






Research and Applications

Evaluation of crowdsourced mortality prediction models as a framework for assessing artificial intelligence in medicine

Timothy Bergquist , PhD^{1,2}, Thomas Schaffter, PhD¹, Yao Yan, PhD^{1,3}, Thomas Yu, BA¹, Justin Prosser⁴, Jifan Gao, MS⁵, Guanhua Chen, PhD⁵, Łukasz, Charzewski, MScEng^{6,7}, Zofia Nawalany, MSc⁶, Ivan Brugere, PhD⁸, Renata Retkute, PhD⁹, Alisa Prusokiene, PhD¹⁰, Augustinas Prusokas¹¹, Yonghwa Choi¹², Sanghoon Lee¹², Junseok Choe¹², Inggeol Lee, MD¹³, Sunkyu Kim , PhD¹², Jaewoo Kang , PhD^{12,13}, Sean D. Mooney , PhD^{2,*}, Justin Guinney , PhD^{1,2,*}; and the Patient Mortality Prediction DREAM Challenge Consortium**

¹Sage Bionetworks, Seattle, WA, United States, ²Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States, ³Molecular Engineering and Sciences Institute, University of Washington, Seattle, WA, United States, ⁴Institute of Translational Health Sciences, University of Washington, Seattle, WA, United States, ⁵Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, United States, ⁶Proacta, Warsaw, Poland, ⁷Division of Biophysics, University of Warsaw, Warsaw, Poland, ⁸Department of Computer Science, University of Illinois at Chicago, Chicago, IL, United States, ⁹Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom, ¹⁰Plant and Molecular Sciences, School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom, ¹¹Department of Life Sciences, Imperial College London, London, United Kingdom, ¹²Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul, Republic of Korea, and ¹³Department of Interdisciplinary Program in Bioinformatics, College of Informatics, Korea University, Seoul, Republic of Korea

*Corresponding authors: Justin Guinney, PhD, Sage Bionetworks, 2901, 3rd Ave, #330, Seattle, WA 98121 (justin.guinney@tempus.com); Sean D. Mooney, PhD, University of Washington, 850 Republican St, Seattle WA 98109 (sdmooney@uw.edu)

Author contributions: T. Bergquist and T. Schaffter contributed equally and are considered co-first authors of this work.

**A list of authors and their affiliations appears at the end of the article.

ABSTRACT

Objective: Applications of machine learning in healthcare are of high interest and have the potential to improve patient care. Yet, the real-world accuracy of these models in clinical practice and on different patient subpopulations remains unclear. To address these important questions, we hosted a community challenge to evaluate methods that predict healthcare outcomes. We focused on the prediction of all-cause mortality as the community challenge question.

Materials and methods: Using a Model-to-Data framework, 345 registered participants, coalescing into 25 independent teams, spread over 3 continents and 10 countries, generated 25 accurate models all trained on a dataset of over 1.1 million patients and evaluated on patients prospectively collected over a 1-year observation of a large health system.

Results: The top performing team achieved a final area under the receiver operator curve of 0.947 (95% CI, 0.942-0.951) and an area under the precision-recall curve of 0.487 (95% CI, 0.458-0.499) on a prospectively collected patient cohort.

Discussion: *Post hoc* analysis after the challenge revealed that models differ in accuracy on subpopulations, delineated by race or gender, even when they are trained on the same data.

Conclusion: This is the largest community challenge focused on the evaluation of state-of-the-art machine learning methods in a healthcare system performed to date, revealing both opportunities and pitfalls of clinical AI.

Key words: evaluation; machine learning; health informatics

Introduction

Applications of machine learning applied to patient data are undergoing wide development and implementation in healthcare.^{1,2} The performance of these methods as they are used in the clinic—and their associated impact on patient and provider outcomes—are not well understood. An important risk in the design and implementation of machine learning

algorithms is the self-assessment bias, where the implementer and evaluator are the same person or team, which can result in overfitting and poor generalization.³ At the same time, health systems and journals are inundated with new methods that overwhelm the ability of healthcare providers to assess effective solutions. Clinical practice and data collection practices change over time, in some cases, rendering EHR data

obsolete in as little as 3-6 months, as it no longer reflects current data distributions.⁴ Clinical data can also contain hidden biases that reflect social and institutional disparities.⁵ Risks of biases in medicine have been well documented, and models built using biased data will propagate these biases into practice through model recommendations.^{5,6} Addressing these issues requires a rigorous, unbiased framework that can evaluate algorithm performance using independent honest brokers, assess generalizability over time and across institutions, and report on performance disparities across subpopulations.

