

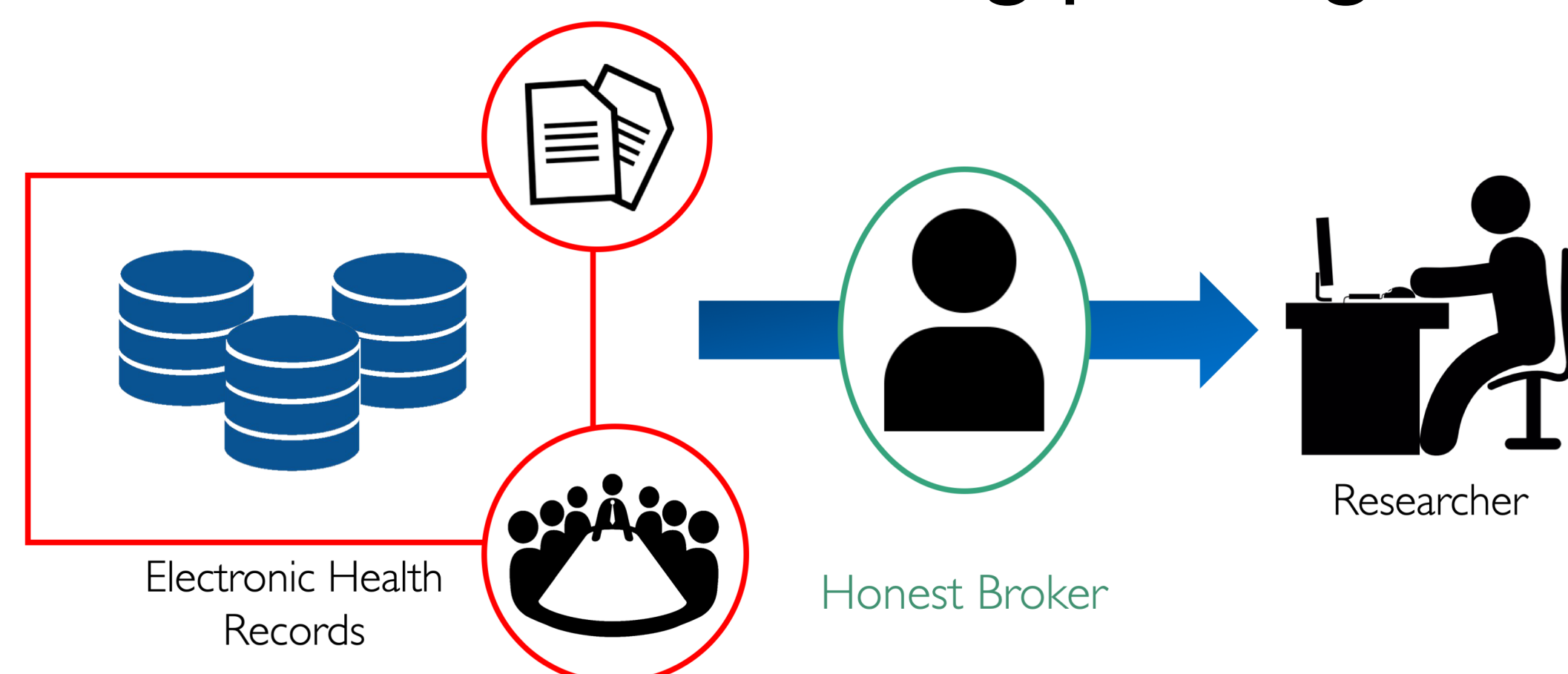
A framework for studying machine learning methods in healthcare: The First EHR DREAM Challenge

Timothy Bergquist^{1,2,*}, Thomas Schaffter^{1,*}, Yao Yan^{1,3}, Thomas Yu¹, Vikas Pejaver¹, Noah Hammarlund², Justin Prosser⁴, Sean D. Mooney^{2,4}, Justin Guinney^{1,2}

The Model-to-data paradigm enables researchers to train and test models on sensitive health data and allows model assessors to evaluate model bias and accuracy.

An alternative model for sharing sensitive biomedical data

Traditional data sharing paradigm



Due to its highly sensitive nature, electronic health record data (EHR) are protected behind firewalls, Institutional Review Boards, privacy regulations, and Data Use Agreements. These make using EHRs for building novel predictive models difficult. In addition, when researchers are able to gain access to these data, often through an honest broker who de-identifies the data, re-identifying patients is still a risk even in these de-identified datasets.

