Webinar Series

Improving Research Integrity:
Managing and Sharing Research Data

06 October | SGT 11AM

Keith Russell
Manager, Engagements,
Australian Research Data Commons
Increasing integrity of your research: managing and sharing research data

PRESENTED BY
Keith Russell, Manager, Engagements, 6 October 2020
NCRIS

The National Collaborative Research Infrastructure Strategy (NCRIS) drives research excellence and collaboration between 40,000 researchers, government and industry to deliver practical outcomes.

3.3 billion since 2004
Coronavirus update: rapid sharing of data vital

Cate Swannell
Med J Aust
Published online: 9 March 2020

THE Australian laboratory which was the first to isolate the COVID-19 virus outside China has shared its methods in a research letter published online today by the Medical Journal of Australia as part of the Journal’s rapid review policy for coronavirus data.

Professor Mike Catton, Deputy Director of the Peter Doherty Institute for Infection and Immunity, described the case of a 58-year-old man who arrived in Melbourne from China on 19 January 2020 and was admitted to hospital on 24 January with “symptoms of fever, cough, and progressive shortness of breath”.

“A nasopharyngeal swab and sputum taken on presentation tested positive for SARS-CoV-2,” Catton and colleagues wrote. “Inoculation with material from the initial nasopharyngeal swab led to the isolation of SARS-CoV-2 virus in culture. Electron microscopy of the supernatant confirmed the presence of virus particles displaying morphology characteristic of the family Coronaviridae.

“Whole genome sequencing of the viral isolate ... revealed that the genome sequence from this patient exhibited >99.99% sequence identity to other publicly
Two papers relying on hospital records of COVID-19 patients have been retracted because the company that purportedly analyzed the raw data won't allow their validity to be independently validated. AP PHOTO/MANU FERNANDEZ

Two elite medical journals retract coronavirus papers over data integrity questions

By Charles Piller, Kelly Servick | Jun. 4, 2020, 5:36 PM

Science's COVID-19 reporting is supported by the Pulitzer Center.

In the first big research scandal of the COVID-19 era, The Lancet and The New England Journal of Medicine (NEJM) today retracted two high-profile papers after a company declined to make the underlying data for both available for an independent audit, following questions being raised about the research. The Lancet paper, which claimed an antimalarial drug touted by President Donald Trump for treatment of COVID-19 could cause serious harm without helping patients, had had a global impact, halting trials of one of the drugs by the World Health Organization (WHO) and others.

Three authors on the Lancet paper requested the retraction, after initiating an independent review of the raw hospital patient data summarized and provided by Surgisphere, a small Chicago-based company that was paid $12 million by Merck to analyze its COVID-19 drug remdesivir. The independent review could not verify the results, the Lancet authors said.
Why is properly managing and sharing data so crucial?
IS THERE A REPRODUCIBILITY CRISIS?

- 7% Don’t know
- 3% No, there is no crisis
- 38% Yes, a slight crisis
- 52% Yes, a significant crisis

1,576 researchers surveyed

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
No raw data, no science: another possible source of the reproducibility crisis

Tetsuya Miyakawa

Abstract

A reproducibility crisis is a situation where many scientific studies cannot be reproduced. Inappropriate practices of science, such as HARKing, p-hacking, and selective reporting of positive results, have been suggested as causes of irreproducibility. In this editorial, I propose that a lack of raw data or data fabrication is another possible cause of irreproducibility.

As an Editor-in-Chief of Molecular Brain, I have handled 180 manuscripts since early 2017 and have made 41 editorial decisions categorized as "Revise before review," requesting that the authors provide raw data. Surprisingly, among those 41 manuscripts, 21 were withdrawn without providing raw data, indicating that requiring raw data drove away more than half of the manuscripts. I rejected 19 out of the remaining 20 manuscripts because of insufficient raw data. Thus, more than 91% of the 41 manuscripts did not present the raw data supporting their results when requested by an editor, suggesting a possibility that the raw data did not exist from the beginning, at least in some portions of these cases. Considering that any scientific study should be based on raw data, and that data storage space should no longer be a challenge, journals, in principle, should try to have their authors publicize raw data in a public database or journal site upon the publication of the paper to increase reproducibility of the published results and to increase public trust in science.

Keywords: Raw data, Data fabrication, Open data, Open science, Misconduct, Reproducibility

Introduction

The reproducibility or replicability crisis is a serious issue in which many scientific studies are difficult to reproduce or replicate. It is reported that, in the field of cancer research, only about 20–25% [1] or 11% [2] of published studies could be validated or reproduced, and that only about 36% were reproduced in the field of psychology [3]. Inappropriate practices of science, such as HARKing (Hypothesizing After the Results are Known) [4], p-hacking [5], selective reporting of positive results and poor research design [6–8], have been proposed to be a cause of such irreproducibility. Here, I argue that a lack of raw data is another serious possible cause of irreproducibility, by showing the results of analyses on the manuscripts that I have handled over the last 2 years for Molecular Brain. The analysis shows that many researchers did not provide the raw data, suggesting that raw data may not exist in some cases and that the lack of data may constitute a non-negligible part of the causes of the reproducibility crisis [9]. In this editorial, I argue that making raw data openly available is not only important for reuse and data mining but also for simply confirming that the results presented in the paper are truly based on actual data. With such concept, the data sharing policy of Molecular Brain has been changed and I introduce this update.

Raw data rarely comes out

As Editor-in-Chief of the journal, I have handled 180 manuscripts since early 2017 to September 2019 and have made 41 editorial decisions categorized as "Revise before review", with comments asking the authors to provide raw data (Fig. 1). See Additional file 2: Table S1 for details.
Funders recognise data as an essential output

- Funders are seeing research data as a valuable output from the funding
- They expect data to be managed
- They expect it to be available to enable verification of findings and further research
What are publishers saying about FAIR?

January 2019 - *Nature* and *Scientific Data* required that authors of papers in these disciplines make supporting data available to others through community repositories where available.

Nature was not alone

Signatories to the Enabling FAIR Data Commitment Statement were many publishers including but not limited to:

So how to share in a valuable way?
F.A.I.R. Data Principles

- Drafted in a Lorentz center workshop in 2014 and in Nature article
- Received international recognition
- Making data usable by humans and machines
- Technology agnostic
- Both the data and the metadata
- Discipline independent...
... but how to put this into practice?
Step 1: Plan ahead

Consider writing a Data Management plan, this will prepare you for steps further on in the research process.
Step 2: Use standard data formats

As you conduct your research try to use community agreed standard data formats and metadata formats

Capture choices/settings along the way off instruments/tools (e.g. calibration settings)

Keep references to reference data sets and tools that you are using including the version used

Document decisions you make when manipulating the data and keep this to publish alongside the data
Step 3: Make your data Findable ....

Deposit your data in a repository, preferably a CoreTrustSeal certified one.

- That will ensure you get a DOI (so you can link from your article and get citations)

- That will ensure you have adequate metadata

https://www.re3data.org/
Step 4: ... and that will make it Accessible

Not all data can be made Open due to privacy, national security constraints or commercial interests ... 

... but that does not mean it cannot be Accessible

There are repositories that allow for mediated access

But do clearly describe how access needs to be requested
Step 5: and make it Reusable

Choose and assign a standard usage licence

We recommend [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)

And publish supporting information alongside the data (codebooks, provenance information, workflows and protocols, syntax used in the your tool)
You don’t have to start from scratch

Ask your library for help

What do journals require?

What does your funder require?

Are there discipline repositories?

Does your institution have a repository?

Use Figshare, Dryad, Mendeley Data, Zenodo
... and if you think you are almost there
