The model-to-data paradigm: overcoming data access barriers in biomedical competitions

Justin Guinney
Dept. of Computational Oncology
Sage Bionetworks
Seattle, WA 98121
justin.guinney@sagebionetworks.org

Abstract

Data competitions often rely on the physical distribution of data to challenge participants, a significant limitation given that much data is proprietary, sensitive, and often non-shareable. To address this, the DREAM Challenges has advanced a challenge framework called *model-to-data* (MTD), requiring participants to submit re-runnable algorithms instead of model predictions. The DREAM organization has successfully completed multiple MTD-based challenges, and is expanding this approach to unlock highly sensitive and non-distributable human data for use in biomedical data challenges.

1 Introduction

A common hurdle in the organization of data challenges is the acquisition of datasets that can be used for model training and validation. This problem is amplified in biomedical data challenges due to the privacy and security concerns associated with human data. Even with HIPAA identifiers scrubbed from data, concerns of reidentifiability remain, along with issues of intellectual property and commercial risks. Consequently, large quantities of biomedical data remain hidden behind firewalls, out of reach from the research community and barred from use within data challenges. At a time when AI in healthcare is touted as critical to improving healthcare inefficiencies and patient outcomes, biomedical data challenges - and clinical benchmarking initiatives more broadly - can serve an important role in the development and assessment of clinical AI. To achieve this, alternative paradigms of data access are needed to reduce the financial, legal and technical costs of data utilization, and to unlock critical datasets for method development and assessment using sensitive biomedical data.

2 The Model-to-Data Paradigm

The Model to Data (MTD) framework for enabling research on private data was described by Guinney et al. as an alternative to traditional data sharing methods [1]. The focus of MTD is to enable the development of analytic tools and predictive models without needing to provide a researcher direct access to the data. Instead of having data distributed directly to the researcher, a researcher will send a containerized model to the data owners - or a third-party honest broker - who are then responsible for applying the model on secure and protected data. The outputs of the algorithm, along with performance metrics, are then returned back to the researcher. MTD bypasses the need for complex data-use agreements and other costly legal protections that accompany data transfers. While there are clear drawbacks to this framework, as it does not allow unconstrained data exploration and modeling, the MTD framework provides an important middle-ground that allows access and utilization of data that would otherwise remain hidden and inaccessible to the reserach community.

Technically, the MTD framework relies on modern containerization software such as Docker or Singularity for model packaging and transfer. These technologies have greatly simplified the problem of algorithm portability

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

and reproducibility in heterogeneous compute environments. MTD is also enabled by modern cloud platforms (e.g. Amazon Web Services, Google Cloud), which have commoditized data storage and compute, and provide almost limitless scaling of compute resources along with scaleable costs.

3 Demonstrations of the MTD Framework in Biomedical Data Challenges

MTD has been successfully implemented and demonstrated in a series of community challenges [2], including a challenge to predict patient outcomes in multiple myeloma using genomic and clinical data, and a proteomic challenge to predict protein abundance from transcriptomic data. The most ambitious applications of MTD include the completed Digital Mammography DREAM Challenge, and the recently launched Electronic Healthcare Challenge, described here in more detail.

3.1 Digital Mammography (DM) DREAM Challenge

The objective of the DM Challenge was to determine whether AI could overcome human mammography interpretation limitations with a rigorous, unbiased evaluation of machine learning algorithms[3]. Data was contributed by two institutions - Kaiser Permanente and the Karolinska Institute - with an aggregate of over 160k woman exams and 1.1 million DM images ($\tilde{1}$5TB). The data contributors required that data could not be directly distributed to and accessed by challenge participants. To overcome this limitation, the MTD framework was employed: participants were asked to submit containerized algorithms for application on these hidden datasets. The Kaiser cohort was split into datasets for model tuning and model validation, and the Karolinska cohort was used exclusively for independent model validation.

Over 1,100 participants comprising 126 teams from 44 countries participated in the challenge. While no single AI algorithm outperformed radiologists, an ensemble of AI algorithms combined with radiologist assessment significantly outperformed the radiologist alone, demonstrating how deep learning approaches could augment the radiologist interpretation. This study also demonstrated the feasibility and scaleability of the MTD approach in the context of highly sensitive, clinical datasets.

3.2 Electronic Healthcare Record (EHR) DREAM Challenge

The EHR DREAM Challenges represents a series of data challenges to address important clinical questions using patient healthcare data. In the first recently launched challenge, participants are asked to predict patient mortality within six months of their last hospital visit. The data host for this Challenge is the University of Washington Medical System, which has prepared a curated dataset from their EHR enterprise data warehouse. The data collected span 10 years (2009-2019), with 1.3 million patients, 22 million visits, 33 million procedures, 5 million drug exposure records, 48 million condition records, 10 million observations, and 221 million measurements. Data is formatted using the OMOP common data model. Given the highly sensitive nature of EHR data and associated risks of re-identifiability, the MTD approach is being used to enable the use of this data.

Prior to launching the EHR Challenge, we completed a feasibility study to assess both the technical infrastructure and to establish baseline patient mortality models using the EHR data. We developed 3 models using the same infrastructure available to challenge participants: model 1 used only demographic information, model 2 used only demographic information and 4 common chronic diseases, and model 3 used only demographic information and the top 20 indications. The performance of these models - measured using area under the receiver-operating-curve - was 0.682, 0.794, and 0.723, respectively. This successfully demonstrated the technical robustness of the challenge architecture, and the ability to generate and evaluate predictive algorithms in a secure manner. The EHR Challenge was launched in September 2019, and will close in January 2020.

4 Conclusion

The MTD framework is a powerful framework for using sensitive and difficult-to-share data in competitions, and has been successfully demonstrated in multiple challenges including the Digital Mammography and EHR DREAM Challenges. In the future, we intend to expand MTD within the biopharma and healthcare industries to unlock important biomedical data for the development and benchmarking of clinical algorithms.

References

[1] Guinney J, & Saez-Rodriguez J. (2018) Alternative models for sharing confidential biomedical data. *Nat Biotechnol* **36**(5):391–2.

[2] Ellrott K Buchanan A Creason A, et al. (2019) Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. *Genome Biol* **20**(1):195

[3] Trister AD Buist DSM Lee CI. (2017) Will Machine Learning Tip the Balance in Breast Cancer Screening? *JAMA Oncol*