Concurrently, over the past several decades, “big data” open science experiments have been established that leverage a community of data scientists to work together to competitively solve a specific problem where the answer is unknown to the data scientists. Examples include the DREAM Challenges⁷ and the Critical Assessments.⁸⁻¹² Large-scale clinical data has not been broadly utilized in these experiments because of HIPAA, privacy concerns, and business risks. However, recent technological advances can enable open science on clinical data by restricting data scientist access to protected data and instead of traditional data sharing where the data is given to data scientists, the methods are shared and applied to restricted data by the managers of that data.

We have developed an approach, called Model-to-Data (MTD), that delivers analytical models to protected data without sharing the data directly with model developers.¹³ Instead, developers receive synthetic patient data that conforms to the same data model and architecture as the source protected data and containerized models developed by the participants are delivered to the protected health data for training and evaluation. We previously piloted this method on an EHR dataset from the University of Washington and demonstrated the feasibility of accurate model development without the model developer having direct access to the patient data.¹⁴ This approach has 2 benefits: (1) it protects patient data while allowing researchers to build machine learning methods and (2) it forces a more standardized and transferable approach to building models allowing the data host to perform rigorous evaluations of submitted models.

We leveraged this platform to implement the *EHR DREAM Challenge: Patient Mortality Prediction* to assess machine learning approaches applied to a clinical data warehouse while protecting patient privacy. DREAM Challenges are open competitions, where the challenge organizers solicit the broader research community to develop methods to answer a specific set of biomedical questions,⁷ and to assess these methods using hidden, gold standard datasets. Community challenges have proven to be a robust setting for the objective evaluation of prediction models since they remove the researcher from the evaluation process,^{8,9,15-17} limiting the self-assessment bias.³ We focused on the clinical question of predicting all-cause mortality, as the clinical phenotype is clearly defined and complete (University of Washington merges patient records with state death records to minimize missingness) and previous mortality prediction methods have been developed.¹⁸⁻²¹ In this Challenge, we asked participants to predict whether patients would pass away within 180 days of their last visit to the UW medical system based on that patient’s previous medical history. We evaluated models for population level

accuracy and longitudinal generalizability by evaluating models on a prospectively collected data set. We also assessed demographic generalizability by evaluating model performance across sensitive demographic strata.

Methods

The University of Washington clinical data repository

The UW Medicine enterprise data warehouse (EDW) includes patient records from clinical sites within the UW Medicine system, including more than 300 specialty and primary care clinics. The EDW gathers data from more than 60 sources, including laboratory results, demographic data, diagnosis codes, and medications prescribed. Patient records from 2010 to 2019 in the EDW were transformed into a standardized data format, the Observational Medical Outcomes Partnerships Common Data Model (OMOP CDM v5.0).²² For the EHR DREAM Challenge, we used all patients who had at least 1 visit in the UW OMOP repository, which represented 1.3 million patients with 22 million visits covering approximately 10 years of patient histories.

The EHR DREAM challenge: patient mortality prediction

Challenge question

For this challenge, we asked participants to predict 180-day all-cause mortality from the last patient visit at UW Medicine. True positives were defined as patients who had a death record in the first 180 days of their last visit record and true negatives were defined as patients who either had a death record more than 180 days from their last visit, or who did not have a death record. Death records were derived from the UW medical record and the Washington State death records which were mapped to a patients’ EHR record using their name, address, date of birth, and social security number.

The challenge infrastructure

The EHR DREAM Challenge was developed and run using a “Model to Data” (MTD) approach.^{13,14} This method relies on containerization software (Docker),²³ a common data model (OMOP),²² a model intake mechanism (Synapse),²⁴ and a synthetic dataset for low risk technical validation of submitted models (Synpuf).²⁵ Information on the synthetic data is available in the [Supplementary Material](#). Challenge participants were required to submit “containerized” models to be applied to protected data by the Challenge organizers. At no time during the Challenge did participants have direct access to real patient data and models never had access to direct patient identifiers. Participants were allowed to download the synthetic data to locally test and debug their docker container. Synthetic data was not used to train their models (see [Supplementary Material](#) for synthetic data methods). Participants were allowed to submit pretrained models using data to which they had access, such as their own institution’s clinical data warehouse. The containerized algorithms submitted by participants were able to use a training split of the UW dataset to train a predictive model *de novo*, or to further optimize a pretrained model. Submitted models were first applied to synthetic data to check for technical compliance ([Figure 1](#), Stage 1: Model Validation) and the log files generated by the models in the synthetic data environment were returned to

