Pacific Symposium on Biocomputing 26:341-345 (2021)

Establishing the reliability of algorithms

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

Sean D. Mooney
Department of Biomedical Informatics and Medical Education, University of Washington
Seattle, WA, USA
Email: sdmooney@uw.edu

Iddo Friedberg
Bioinformatics and Computational Biology Program, Department of Veterinary Microbiology and Preventive Medicine, Iowa State University
Ames, IA, USA

Justin Guinney
Sage Bionetworks
Seattle, WA, USA
Email: justin.guinney@sagebionetworks.org
idoerg@gmail.com

As rich biomedical data streams are accumulating across people and time, they provide a powerful opportunity to address limitations in our existing scientific knowledge and to overcome operational challenges in healthcare and life sciences. Yet the relative weighting of insights *vs.* methodologies in our current research ecosystem tends to skew the computational community away from algorithm evaluation and operationalization, resulting in a well-reported trend towards the proliferation of scientific outcomes of unknown reliability. Algorithm selection and use is hindered by several problems that persist across our field. One is the impact of the self-assessment bias, which can lead to mis-representations in the accuracy of research results. A second challenge is the impact of data context on algorithm performance. Biology and medicine are dynamic and heterogeneous. Data is collected under varying conditions. For algorithms, this means that performance is not universal -- and need to be evaluated across a range of contexts. These issues are increasingly difficult as algorithms are trained and used on data collected in the real-world, outside of the traditional clinical research lab. In these cases, data collection is

neither supervised nor well controlled and data access may be limited by privacy or proprietary reasons. Therefore, there is a risk that algorithms will be applied to data that are outside of the scope of the intent of the original training data provided. This workshop will focus on approaches that are emerging across the researcher community to quantify the accuracy of algorithms and the reliability of their outputs.

Keywords: benchmarking; algorithm assessment; open science; translational research

1. Introduction

Despite intensive efforts to utilize this data to optimize healthcare, relatively few methods have been adequately validated and clinically deployed. The reasons for this are technical, scientific, social and business related. On the technical side this includes inaccessibility of gold-standard datasets for robust validation, heterogeneity in data collected from distributed sources, contextual relevance of biological observations across samples, poor algorithmic reproducibility and community-acceptance of biased approaches for assessing methods. Reproducibility and transparency are two methods which support development of reliable biomedical claims that can both generate new knowledge and apply it to advance health care. Although these approaches have become firmly established and increasingly practiced over the past decade, they do not fully address the question of transferability in biomedical research findings or algorithms. This topic builds from the types of work described in the PSB 2017 Session on Methods to Ensure Reproducibility in Biomedical Research, which was developed in reaction to both the announcement of the data sharing initiatives of the Biden Cancer program and the NEJM data parasite commentary, focused on methods that individual researchers were taking to assure reproducibility within their own work. This session will discuss general methods for open community-based methods to benchmark algorithms, including the use of crowd-sourced challenges[1-3] as a tool for the unbiased assessment of tools and algorithms.

The public health, economic, and social justice crises that have occurred in 2020 have brought an urgency to the question of rapid, reliable algorithm assessments. The global COVID-19 pandemic has provided an urgent need to rapidly optimize healthcare practices, establish public health practices for prevention and monitoring, and identify drugs and vaccines to use in prevention and treatment. The urgency of this situation is at odds with the typical pace through which scientific knowledge is developed, established and integrated into care. Further, the social justice crisis underlies the known issues with medical algorithms that initiate biases or may propagate those established in the underlying data.

2. Workshop goals and organization

This workshop, **Establishing Reliability in Algorithms**, at the 2021 Pacific Symposium on Biocomputing is designed to stimulate conversation around mechanisms that our community can use to objectively establish the reliability of algorithms. This will include community mechanisms for evaluation as well as mechanisms for use by individual researchers within the context of independent research programs. The workshop will provide three examples of existing approaches and then stimulate an open discussion that will be actively guided and moderated by the organizers. The conversation should extend to a discussion of potential mechanisms for establishing standards that enforce greater accountability across the community.

