On the Importance of Data Transparency

On June 13, 2020, the Lancet issued a retraction for the paper “Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis” after an independent third-party peer review to check on the veracity of the data and analyses conducted by Surgisphere Corporation. The Lancet was informed by the independent reviewers “…that Surgisphere would not transfer the full dataset, client contracts, and the full ISO audit report to their servers for analysis as such transfer would violate client agreements and confidentiality requirements.” The Lancet therefore issued the following statement: “Based on this development, we can no longer vouch for the veracity of the primary data sources. Due to this unfortunate development, the authors request that the paper be retracted” (https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31324-6/fulltext).

The New England Journal of Medicine issued a similar retraction to another paper based on the Surgisphere database on June 4, 2020. On behalf of the authors, the retraction stated: “Because all the authors were not granted access to the raw data and the raw data could not be made available to a third-party auditor, we are unable to validate the primary data sources underlying our article, ‘Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19’” (https://www.nejm.org/doi/full/10.1056/NEJMct2021225).

These retractions have had significant ramifications for several clinical trials, which were immediately stopped and have now since restarted (https://time.com/5847664/who-hydroxychloroquine-covid-19/), resulting in people losing confidence in medical testing, how science is done, and correspondingly, research’s ability to solve the problems we’re facing.

Unfortunately, this is not a new story for data professionals or academic publishers, who have seen many retractions made because of concerns with data. The Retraction Watch database currently contains 1,502 entries tagged with “Concerns/Issues About Data” as of June 17, 2020 (http://retractiondatabase.org/RetractionSearch.aspx). The high-profile nature of these retractions, however, has raised awareness about issues with the transparency, reproducibility, and verifiability of research in a wider space than before.

Science is the act of formulating a hypothesis, then coming up with a method to test that hypothesis, often by making measurements and, hence, collecting data. If an experiment can’t be repeated, the conclusions from that experiment cannot be verified, and therefore, the likelihood of the conclusions being correct is low. This isn’t a revolutionary thing to say in the slightest.

As science has progressed, experiments, data collection, and analysis have all gotten increasingly complicated and reliant on technology. Modern studies rely on massive databases that are physically impossible to replicate in exactly the same way with new measurements. No one has the ability or, for that matter, funding to build a copy of the Large Hadron Collider in a different location in order to verify the LHC’s results from scratch. Similarly, for observational data, like medical records or weather data, it’s not possible to retake measurements from the past. For the conclusions of a study to stand, the underlying data must be made available for scrutiny by qualified professionals with appropriate expertise.

The data science community uses the acronym FAIR to discuss how data should be managed and preserved. FAIR stands for: findable, accessible, interoperable, and reusable (https://www.go-fair.org/fair-principles/).

The FAIR principles are closely related. Findability and accessibility are the first mentioned, because it’s no use to have a wonderfully reusable and interoperable dataset locked away in a drawer where no one can find or use it. The other side of this coin is when dataset creators make their data open (usually as a result of funder mandate) while deliberately obfuscating them to prevent reuse and interoperability.

Data Must Be Accessible

The technological challenges to sharing data are decreasing as time goes by, although this is countered to a certain extent by the increase in the volume of data being created. No longer do researchers have to share data by burning copies of CDs and putting them in the post. Instead data can be shared using a click of the mouse, in real time and with a rapidity that would be astonishing if it weren’t so commonplace.

When it comes to medical data, technology is not the limiting factor. Issues of confidentiality and privacy are rightly a major concern and should be respected. A major part of Patterns’ ethos is the understanding that behind the data are real people, with the potential to be really harmed if their data are misused. That is why safeguards and restrictions need to be put in place for sensitive data and why accessibility needs to be restricted to those with a “need to know.”

Plenty of valuable scientific research has been proposed and built on using proprietary and closed datasets—that isn’t the issue. The conclusions based on closed data are maintained because of selective accessibility, i.e., the fact that expert peer reviewers were permitted access to the data in order to check it. In these cases, the reviewers act as proxy for the rest of the community, providing quality control checks and assurances that the data do indeed back up what the conclusions say.

Reviewers need to be given access to data as standard. Yes, there are significant and complicated legal aspects associated with medical and business data, including client agreements and confidentiality requirements, and intellectual property rights. For a paper to be published in the scientific literature, these issues need to be addressed if the results are to be trusted. If the data and the results cannot be trusted, then they should not be published.

Data Must Be Reusable

No one can be an expert at everything. Data collection, annotation, and validation is a complicated and detailed business and is often unrewarded in existing academic recognition structures.
This situation is slowly changing, with journals offering the ability to publish descriptor articles about research objects like data-sets and code in order to make their use and production more transparent and understandable.

Data must be made reusable to the community that is most likely to use it, which includes those who are most likely to be peer reviewing it. Putting documentation together, annotating data, acknowledging and fixing problems in a dataset—these all take time and effort but are necessary and important work.

Researchers acting as peer reviewers don’t have to do the quality control work of checking data in isolation, as data scientists have the knowledge and ability to quickly analyze data and identify the outliers that might indicate problems with the dataset. Researchers and data scientists do have to work in collaboration, however. An example of a potential way a domain expert could work with a data scientist to validate data is as follows: a data scientist would be able to find the negative numbers in a massively long time series, but it’s the domain expert who would know that negative numbers in this particular stream are a sign that the data are corrupt, because negative numbers for this particular quantity are physically impossible.

Common standards make data more understandable. Common tools and services make data more usable. Making these things open allows others to build on them without re-inventing the wheel, instead connecting data and services together in a modular way.

How Do We Make Science More Robust?
Fundamentally, it’s by acknowledging that everyone who works in science has the responsibility to ensure that science is carried out effectively and ethically. It is by changing research culture so that the issues around FAIR data are discussed early in projects and steps are taken to allow people with bona fide credentials and reasons to access it, including peer reviewers. It is by quite simply taking more care in checking, even knowing that these checks will cost us time and effort in an increasingly more hectic world.

We all need to shoulder this burden. Researchers have the responsibility to ensure that their conclusions are backed up by their data and that their data are in a state where it is easy for those with domain expertise to understand. Peer reviewers must ask questions about the data in their reviews and dig a little deeper into the databases, if something doesn’t seem right. Journal editors need to ensure that these checks on data are carried out before publication and that the policies on data accessibility statements are adhered to.

The lack of transparency around data and the resulting retraction of peer-reviewed papers show that we cannot afford to ignore everyday issues regarding accessibility of data any longer. The good news is that these issues are well understood and policies are already in place to deal with them. What we need to do now is find ways of making it easy and quick to abide by those policies, and that will require time, investment, and a willingness to engage from the entire research community.

The road to scientific transparency is long, but we’re already on our way.

WEB RESOURCES
FAIR Principles, https://www.go-fair.org/fair-principles/
“Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. https://doi.org/10.1056/NEJMo2007621,” https://www.nejm.org/doi/full/10.1056/NEJMc2021225
“Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis,” https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31324-6/fulltext
Retraction Watch Database, http://retractiondatabase.org/Retraction Search.aspx
“WHO Resumes Study of Hydroxychloroquine for Treating COVID-19,” https://time.com/5847664/who-hydroxychloroquine-covid-19/

Sarah Callaghan
Editor-in-Chief, Patterns
https://doi.org/10.1016/j.patter.2020.100070