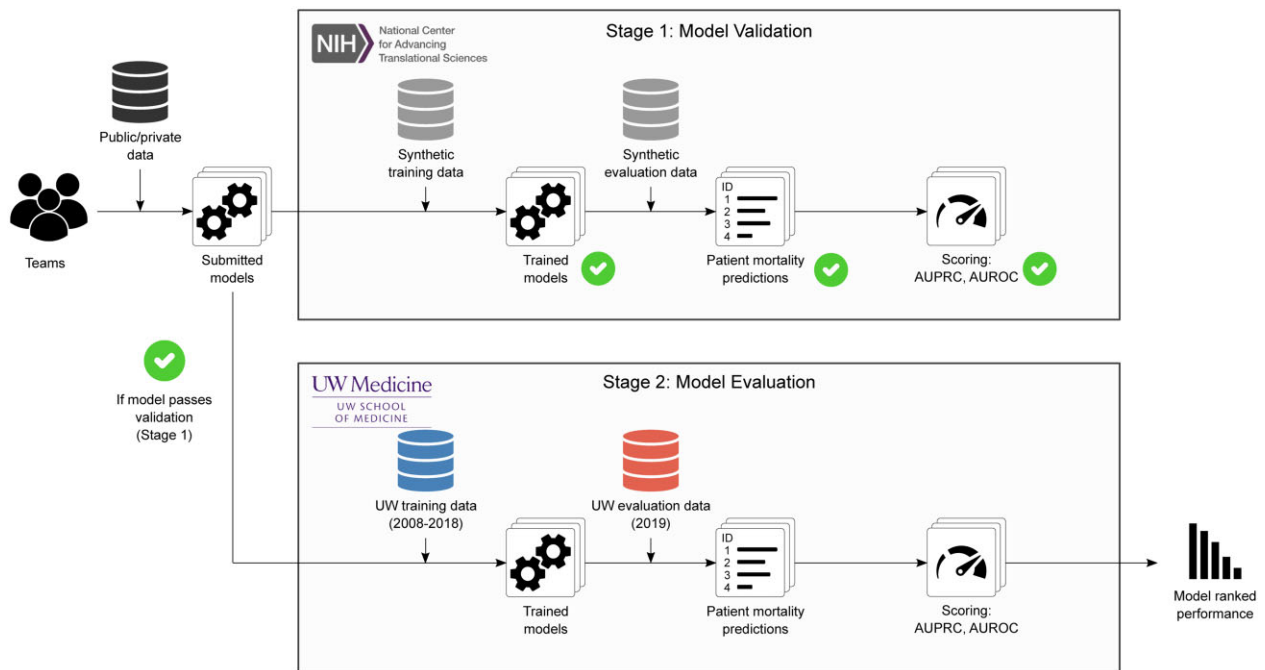


Figure 1. Model-to-data architecture to evaluate the performance of EHR prediction models in the Patient Mortality DREAM Challenge. Models were developed on local environments using synthetic data that resembled the real private EHR data. Docker images were submitted through the Synapse collaboration platform to a submission queue. Images were pulled into the National Center for Advancing Translational Sciences (NCATS) provided AWS cloud environment and run against a synthetic dataset for technical validation (Stage 1). Once validated, images were pulled into the UW Medicine secure infrastructure and run against the private EHR data. Model predictions were evaluated using area under the receiver operator curve (AUROC) and area under the precision recall curve (AUPRC) which were returned to participants through Synapse.

participants. Following successful execution on the synthetic data, the models were pulled into a University of Washington secure environment that was disconnected from the internet where they were trained on the UW OMOP repository. The UW OMOP repository did not contain patient identifiers. The models had no access to the internet during training and evaluation (Figure 1, Stage 2: Model Evaluation). The trained models were tested on a holdout set and the area under the receiver operating curve (AUROC) and area under the precision recall curve (AUPRC) were returned to participants via the Synapse platform. No logs, model parameters, or other information other than the performance metrics, were returned to participants after models were applied to the UW patient repository (including the final models themselves). Participants were allowed a total of 10 hours to train and test their models in this environment. The models were run on a server environment with access to 70 GB of RAM, 32 2.3 GHz CPU cores and no GPUs during this process. Model predictions were never linked to identified patient records.