Topics to be covered in the presented materials will include:

Predictive analytics in healthcare: The COVID-19 pandemic has highlighted an urgent need for healthcare systems to learn from and with each other. Clinical analytics teams are implementing predictive analytics methods that use algorithms trained on electronic health record (EHR) and other data to improve patient care and lower costs. While these methods have the promise of being impactful in delivering on precision medicine and managing population health, their real world accuracy over time is not well understood[4-7]. It is the case in most areas of biomedicine that the evaluation of methods across multiple data sets should be transparent and used to establish their replicability and reliability. Due to differences across clinical sites in practice, population, and data capture, the question of reliability may be less evident and requires an understanding of the context - and potential impact - of deploying an algorithm within a particular system.

Regulatory Science: Another area where analytical methods are directly impacting health care is in support of regulatory filings for new drugs and devices. Data derived from both EHR and from remote monitoring devices are increasingly utilized in this capacity. Recognizing the need to objectively assess the accuracy of methodologies used in the development of regulatory filings, the FDA introduced PrecisionFDA[8], an objective benchmarking program in 2015, which has built from an original focus on genomic processing methods. The proprietary nature of this work introduces barriers to data collection or sharing that make traditional approaches to algorithm assessment unsatisfying. Approaches that can support objective evaluation of results arising from closed data sources are required. Acknowledgement of these needs are represented from the FDA by their Spring 2020 solicitation for community input towards the modernization of their data strategy[9].

Molecular Modeling and Analytics: Biomedical researchers are routinely generating genomic, proteomic, epigenomic, imaging, and other emerging molecular data types comprising billions of data-points. Community benchmarking approaches such as the DREAM Challenges or the

Critical Assessment experiments have predominantly focused in this domain[10-14], where fewer commercial interests impact the sharing of data or knowledge. An evaluation of benchmarking practices within this domain can help to identify lessons from the successes, current gaps in practice, and the development of sustained standards for community-based algorithm assessment.

A moderated discussion will follow that will cover the following topics:

- successes and lessons learned to-date from community benchmarking practices - indicating what impact these approaches to establish reliable outcomes have had on subsequent research or translation practices
- early lessons learned from healthcare method implementation
- emerging approaches in data sharing and in algorithm development and assessment that are addressing the issue of appropriate algorithm interpretation
- community needs and potential solutions for addressing algorithm reliability
- Development of better gold standards in biomedicine and approaches to overcome sub-optimal gold standards

3. References:

1. Ellrott, K. et al. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. Genome Biol. 20, 195 (2019).
2. Bender, E. Challenges: Crowdsourced solutions. Nature 533, S62–4 (2016).
3. Saez-Rodriguez, J. et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. Nat. Rev. Genet. 17, 470–486 (2016).
4. EHR DREAM Challenge. synapse.org/ehr_dream_challenge_mortality doi:10.7303/syn18405991.
5. Kahn, M. G. et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC) 4, 1244 (2016).
6. Beaulieu-Jones, B. K. et al. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. Circ. Cardiovasc. Qual. Outcomes 12, e005122 (2019).
7. Chen, J., Chun, D., Patel, M., Chiang, E. & James, J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC Med. Inform. Decis. Mak. 19, 44 (2019).
8. https://precision.fda.gov/
9. Modernizing the Food and Drug Administration's Data Strategy; Public Meeting; Request for Comments. https://www.federalregister.gov/documents/2020/01/08/2020-00071/modernizing-the-food-and-drug-administrations-data-strategy-public-meeting-request-for-comments

10. Marbach D. et. Al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012 Aug; 9(8): 796–804.
11. Trister, A. D., Buist, D. S. M. & Lee, C. I. Will Machine Learning Tip the Balance in Breast Cancer Screening? JAMA Oncol (2017) doi:10.1001/jamaoncol.2017.0473.
12. Keller, A. et al. Predicting human olfactory perception from chemical features of odor molecules. Science 355, 820–826 (2017).
13. Radivojac, P., Clark, W., Oron, T. *et al*. A large-scale evaluation of computational protein function prediction. Nat Methods 10, 221–227 (2013).
14. Zhou N, Siegel ZD, Zarecor S, Lee N, Campbell DA, et al. (2018) Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning. PLOS Computational Biology 14(7): e1006337