publishers require that only pre-prints are self-archived
- Some publishers allow post-print publishing.

If you’re interested read Gadd and Troll Covey[40].

Negative Results

Positive Publication Bias

There is a significant and very clear bias to researchers publishing “positive results”[41] - which you can even see in article titles.

This poses significant issues in the literature.

- It’s really useful to know X doesn’t work! It means others don’t need to repeat the result.
- Negative effects can be under reported, as per the GSK lawsuit in 2004.
- A lack of negative results introduces bias to meta-analyses.

The MRC Open Research Data policy[36] explicitly requires both positive and negative results of studies be published within 24 months of the trial end.

Publishing negative results

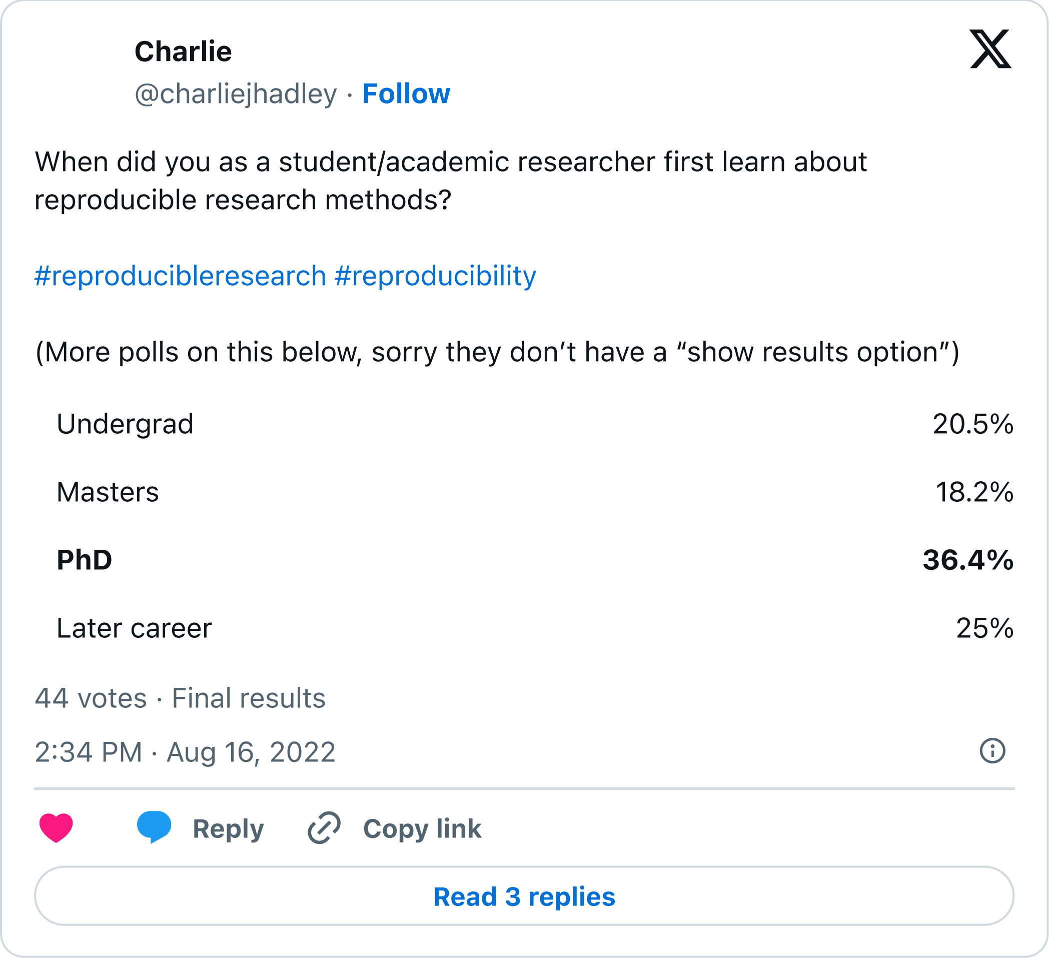
This is an open problem.

There was a push for new negative result journals in the early 2010s but several of these folded, including the Journal of Negative Results in Biomedicine The remaining journals have very low impact factors as they are published by smaller publishers.

In general the journals with the highest impact factors are getting better at publishing negative results.