Challenge timeline and data

The EHR DREAM Challenge lasted from September 9, 2019 to February 23, 2020 and was conducted in 3 phases: the open phase, the leaderboard phase, and the validation phase. During the open phase, participants could submit models for technical validation using only synthetic data in the Challenge cloud environment (Figure 1, Stage 1). During the leaderboard phase, models that were technically validated against the synthetic data were applied to the leaderboard training data and evaluated against the leaderboard validation data (Table 1). The leaderboard training and validation data represented patients who had a clinical visit from January 2010 to August 2018 with possible death records spanning into

February 2019. During the open and leaderboard phases, new data accumulated in the UW EHR. We gathered this data (holdout test data) which represented patients who visited UW medical facilities between January and June of 2019 and whose last visit record was at least 180 days prior to December 31, 2019 (the end of the prospectively collected data). Patients who were originally in the training data but who later had clinical records in 2019 were removed from the training data and included in the holdout test data. During the validation phase, models were retrained on the leaderboard training data (excluding the patients transitioned to the holdout test data) and were tested on this new prospectively collected holdout test data. (See [Supplementary Material](#) for details on the timeline and data generation processes.)

Model evaluation

Challenge evaluation metrics

The AUROC was used as the primary metric for assessing model performance.²⁶ An empirical Bayes factor, K , (bootstrapped distributions $n = 10\,000$) was computed to determine if the AUROCs between 2 models were consistently different. If 2 models were found to have a small Bayes factor ($K < 19$), we used the AUPRC as a tie-breaking metric. Both the AUROC and the AUPRC were computed for all submissions and were used to rank teams on the Challenge leaderboard. During the leaderboard phase, models were scored against the leaderboard validation data to build the initial leaderboard phase model ranking. During the final validation phase, models were scored against the prospectively collected holdout test set. The top performing teams were declared from the resulting validation phase model rankings. This holdout test set served to evaluate models on prospectively

Table 1. Demographic makeup as a percentage of the individual data sizes across the different versions of data used in the DREAM Challenge.

Demographic	Leaderboard phase		Validation phase			Postchallenge Resplit	
	Training (<i>n</i> = 979 184)	Validation (<i>n</i> = 284 883)	Training (<i>n</i> = 942 381)	Validation (<i>n</i> = 200 855)	Holdout test (<i>n</i> = 168 708)	Training (<i>n</i> = 1 067 084)	Validation (<i>n</i> = 273 597)
Age (%)							
0-17	6.12	6.38	6.18	7.31	6.04	5.62	9.42
18-34	23.77	22.18	23.84	24.61	20.98	22.07	29.2
35-64	46.82	45.31	46.68	44.84	45.58	47.36	41.12
65-99	22.86	26.03	22.85	23.13	27.33	24.55	20.09
100+	0.35	0.09	0.37	0.1	0.06	0.34	0.1
Race (%)							
White	54.42	64.05	54.39	62.76	66.74	58.07	53.83
Asian	8.36	10.29	8.36	10.59	9.57	8.9	8.44
Black	6.3	7.41	6.22	6.82	8.39	6.81	5.51
Other/Nan	30.93	18.25	31.03	19.83	15.30	26.22	32.22
Gender (%)							
Female	52	54.2	51.92	53.82	54	52.59	51.6
Male	47.94	45.78	48.02	46.16	46	47.37	48.35
Other/Nan	0.05	0.01	0.05	0.02	0.01	0.04	0.05
Ethnicity (%)							
Hispanic	5.79	6.45	5.78	6.47	7.09	5.80	7.03
Not Hispanic	50.17	77.13	49.90	75.42	80.09	56.09	63.71
Other/Nan	44.04	16.42	44.31	18.11	12.81	38.11	29.26
Mortality Status (%)							
Passed	0.83	0.75	0.90	1.12	1.32	0.93	1.33
Alive	99.17	99.25	99.10	98.88	98.68	92.55	98.67

All values represent the percentage of the total number of patients in the dataset of interest. We include a 100+ category as a standalone age category because that age range is of questionable quality. This gives some idea to the quality of the data made available.

collected clinical data, testing a models ability to generalize over shifting clinical practice and data collection.

