Computational Reproducibility in Medical Research: Toward Open Code and Data

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

R / Medicine
Yale University
September 8, 2018
Today: Technology is driving a reassessment of transparency

1. Big Data / Data Driven Discovery: high dimensional data, $p >> n$,

2. Computational Power: simulation of the complete evolution of a physical system, systematically varying parameters,

3. Deep intellectual contributions now encoded only in software.

*The software contains “ideas that enable biology…”*

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

“Stories from the Supplement” from the Genome Informatics meeting 11/1/2013
https://youtu.be/5NiFibnbE8o
The digital age in science

Claim 1:
Virtually all published discoveries today have a computational component.

Claim 2:
There is a mismatch between the traditional scientific process and computation, leading to reproducibility concerns.
Parsing Reproducibility

“Empirical Reproducibility”

“Statistical Reproducibility”

“Computational Reproducibility”

V. Stodden, IMS Bulletin (2013)
Empirical Reproducibility

Sorting Out the FACS: A Devil in the Details

William C. Hines, Ying Su, Irene Kuhn, Kornelia Polyak, and Mina J. Bissell

The reproduction of results is the cornerstone of science; yet, at times, reproducing the results of others can be a difficult challenge. Our two laboratories, one on the East and the other on the West Coast of the United States, decided to collaborate on a problem of mutual interest—namely, the heterogeneity of the human breast. Despite using seemingly identical methods, reagents, and specimens, our two laboratories quite reproducibly were unable to replicate each other’s fluorescence-activated cell sorting (FACS) profiles of primary breast cells. Frustration of studying cells close to their context in vivo makes the exercise even more challenging.

Paired with in situ characterizations, FACS has emerged as the technology most suitable for distinguishing diversity among different cell populations in the mammary gland. Flow instruments have evolved from being able to detect only a few parameters to those now capable of measuring up to—and beyond—an astonishing 50 individual markers per cell (Cheung and Uth, 2011). As with any exponential increase in data complexity, breast reduction mammoplasties. Molecular analysis of separated fractions was to be performed in Boston (K.P.’s laboratory, Dana-Farber Cancer Institute, Harvard Medical School), whereas functional analysis of separated cell populations grown in 3D matrices was to take place in Berkeley (M.J.B.’s laboratory, Lawrence Berkeley National Lab, University of California, Berkeley). Both our laboratories have decades of experience and established protocols for isolating cells from primary normal breast tissues as well as the capabilities required for

Reproducibility Issues in Research with Animals and Animal Models

The missing “R”: Reproducibility in a Changing Research Landscape

A workshop of the Roundtable on Science and Welfare in Laboratory Animal Use

National Academy of Sciences, NAS 125
2100 C Street NW, Washington DC
June 4-5, 2014

The ability to reproduce an experiment is one important approach that scientists use to gain confidence in their conclusions. Studies that show that a number of significant peer-reviewed studies are not reproducible has alarmed the scientific community. Research that uses animals and animal models seems to be one of the most susceptible to reproducibility issues.

Evidence indicates that there are many factors that may be contributing to scientific irreproducibility, including insufficient reporting of details pertaining to study design and planning; inappropriate interpretation of results; and author, reviewer, and editor abstracted reporting, assessing, and accepting studies for publication.

In this workshop, speakers from around the world will explore the many facets of the issue and potential pathways to reducing the problems. Audience participation portions of the workshop are designed to facilitate understanding of the issue.
Statistical Reproducibility

- False discovery, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values, lack of multiple testing adjustments.
- Low power, poor experimental design, nonrandom sampling,
- Data preparation, treatment of outliers, re-combination of datasets, insufficient reporting/tracking practices,
- inappropriate tests or models, model misspecification,
- Model robustness to parameter changes and data perturbations,
- ...
In January 2014 Science enacted new manuscript submission requirements:

- a “data-handling plan” i.e. how outliers will be dealt with,
- sample size estimation for effect size,
- whether samples are treated randomly,
- whether experimenter blind to the conduct of the experiment.

Also added statisticians to the Board of Reviewing Editors.
Computational Reproducibility

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete set of instructions [and data] which generated the figures.

