

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

Thomas Schaffter^{1,*}, Timothy Bergquist^{2,*}, Yao Yan³, Thomas Yu¹, Vikas Pejaver², Noah Hammarlund², Justin Prosser⁴, Sean Mooney², Justin Guinney^{1,2}

Model-to-data paradigm enables researchers to train and evaluate models on critical data derived from health care and clinical trials

Alternative models for sharing confidential biomedical data

Critical patient information derived from academic research, health care and clinical trials are off limit for traditional data-to-model challenges. Existing barriers include:

- Access to big and sensitive data
- Lack of effective frameworks for assessing performance & generalizability

Challenge participants (public)

Training data Validation data

Model Prediction

Challenge teams

Challenge cloud platform (private)

Predictions Ground truth

Scoring

Leaderboards & benchmarks

Traditional data-to-model challenges

Our solution: The model-to-data paradigm

Containerized models submitted by participants are trained and/or evaluated on sensitive data in secure environments. The reproducibility of the results is also easier to achieve.

Challenge participants (public)

Training data

Model Containerized model

Challenge teams

Challenge cloud platform (private)

Model submissions Validation data sets

Future models Prospective data sets

Scoring

Leaderboards & benchmarks

EHR DREAM Challenge: Patient Mortality Prediction

Our vision is to apply the MTD paradigm to build a **network of EHR Data Nodes** to enable participants to develop predictive models.

For this first EHR Challenge, participants are addressing the following question:

Given all the past EH Records of a patient, predict the probability that the patient will pass away within 180 days following his/her most recent exam.

Challenge data

- Synthetic data (SynPUFF)
- University of WA EHR data warehouse
 - 1.3 million patients with at least one visit
 - 22 million visits
 - 5 million drug exposure
 - 10 million observations
 - 48 million conditions
 - 221 million measurements
- Format: OMOP common data model

Submission workflow

Participation & preliminary results

- 320 registered participants
- 282 models submitted from 29 teams

Rank	Performer	AUROC	AUPR
1	ivanbrugere	0.953	0.255
2	Georgetown - ESAC	0.930	0.124
3	LCSB_LUX	0.927	0.116
4	avati	0.926	0.153
5	DMIS_EHR	0.926	0.148

Evolution of the AUROC of submitted models

Model-To-Data DREAM Challenges

The MTD approach has already been implemented in several DREAM Challenges such as **The Digital Mammograph Challenge**.

Clinical data (longitudinal)

Digital mammograms (longitudinal)

Dockerized model

Probability to develop cancer within 6 months after the most recent exam

- Participants do not have direct access to patient data at any point.

Contact

SageBionetworks

DREAM CHALLENGES

CD2H NATIONAL CENTER FOR DATA TO HEALTH

UW Medicine BIOMEDICAL INFORMATICS AND MEDICAL EDUCATION

^{*} Are contributing equally

¹ Sage Bionetwork, Seattle, WA 98121, USA

² Biomedical Informatics and Medical Education, University of Washington

³ Molecular Engineering & Sciences Institute, University of Washington

⁴ Institute for Translational Health Sciences, University of Washington

justin.guinney@sagebionetworks.org

dreamchallenges.org

synapse.org/ehr_dream_challenge_mortality

Guinney, J., Saez-Rodriguez, J. Alternative models for sharing confidential biomedical data. Nat Biotechnol 36, 391–392 (2018), doi:10.1038/nbt.4128

Ellrott, K., Buchanan, A., Creason, A., Mason, M., Schaffter, T., Hoff, B., ... & Saez-Rodriguez, J. (2019). Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. Genome biology, 20(1), 1-9. doi.org/10.1186/s13059-019-1794-0

figshare

(this poster!)