Our Solution: The model-to-data paradigm

1. Researchers build models on local environments with synthetic data.
2. Models are containerized (Docker) and submitted to the Synapse collaboration platform.
3. Models are pulled into a cloud environment and validated against synthetic data.
4. Validated models are pulled into the UW secure environment to be trained and evaluated against real clinical records.
5. Predictions are assessed against goldstandard clinical outcomes.
6. Model accuracy metrics are returned to researchers.

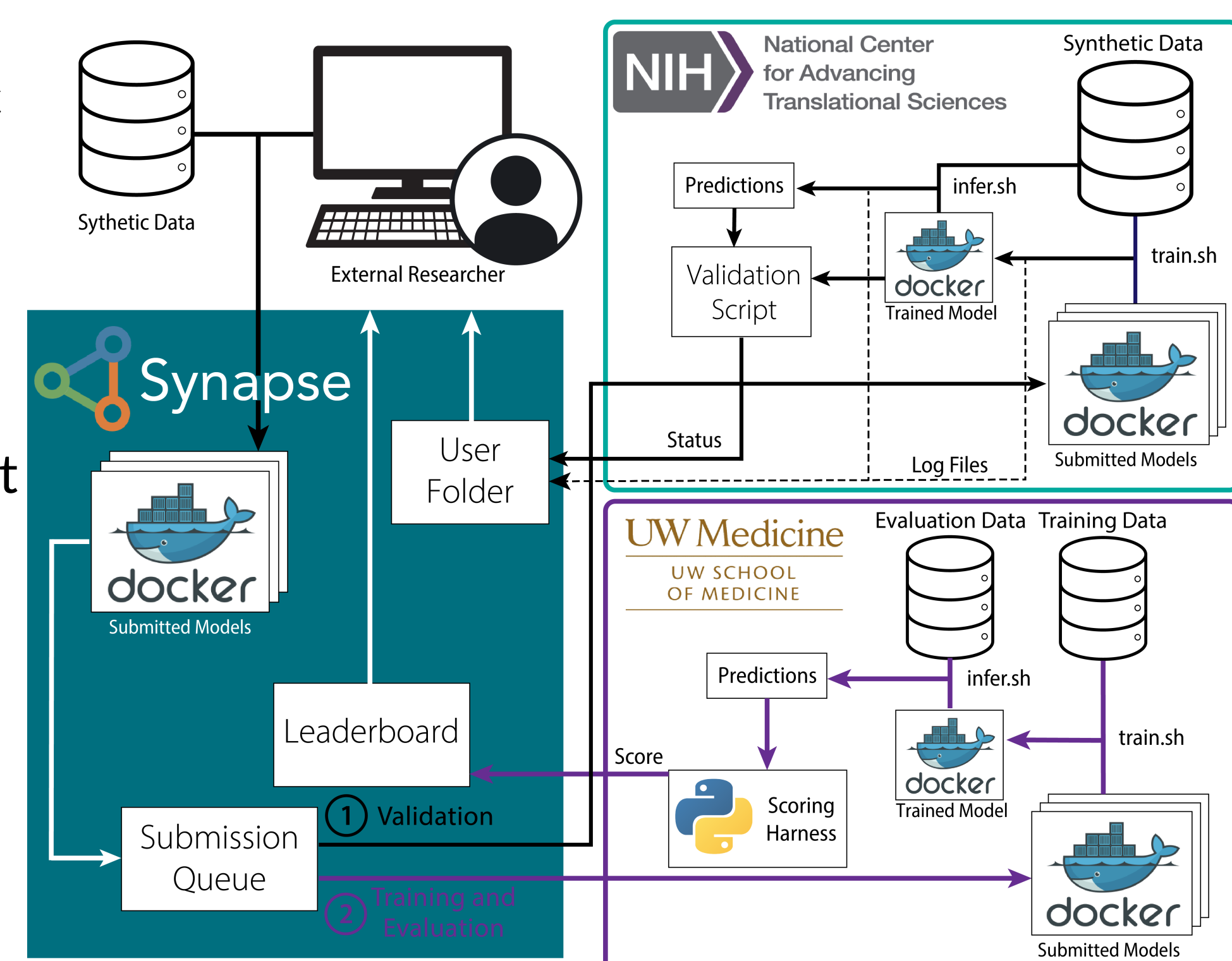


Figure 2. Our implementation of the model-to-data workflow combining the Synapse collaboration platform and the University of Washington Clinical Data Warehouse.