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

By Victoria Stodden,1 Marcia McNutt,2 David H. Bailey,3 Ewa Deelman,4 Yolanda Gil,4 Brooks Hanson,5 Michael A. Heroux,6 John P.A. Ioannidis,7 Michela Taufer8

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency of the computational process. To understanding how computational results were derived and to reconciling any differences that might arise between independent replications (4). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (5, 6), which includes workflow information that explains what raw data and intermediate results are input to which computations (7). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://

Access to the computational steps taken to process data and generate findings is as important as access to data themselves.

Reproducibility Enhancement Principles

1: To facilitate reproducibility, **share the data, software, workflows**, and details of the computational environment in open repositories.

2: To enable discoverability, **persistent links** should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

3: To enable credit for shared digital scholarly objects, **citation** should be standard practice.

4: To facilitate reuse, adequately **document** digital scholarly artifacts.

5: Journals should conduct a **Reproducibility Check** as part of the publication process and enact the TOP Standards at level 2 or 3.

6: Use **Open Licensing** when publishing digital scholarly objects.

7: Funding agencies should instigate **new research** programs and pilot studies.
<table>
<thead>
<tr>
<th>Standard</th>
<th>Level 0</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation standards</td>
<td>Journal encourages citation of data, code, and materials—or says nothing.</td>
<td>Journal describes citation of data in guidelines to authors with clear rules and examples.</td>
<td>Article provides appropriate citation for data and materials used, consistent with journal’s author guidelines.</td>
<td>Article is not published until appropriate citation for data and materials is provided that follows journal’s author guidelines.</td>
</tr>
<tr>
<td>Data transparency</td>
<td>Journal encourages data sharing—or says nothing.</td>
<td>Article states whether data are available and, if so, where to access them.</td>
<td>Data must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Analytic methods (code) transparency</td>
<td>Journal encourages code sharing—or says nothing.</td>
<td>Article states whether code is available and, if so, where to access them.</td>
<td>Code must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Research materials transparency</td>
<td>Journal encourages materials sharing—or says nothing</td>
<td>Article states whether materials are available and, if so, where to access them.</td>
<td>Materials must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Design and analysis transparency</td>
<td>Journal encourages design and analysis transparency or says nothing</td>
<td>Journal articulates design transparency standards.</td>
<td>Journal requires adherence to design transparency standards for review and publication.</td>
<td>Journal requires and enforces adherence to design transparency standards for review and publication.</td>
</tr>
<tr>
<td>Preregistration of studies</td>
<td>Journal says nothing.</td>
<td>Journal encourages preregistration of studies and provides link in article to preregistration if it exists.</td>
<td>Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.</td>
<td>Journal requires preregistration of studies and provides link and badge in article to meeting requirements.</td>
</tr>
<tr>
<td>Preregistration of analysis plans</td>
<td>Journal says nothing.</td>
<td>Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.</td>
<td>Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.</td>
<td>Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.</td>
</tr>
<tr>
<td>Replication</td>
<td>Journal discourages submission of replication studies—or says nothing.</td>
<td>Journal encourages submission of replication studies.</td>
<td>Journal encourages submission of replication studies and conducts blind review of results.</td>
<td>Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.</td>
</tr>
</tbody>
</table>
The LifeCycle of Data Science as a Framework
Lifecycle of Data

Berman et al., “Realizing the Potential of Data Science,” CACM, April 2018
Lifecycle of *Data Science*

- Framework to incorporate data science contributions from different fields,
- Explicit emphasis on re-use and reproducibility,
- Explicit emphasis on computational tools (e.g. Kubernetes), hardware (e.g. Google Edge TPUs) and software (e.g. Jupyter Notebooks)
- Surfaces ethics (human subjects, privacy), social context (interpretations of “bias”), scholarly communication and reproducible research.
Lifecycle of Data Science: An Abstraction

- Data generation and collection
- Data exploration and hypothesis generation
- Data cleaning and organization
- Feature selection and data preparation
- Model building and statistical inference
- Simulation and cross-validation
- Notebooks
- Notebooks, visualization software
- Workflow software, artifact linking tools

- System level: hardware, cloud computing infrastructure, systems and system management, data structures, storage
- Infrastructure level: notesbooks and workflow software; database structures; workflow software and preregistration tools; data management tools
- Application level: experimental design; data generation and collection; data exploration and hypothesis generation; data cleaning and organization

- The science of data science: ethics, documentation and metadata creation, best practices, policy;
  - The study of data science: the science of data science
Progress on computational reproducibility is enabled through coordination by a variety of stakeholders.