Resplit data validation

During the challenge, we used the validation phase holdout test set to build a final ranking of model performance. This holdout set only contained patients who appeared in the UW medical system between January and June of 2019. This left a 6-month longitudinal gap between the end date of the training data and the start date of the holdout test data. We combined all the datasets (training, validation, and holdout test) and redivided the dataset into an 80/20 split between training and testing data using the same prospective splitting method we used to split the initial leaderboard data (details in the [Supplementary Material](#)). We trained the models on the 80% training data and evaluated the trained models against the 20% test data. This allowed us to compare the effect of the 6-month gap on model performance.

Subpopulation accuracy comparison

For each model, we evaluated how well that model performed across various subpopulations which were defined by different demographic or clinical features including race, gender, ethnicity, age, and type of last visit. Between each demographic strata, we calculated an empirical Bayes factor, K , (bootstrapped distributions $n = 10\ 000$) to determine if a model's AUROCs between demographic strata were consistently different. We ran this experiment on the prospectively gathered validation phase data. Additionally, we evaluated the impact of numerous clinical subcohorts on model accuracies (see [Supplementary Material](#)).

Model features

The top 4 highest scoring teams were asked to adjust their dockerized models to output their trained features as a list of codes/values with associated weights from their trained models. In order to compare features across models, these teams reported which terms (SNOMED, RxNorm, LOINC, etc.) were used during any feature engineering.

Results

The EHR DREAM Challenge on all-cause patient mortality prediction was held between September 9, 2019 and February 23, 2020. Participants were asked to submit software programs that the Challenge organizers—acting as an honest broker—applied to hidden EHR data for training and model validation ([Figure 1](#), [Figure S2](#)). Data was split into training, validation, and holdout testing data sets that were used across 2 phases of the Challenge: a leaderboard phase and a final validation phase. Within the final validation phase, 942 381 patients were available for model training and 168 708 patients were used for model validation, with mortality rates of 0.90% and 1.32%, respectively ([Table 1](#)). The Challenge received a total of 132 submissions from 25 teams that were able to be successfully executed and produced valid predictions.

During the leaderboard phase, of the 25 successfully validated models, 10 teams exceeded AUROC >0.9 . AI4Life led the leaderboard phase—achieving an AUROC = 0.979 (0.977-0.981) and AUPR = 0.614 ([Table 2](#)). In the final validation phase, 15 teams submitted successfully validated models, with 3 teams achieving an AUROC >0.9 ([Table 2](#)). The top performing team, UW-biostat, achieved an AUROC = 0.947 (0.924-0.952) and an AUPR = 0.478. Given the tendency of models to overfit on the leaderboard data, we

Table 2. Top 15 teams and the metrics for their highest performing models.

Team	Leaderboard phase			Validation phase				Postchallenge Resplit		
	AUROC	AUROC 95% CI	AUPR	AUROC	AUROC 95% CI	Delong P-value	AUPR	AUROC	AUROC 95% CI	AUPR
UW-biostat	0.972	(0.969-0.975)	0.524	0.947	(0.942-0.951)	1.70E-04	0.478	0.964	(0.961-0.967)	0.43
Ivanbrugere	0.968	(0.964-0.971)	0.474	0.938	(0.933-0.942)	1.96E-07	0.3	0.956	(0.953-0.96)	0.409
ProActa	0.943	(0.937-0.948)	0.458	0.91	(0.903-0.918)	2.84E-03	0.383	0.904	(0.898-0.91)	0.43
AMbeRland	0.942	(0.937-0.947)	0.288	0.897	(0.89-0.903)	4.18E-02	0.163	0.929	(0.924-0.934)	0.284
DMIS_EHR	0.915	(0.91-0.92)	0.111	0.887	(0.88-0.89)	5.95E-02	0.093	0.939	(0.936-0.943)	0.347
PnP_India	0.958	(0.954-0.963)	0.449	0.876	(0.87-0.883)	1.26E-01	0.182	—	—	—
ultramangod671	0.882	(0.874-0.891)	0.289	0.865	(0.856-0.875)	2.45E-03	0.264	0.868	(0.86-0.876)	0.37
HELM	0.951	(0.948-0.955)	0.323	0.842	(0.834-0.85)	5.65E-01	0.135	—	—	—
AI4Life	0.979	(0.977-0.981)	0.614	0.831	(0.82-0.841)	5.28E-01	0.302	0.971	(0.969-0.974)	0.63
Georgetown— ESAC	0.938	(0.933-0.942)	0.168	0.839	(0.832-0.848)	1.64E-02	0.073	0.938	(0.933-0.941)	0.272
LCSB_LUX	0.956	(0.952-0.959)	0.307	0.82	(0.81-0.829)	8.41E-01	0.116	0.936	(0.932-0.94)	0.201
QiaoHezhe	0.925	(0.92-0.93)	0.16	0.819	(0.81-0.827)	2.92E-01	0.073	—	—	—
chk	0.903	(0.896-0.908)	0.159	0.808	(0.8-0.817)	1.38E-05	0.062	0.811	(0.804-0.818)	0.061
moore	0.955	(0.951-0.958)	0.313	0.771	(0.757-0.784)	9.51E-45	0.122	0.947	(0.943-0.95)	0.377
tgaudelet	0.904	(0.898-0.91)	0.278	0.807	(0.798-0.817)	—	0.201	0.158	(0.151-0.166)	0.007