Why are we learning all of this now?

```r
tweetrmd::tweet_embed("https://twitter.com/charliejhadley/status/1559534088838647808?s=20&t=M0f0BvpbqLiKpDxBH
```



Charlie

@charliejhadley · **Follow**

When did you as a student/academic researcher first learn about reproducible research methods?

#reproducibleresearch #reproducibility

(More polls on this below, sorry they don't have a "show results option")

Undergrad	20.5%
Masters	18.2%
PhD	**36.4%**
Later career	25%

44 votes · Final results

2:34 PM · Aug 16, 2022

♥ Reply Copy link

Read 3 replies

References

1. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Medicine* **2**, e124 (2005).
2. Douglas Heaven, W. Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review* (2021).
3. Broman, K. *et al. Recommendations to Funding Agencies for Supporting Reproducible Research*. 1–4 (2017).
4. Fanelli, D. How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLOS ONE* **4**, e5738 (2009).
5. John, L. K., Loewenstein, G. & Prelec, D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science* **23**, 524–532 (2012).
6. Aschwanden, C. We're All 'P-Hacking' Now. *Wired* (2019).
7. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* **22**, 1359–1366 (2011).
8. Hu, Y., Liu, H., Wu, J. & Fang, G. Factors influencing self-care behaviours of patients with type 2 diabetes in China based on the health belief model: A cross-sectional study. *BMJ Open* **12**, e044369 (2022).
9. Clifton, C. *et al.* Eye movements in reading and information processing: Keith Rayner's 40year legacy. *Journal of Memory and Language* **86**, 1–19 (2016).
10. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly work in Medical Journals (ICMJE). (2022).
11. World Medical Association. Declaration of Helsinki, Ethical Principles for Medical Research involving human subjects. (2013).
12. Bradley, S. H., Lloyd, K. E. & DeVito, N. J. Automatic registration for UK trials. *BMJ* **376**, o41 (2022).
13. Bruckner, T. UK launches new system to achieve 100% clinical trial registration (updated). *transparimed* (2021).
14. Dyer, O. GlaxoSmithKline faces US lawsuit over concealment of trial results. *BMJ : British Medical Journal* **328**, 1395 (2004).
15. Spurgeon, D. GlaxoSmithKline staff told not to publicise ineffectiveness of its drug. *BMJ : British Medical Journal* **328**, 422 (2004).
16. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences* **115**, 2600–2606 (2018).

17. Dreber, A. *et al.* Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences* **112**, 15343–15347 (2015).
18. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* **2**, 637–644 (2018).
19. Tedersoo, L. *et al.* Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data* **8**, 192 (2021).
20. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology* **13**, e1002295 (2015).
21. Tenopir, C. *et al.* Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE* **6**, e21101 (2011).
22. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. & McGillivray, B. The citation advantage of linking publications to research data. *PLOS ONE* **15**, e0230416 (2020).
23. DATA in the time of COVID-19. *Open Data Watch* (2020).
24. Roth, J., Lim, B., Jain, R. K. & Grueneich, D. Examining the feasibility of using open data to benchmark building energy usage in cities: A data science and policy perspective. *Energy Policy* **139**, 111327 (2020).
25. CDC - BRFSS Annual Survey Data. (2021).
26. The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge. (2022).
27. UK Government. Open Government Licence. (2022).
28. World Bank. Open Data Essentials I Data. (2021).
29. UK Government. Data Protection Act 2018. (2018).
30. Information Commissioner's Office. Health data. (2021).
31. UKRI. UKRI Common Principles on Research Data. (2022).
32. AHRC. AHRC Research Funding Guide. (2022).
33. BBSRC. BBSRC Data Sharing Policy. (2022).
34. EPSRC. ESPRC Policy Framework on Research Data. (2022).
35. ESRC. ESRC Research Data Policy. (2022).
36. MRC. MRC Open Research Data Advice. (2022).
37. NERC. NERC Data Policy. *NERC Data Policy* (2022).
38. STFC. STFC Open Data Policy. (2021).
39. Nature. Data Repository Guidance I Scientific Data. (2022).
40. Gadd, E. & Troll Covey, D. What does "green" open access mean? Tracking twelve years of changes to journal publisher self-archiving policies. *Journal of Librarianship and Information Science* **51**, 106–122 (2019).
41. Mlinarić, A., Horvat, M. & Smolčić, V. Šupak. Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia Medica* **27**, 030201 (2017).