- **Scientific Societies**
- **Funders** (policy)
- **Publishers** (TOP guidelines)
- **Regulatory Bodies** (OSTP Memos)
- **Researchers** (processes)
- **The Public/Press**
- **Universities/institutions** (hiring/promotion)
- **Universities/libraries** (empowering w/tools, support)
Does artifact access on demand work?

February 11, 2011:

“All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. All computer codes involved in the creation or analysis of data must also be available to any reader of Science. After publication, all reasonable requests for data and materials must be fulfilled.”


- Obtained a random sample of 204 scientific articles with computational findings. Asked for the data and code!

Stodden et al., “Journal Policy for Computational Reproducibility,” PNAS, March 2018
<table>
<thead>
<tr>
<th>Response</th>
<th>Count</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No response</td>
<td>46</td>
<td>26%</td>
</tr>
<tr>
<td>Email bounced</td>
<td>3</td>
<td>2%</td>
</tr>
<tr>
<td>Impossible to share</td>
<td>3</td>
<td>2%</td>
</tr>
<tr>
<td>Refusal to share</td>
<td>12</td>
<td>7%</td>
</tr>
<tr>
<td>Contact to another person</td>
<td>20</td>
<td>11%</td>
</tr>
<tr>
<td>Asks for reasons</td>
<td>20</td>
<td>11%</td>
</tr>
<tr>
<td>Unfulfilled promise to follow up</td>
<td>5</td>
<td>3%</td>
</tr>
<tr>
<td>Direct back to SOM</td>
<td>6</td>
<td>3%</td>
</tr>
<tr>
<td>Shared data and code</td>
<td>65</td>
<td>36%</td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td>100%</td>
</tr>
</tbody>
</table>

24 articles provided direct access to code/data.
Replicating Computational Findings

• We deemed 56 of the 89 articles for which we had data and code potentially reproducible

• We chose a random sample of 22 from these 56 to replicate
Computational Replication Rates

We were able to obtain data and code from the authors of 89 articles in our sample of 204,

- overall **artifact recovery rate** estimate: 44% with 95% confidence interval [0.36, 0.50]

Of the 56 potentially reproducible articles, we randomly choose 22 to attempt replication, and all but one provided enough information that we were able to reproduce their computational findings.

- overall **computational reproducibility** estimate: 26% with 95% confidence interval [0.20, 0.32]
When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.

The data files remain our property and are not deposited for free access. Please, let me know the purpose you want to get the file and we will see how we can help you.

We do not typically share our internal data or code with people outside our collaboration.

The code we wrote is the accumulated product of years of effort by [redacted] and myself. Also, the data we processed was collected painstakingly over a long period by collaborators, and so we will need to ask permission from them too.

Normally we do not provide this kind of information to people we do not know. It might be that you want to check the data analysis, and that might be of some use to us, but only if you publish your findings while properly referring to us.
Thank you for your interest in our paper. For the [redacted] calculations I used my own code, and there is no public version of this code, which could be downloaded. Since this code is not very user-friendly and is under constant development I prefer not to share this code.

I’m sorry, but our computer code was not written with an eye toward distributing for other people to use. The codes are not documented and we don’t have the time or resources to document them. If you have a particular calculation you would like done and it is not a major extension of what we are presently set up to do, we might be able to run the codes for you.

R is a free software package available at www.r-project.org/ I used R for the [redacted] models. As you probably know, [redacted] and [redacted] are quite complicated. But I don’t have to tell you that given that you are a statistics student! I used Matlab for the geometry.
Our program [redacted] is available here [URL redacted] (documentation and tutorials were included)

If you go to [URL redacted], under the publications, I have a link to the gitHub repository. I don’t know if I have all of the raw simulated data, but I certainly have the processed data used to make the plots. What do you need? All of the simulated data could of course be regenerated from the code.

Please find attached a .zip file called [redacted].zip that has the custom MATLAB [redacted] analysis code. If you run Masterrunfigure-one.m this will generate several panels from the paper.

In the next email I will enclose the custom image analysis software. This can also be accessed from [URL redacted] where there is a manual and tutorial.

Please let me know if you have any troubles, or if there is anything else I can help with.