95% confidence intervals were calculated using bootstrapped ($n = 1000$) distributions. The Delong test P -value was generated by comparing each team's model with the team's model ranked below them. Leaderboard phase scores were generated using the models submitted during the final validation phase.

assessed performance degradation on the validation cohort data. Between the leaderboard and validation phases, the average decrease in AUROC was 0.069 with the top 5 models decreasing by an average of 0.032 and the bottom 8 models decreasing by an average of 0.10. The top 5 models from the validation phase were ranked second, third, 11th, 10th, and 7th, respectively, at the end of the leaderboard phase but were ranked in the top 5 due to having the lowest decrease in performance. Of note, none of the models had impressive calibration curves (Figure S5), likely due to the challenge emphasis on threshold-agnostic evaluation metrics like AUROC and AUPR.

Teams used a variety of machine learning techniques in their submitted models. Of the 15 validated models, 12 were boosted methods (LightGBM,²⁷ XGBoost,²⁸ CatBoost,²⁹ Generalized Boosted Regression³⁰), 2 were logistic regression, and 1 was a neural network (Table S4). Of the top 5 models, 2 were LightGBM, 1 was logistic regression, 1 was CatBoost, and 1 was Generalized Boosted Regression. Each model used a different feature selection method ranging from randomly sampling all available concepts (Team IvanBrugere), carefully selecting a few features from the literature (Team LCSB_LUX), and using the structure of the concept ontologies to roll up low-level granular concepts into broad categories of disease and drugs as features (Team UW-biostat). We developed an ensemble model with the top performing models but observed that the ensemble did not meaningfully improve prediction accuracy over the top models (see Supplementary Material).

Top performing model

UW-biostat's (University of Wisconsin-Madison, Biostatistics and Medical Informatics) model achieved the highest AUROC during the final validation phase. While they were not the highest scoring model during the leaderboard phase, their model had the smallest decrease in AUROC (0.025) of any model between the leaderboard phase and the validation phase (Table 2, Figure 2). The team used ontology-rollup to reduce feature dimensionality and used time binning and

sample reweighting to capture longitudinal characteristics. For model development, they trained and tuned a LightGBM model to predict the mortality risk of each patient. To take into account potential data drift in EHRs,^{4,31-33} the team upweighted more recent patients during optimization and training of their model. A more detailed description of this model and of the top 5 models is provided in the Supplementary Material.

Demographic and clinical cohort evaluation

We evaluated whether models generalized across multiple demographic and clinical groups including race, gender, age, ethnicity, last visit type, and clinical condition cohort. Models were consistently more accurate on Asian patients when compared to any other racial group (Figure 3, Table S2), despite Asian patients only making up 8.4% of the validation data and 9.6% of the validation phase training data (Table 2). Methods varied in their accuracies for other races with some models (eg, UW-biostat, IvanBrugere, Proacta, AMbeRland, DMIS_EHR) scoring higher on White patients compared to Black patients, and others scoring higher on Black patients than White patients (PnP_India, HELM, Georgetown-ESAC, AI4Life) (Figure 3, Table S2).