We wanted to test this novel Model-to-Data infrastructure with a community challenge.

EHR DREAM Challenge: Patient Mortality Prediction

Challenge Question

For this EHR DREAM Challenge, we asked participants to address the following question:

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

We made the records of 1.3 million patients available within the secure UW environment for training and testing containerized models

Challenge Results

- 350 registered participants
- Over 30 teams trained and tested models.

Conclusion

The Model to Data approach can be used to develop accurate prediction models on a hidden EHR dataset. This framework leads to more reproducible models and give assessors the ability to evaluate models across different subpopulations.

The highest scoring team (University of Wisconsin biostatistics) achieved a performance of 0.947 AUROC and 0.478 AUPR

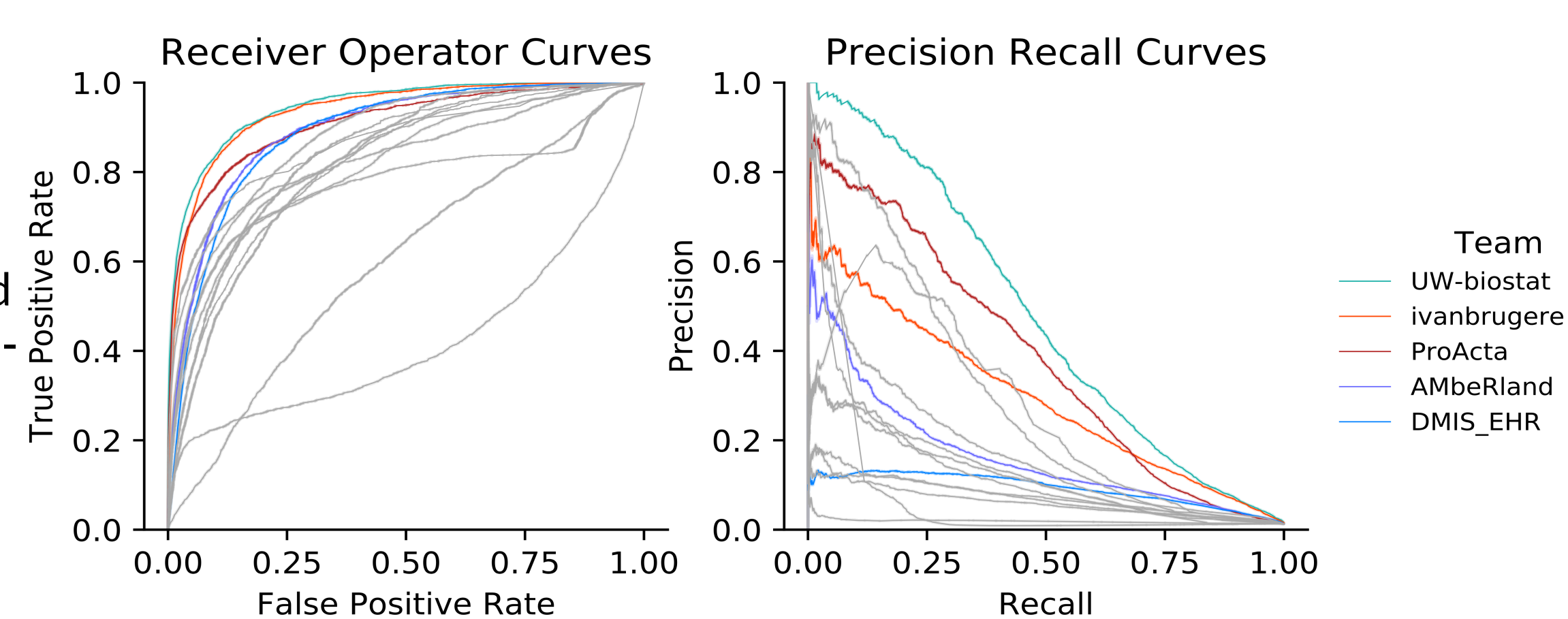


Figure 3. Final results after validating submitted models on a hold out validation set. The top 5 models are colored.

