Democratizing Health Data for Translational Research

Philip R.O Payne
Institute for Informatics, Washington University School of Medicine
St. Louis, MO, USA
Email: prpayne@wustl.edu

Nigam H. Shah
Center for Biomedical Informatics Research, Stanford University School of Medicine
Stanford, CA, USA
Email: nigam@stanford.edu

Jessica D. Tenenbaum
Department of Biostatistics and Bioinformatics, Duke University School of Medicine
Durham, NC, USA
Email: jessie.tenenbaum@duke.edu

Lara Mangravite
Sage Bionetworks
Seattle, WA, USA
Email: lara.mangravite@sagebase.org

There is an expanding and intensive focus on the accessibility, reproducibility, and rigor of basic, clinical, and translational research. This focus complements the need to identify sustainable ways to generate actionable research results that improve human health. The principles and practices of *open science* offer a promising path to address both issues by facilitating: 1) increased transparency of data and methods which promotes research reproducibility and rigor; and 2) cumulative efficiencies wherein research tools and the output of research are combined to accelerate the delivery of new knowledge. While great strides have been in made in terms of enabling the open science paradigm in the biological sciences, progress in sharing of patient-derived health data has been more moderate. This lack of widespread access to common and well characterized health data is a substantial impediment to the timely, efficient, and multi-disciplinary conduct of translational research, particularly in those instances where hypotheses spanning multiple scales (from molecules to patients to populations) are being developed and tested. To address such challenges, we review current best practices and lessons learned, and explore the need for policy changes and technical innovation that can enhance the sharing of health data for translational research.

Keywords: Open Science, Open Data, Translational Research, Data Science

1. Introduction

There is an emergent national and international dialogue concerned with the accessibility, reproducibility, and rigor of all types of biomedical research. Simultaneously, it has been recognized that the scientific community needs new approaches to make its work sustainable during times of both decreased funding and increased demand for timely and actionable outcomes of research programs. One potential solution to both of these challenges is the adoption of *open science* models that allow: 1) increased transparency of data and methods, which promotes research reproducibility and rigor [1-4]; and 2) cumulative efficiencies wherein research tools and the output of research are combined to accelerate the delivery of new knowledge [5-7]. For the purposes of this manuscript, we provide the following working definition for open science:

"Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods."[8]

Unfortunately, contradictory and sometimes conflicting positions on open science - and the way the open science paradigm might best be operationalized - demonstrate the need for greater community engagement to test the theory that open science in the health sciences can indeed improve the rigor and efficiency of research. This challenge is exemplified by the recent controversy regarding research "parasites" [9], and the vigorous debate that ensued as a result. Furthermore, while there have been some notable and early successes in terms of enabling open science frameworks, such as the creation of data sharing "commons" or the enforcement of data sharing policies concomitant with formal publication of research results, most if not all of these efforts have been focused on biologic data sets such as those related to either: 1) –omics focused measurements of bio-molecular phenomena with some small volume of associated clinical phenotype annotations; or 2) limited scale and highly synthesized data derived from clinical trials of new therapies or diagnostic methods [10-14]. Often, these data are used as a "reference" for secondary interrogation, for example, in studies involving the derivation of phenotypic "signatures" that correlate disease risk, severity, or potential response to therapy [15-19], or for the identification of candidates for drug repositioning/repurposing [20-22], to name a few of many potential examples.

In contrast, the wide-spread and comprehensive sharing of "reference" data sets containing patient-derived health data, such as that found in Electronic Health Records (EHRs), Clinical Research Management Systems (CRMS), or Electronic Data Capture (EDC) systems remains far less prevalent or developed. A number of rationales have been given for this lack of sharing and re-use of patient-derived data, including concerns surrounding patient consent and privacy [10, 23], the mechanisms for attribution of such data and its sources [12-14, 24], and uncertainty surrounding the quality and completeness of data sets collected for primarily clinical or administrative purposes [11, 13, 25]. Unfortunately, in the absence of "reference" data sets that include adequate amounts of patient-derived health data and that are well annotated and understood from a content, quality, and

provenance standpoint, we risk substantial inefficiencies as well as the potential for non-reproducible research, where such studies involve the development of novel ways to reason across multiple scales of phenotype, a situation that is intrinsic to what we often refer to as translational research.

Emergent efforts such as All of Us (formerly the Precision Medicine Initiative) in the United States [26] and the Observational Health Data Science and Informatics (OHDSI) that seek to create mechanisms and best practices for the sharing of patient-derived health data [27], as well as an increasing emphasis on the creation and maintenance of registries for "Real World Evidence" (RWE) generation by both diseases focused groups and biotechnology and pharmaceutical firms [2, 3, 7], provide the basis for a path forward. However, all of the preceding efforts remain both formative and early in their development. As such***, there remain substantial and unanswered questions concerning how to achieve a vision of "democratized" health data for translational research*** wherein all of the preceding challenges and opportunities have been adequately addressed. Ideally, such a vision of "democratized" health data would involve the adoption and widespread use of open science approaches that include a full spectrum of data types and assets.

2. Background

Never before has the health and life sciences communities been able to access such a wide variety of open data resources. Such data sets include measurements at the bio-molecular, clinical, and population levels, and are often derived from observational, clinical, and broader public health studies, not to mention an ever-increasing number of patient-reported indicators of health and wellness as well as sensors and other ubiquitous computing sources [4, 7, 11-13, 28, 29]. When viewed as a whole, these open data resources represent an opportunity for a paradigm shifting approach to discovery science, one in which we move away from the collection and curation of high-cost and project-specific data sets in which a small number of hypotheses are tested, and towards a model in which large-scale and heterogeneous data sets are collected, integrated, shared, and interrogated in a high throughput manner. This **open science** paradigm has the potential to substantially increase the speed and impact of research, while also reducing costs and barriers to answering critical questions by making the pursuit of such question cumulative in nature [7, 14, 30].

Given the promise of open science, one must ask why such an approach is not more common and widespread. Unfortunately, there exist a number of notable impediments in the contemporary scientific environment that preclude or inhibit the pursuit of open science, including:

- Confusing, overlapping, or contradictory **regulatory frameworks** governing the sharing of and access to research data, particularly data derived from humans;
- Misaligned **incentives and "community standards"** corresponding to research **career development and peer recognition**; and
- The dearth of **suitable platforms, technologies, and best practices** that can serve to support or enable the pursuit of open science by geographically, temporally, or otherwise distributed research teams that span traditional organizational boundaries and settings.

The papers and presentations associated with our session at the 2018 Pacific Symposium and Biocomputing (PSB 2018), entitled "**Democratizing Health Data for Translational Research**", explore each of these areas in further detail, particularly as they relate to implementing open science paradigms when seeking to understand the critical relationships between bio-molecular and clinical phenotypes in both health and wellness. As part of this collection of papers and presentations, there is both an assessment of the current state-of-the-art as well as evolving approaches and solutions to such impediments. Ultimately, it is our belief that the benefits of, and momentum behind, open science paradigms will overcome these barriers as a result in solutions that address fundamental flaws associated with "traditional" and highly compartmentalized approaches to translational research. This momentum will be amplified by the way in which open science democratizes access to and participation in research endeavors, thus supporting true communities-of-practice and the economy of ideas and thinking such collaborative constructs provide for. However, the trajectory of open science we envision will not be easy, as it represents a fundamental culture change in the health and life sciences communities, and it is well understood and documented that culture change is challenging and fraught with peril for early adopters or advocates of such change. As such, we believe that this session and its content serve as a critical "marker" concerning the future directions and research agendas needed to realize such a vision for high impact discovery science and translational research.

3. Democratizing Health Data for Translational Research

As was noted above, a selection of papers and presentations concerning the current state-of-the-art in terms of democratizing health data for translational research, as well as evolving approaches to fundamental impediments in implementing open science, was curated as part of the proceeding of PSB 2018. Below is a thematic summary of the three major areas of endeavor highlighted by those papers and presentations:

- **Leveraging data models and standard to improve the discoverability and utility of public data repositories:** Exemplars of this theme included: 1) efforts by ***Madhavan*** and colleagues to employ syntactic and semantic standards in order to improve the accessibility and usefulness of comprehensive biomolecular and clinical data sets created by the ClinGen initiative; 2) methods being developed by ***Tenenbaum*** and colleagues to ensure the reproducibility of analyses using publically available data assets relevant to Alzheimer's research; and 3) similar methodological development efforts by ***Sharma*** and colleagues to extract meaningful health outcomes and natural product therapeutic information from adverse event report systems.
- **Synthesizing and simulating data sets that are comparable to human-derived source data:** Exemplary of this theme is the work by ***Moore*** and colleagues to employ simulation techniques to synthesize health data that can be used to inform the design and evaluation of a variety of machine learning algorithms that can identify potentially informative patterns spanning multidimensional data.
- **Educating informatics investigators to systematically and responsibly access and utilize emerging sources of health data for translational research:** Finally, this theme is represented by the work of ***Van Horn*** and colleagues to define pedagogical approaches for

preparing biomedical informatics and data science investigators to responsibly and reproducibly utilize health data as found in a variety of domains in order to support hypothesis generating and testing science.

4. Conclusions

PSB 2018 is a unique venue for thought leadership and technical direction setting that we believe can enable widespread action to "democratize" health data and to support timely as well as high impact translational research. This capability is exemplified by the work presented in our session. Ultimately, the themes and findings presented by our authors chart a path forward for this important area, involving:

- The **creation, verification and validation of tools** and methods that can assist in the sharing, discovery, and analysis of open health data in a primary or secondary manner, including the development of databases, algorithms, and modeling techniques therein;
- The **conduct of discovery science in data-intensive experimental contexts** that leverage such open health data resources across scales from molecules to patient to populations; and
- The **preparation and interaction of multidisciplinary computational, biology, clinical, and population health science teams** to conduct research that serves to identify policy, technical, and socio-cultural needs associated with the implementation of open science paradigms that include patient-derived health data.