Models show varying accuracies across racial groups

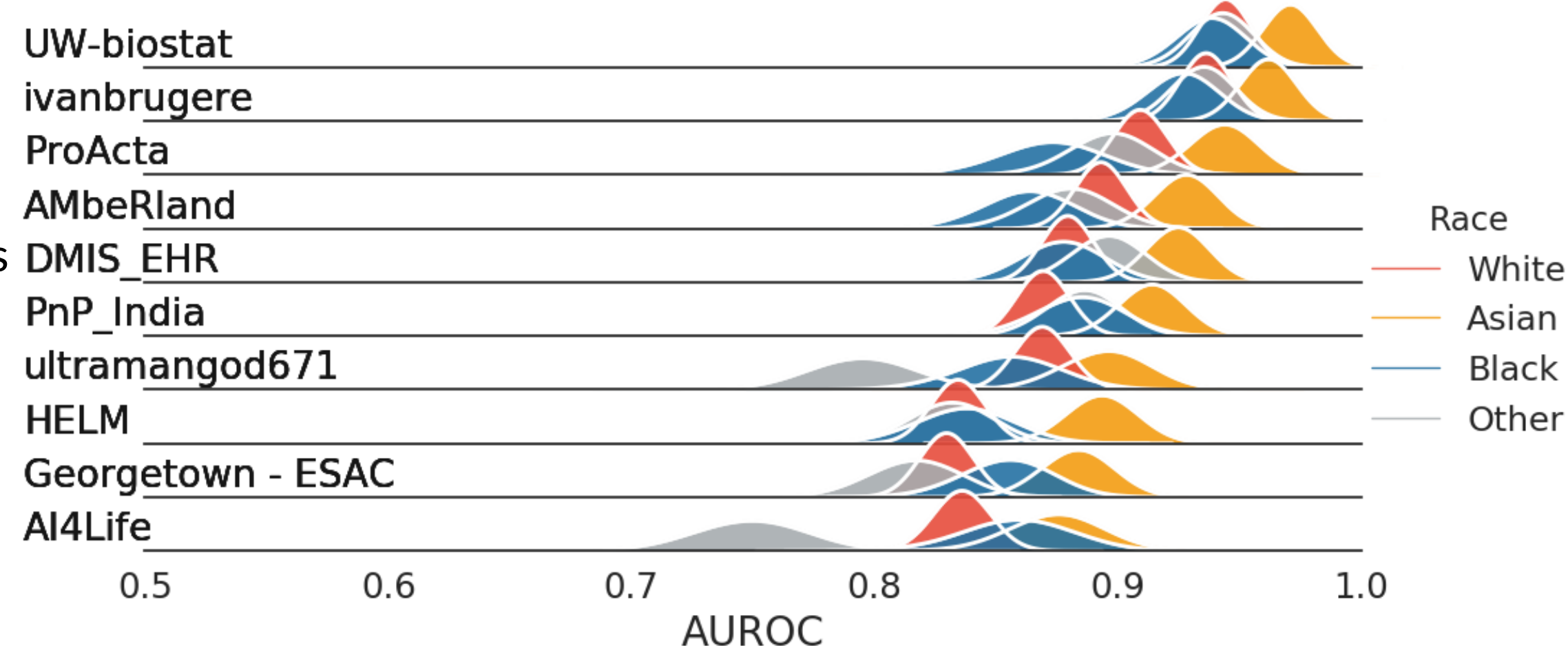


Figure 4. Bootstrapped (n=10,000) distributions of model AUROCs across difference racial groups for the top 10 teams.

Contact  SageBionetworks  CD2H NATIONAL CENTER FOR DATA TO HEALTH

Corresponding Author: justin.guinney@sagebase.org
Corresponding Author: sdmooney@uw.edu

* 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

UW Medicine
BIOMEDICAL INFORMATICS
AND MEDICAL EDUCATION

DREAM
CHALLENGES

ITHS
Institute of Translational
Health Sciences

dreamchallenges.org

synapse.org/ehr_dream_challenge_mortality

Guinney, J., & Saez-Rodriguez, J. (2018). Alternative models for sharing confidential biomedical data. *Nature Biotechnology*, 36(5), 391–392. <https://doi.org/10.1038/nbt.4128>

Bergquist, T., Yan, Y., Schaffter, T., Yu, T., Pejaver, V., Hammarlund, N., Prosser, J., Guinney, J., & Mooney, S. (2020). Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction. *Journal of the American Medical Informatics Association: JAMIA*. <https://doi.org/10.1093/jamia/ocaa083>