This research agenda can and should advance our collective understanding of the role of "democratized" health data in advancing the state-of-the-art in translational research writ large with demonstrable benefit in terms of human health and wellness.

Acknowledgements

The authors wish to acknowledge the contributions of all of the authors who submitted content to this session at PSB 2018, as well as the scientific and editorial oversight provided by the conference's scientific program committee.

References

1. Goodman, S.N., D. Fanelli, and J.P. Ioannidis, *What does research reproducibility mean?* Science translational medicine, 2016. **8**(341): p. 341ps12-341ps12.
2. Iqbal, S.A., et al., *Reproducible research practices and transparency across the biomedical literature.* PLoS Biol, 2016. **14**(1): p. e1002333.
3. Nosek, B.A., et al., *Promoting an open research culture.* Science, 2015. **348**(6242): p. 1422-1425.
4. Warren, E., *Strengthening research through data sharing.* New England Journal of Medicine, 2016. **375**(5): p. 401-403.
5. Holve, E., *Open Science and eGEMs: Our Role in Supporting a Culture of Collaboration in Learning Health Systems.* eGEMs, 2016. **4**(1).
6. McKiernan, E.C., et al., *How open science helps researchers succeed.* Elife, 2016. **5**: p. e16800.
7. Moher, D., et al., *Increasing value and reducing waste in biomedical research: who's listening?* The Lancet, 2016. **387**(10027): p. 1573-1586.
8. FOSTER. *Open Science Taxonomy*. 2016 [cited 2016 September 14]; Available from: https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition.
9. Longo, D.L. and J.M. Drazen, *Data sharing.* New England Journal of Medicine, 2016. **374**(3): p. 276-277.
10. Aitken, M., et al., *Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies.* BMC medical ethics, 2016. **17**(1): p. 73.
11. Ross, J.S. and H.M. Krumholz, *Open Access Platforms for Sharing Clinical Trial Data.* Jama, 2016. **316**(6): p. 666-666.
12. Taichman, D.B., et al., *Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors.* JAMA, 2016. **315**(5): p. 467-468.
13. Vallance, P., A. Freeman, and M. Stewart, *Data Sharing as Part of the Normal Scientific Process: A View from the Pharmaceutical Industry.* PLoS Med, 2016. **13**(1): p. e1001936.
14. Wilbanks, J. and S.H. Friend, *First, design for data sharing.* Nature biotechnology, 2016.
15. Denny, J.C., et al., *Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data.* Nature biotechnology, 2013. **31**(12): p. 1102-1111.
16. Payne, P.R. and P.J. Embi, *An Introduction to Translational Informatics and the Future of Knowledge-Driven Healthcare*, in *Translational Informatics*. 2015, Springer. p. 3-19.

17. Plenge, R.M., E.M. Scolnick, and D. Altshuler, *Validating therapeutic targets through human genetics.* Nature reviews Drug discovery, 2013. **12**(8): p. 581-594.
18. Ritchie, M.D., et al., *Methods of integrating data to uncover genotype-phenotype interactions.* Nature Reviews Genetics, 2015. **16**(2): p. 85-97.
19. Shah, N.H., *Mining the ultimate phenome repository.* Nature biotechnology, 2013. **31**(12): p. 1095-1097.
20. Chen, B. and A.J. Butte, *Leveraging big data to transform target selection and drug discovery.* Clinical Pharmacology & Therapeutics, 2016. **99**(3): p. 285-297.
21. Li, J., et al., *A survey of current trends in computational drug repositioning.* Briefings in bioinformatics, 2016. **17**(1): p. 2-12.
22. Liu, Z., et al., *In silico drug repositioning–what we need to know.* Drug discovery today, 2013. **18**(3): p. 110-115.
23. Joly, Y., et al., *Are Data Sharing and Privacy Protection Mutually Exclusive?* Cell, 2016. **167**(5): p. 1150-1154.
24. Krumholz, H.M., S.F. Terry, and J. Waldstreicher, *Data acquisition, curation, and use for a continuously learning health system.* Jama, 2016. **316**(16): p. 1669-1670.
25. Krumholz, H.M. and J. Waldstreicher, *The Yale Open Data Access (YODA) project—a mechanism for data sharing.* New England Journal of Medicine, 2016. **375**(5): p. 403-405.
26. Ashley, E.A., *Towards precision medicine.* Nature Reviews Genetics, 2016. **17**(9): p. 507-522.
27. *OHDSI*. 2017; Available from: http://www.ohdsi.org/.
28. Frasier, M., *Perspective: Data sharing for discovery.* Nature, 2016. **538**(7626): p. S4-S4.
29. Payne, P., et al., *Enabling Open Science for Health Research: Collaborative Informatics Environment for Learning on Health Outcomes (CIELO).* Journal of Medical Internet Research, 2017. **19**(7).
30. Watson, M., *When will 'open science' become simply 'science'?* Genome biology, 2015. **16**(1): p. 101.