Without exception, models were more accurate on female patients than on male patients with Bayes factors greater than 10 (strong evidence) for 9 of the top 15 models (Figure S5). As the challenge asked participants to predict mortality status 180 days from the last visit, we examined whether there were differences in model performances based on whether the last visit was inpatient, outpatient, or an emergency room visit. Most models had lower accuracy when the last visit was an outpatient visit, with the exception of 3 models (ultramangod671, Georgetown—ESAC, AI4Life in Figure S9). On patients where the last visit was an emergency room visit, models showed a wide variety of accuracies. In a few cases, models that had an overall lower model accuracy had higher accuracies on patients in the emergency room (compare ProActa to PnP_India in Figure S9). The results of evaluating

Comparison of Model Accuracies Across Challenge Phases

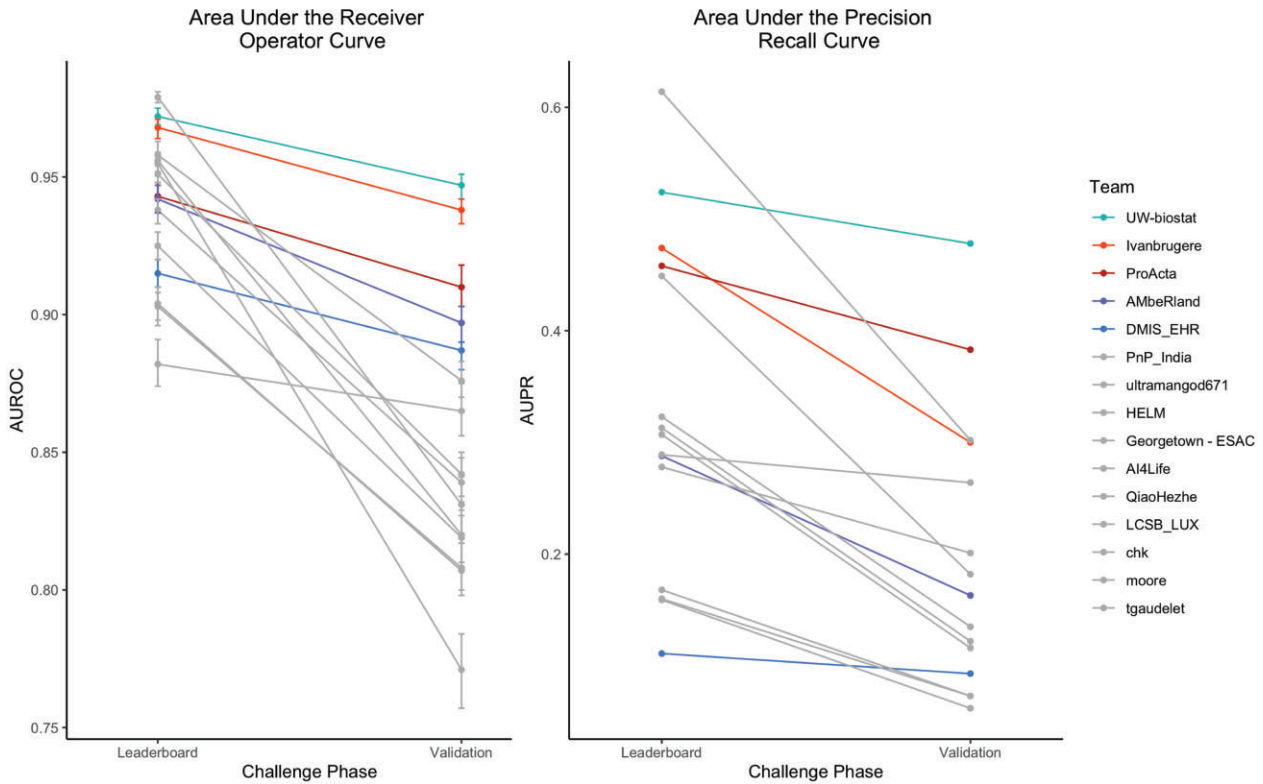


Figure 2. Comparison of model performance between the leader phase and the validation phase (Data in Table 2). All models decreased in AUROC and AUPRC. The top 5 teams' AUROCs decreased the least between the 2 phases. Only the top 5 team's performances are colored. The error bars for the AUROCs represent the 95% confidence interval.

model accuracy across clinical condition cohorts can be found in the [Supplementary Material](#).

Assessment of important features

Table 3 reports the top 10 features of each model, including engineered features (ie, presence or absence of a category of diagnosis or drug) and raw concepts from the data (ie, granular SNOMED or LOINC codes). Some of the highly weighted features included the age of the patient at their last visit, systolic and diastolic blood pressure, heart rate, and a code for Do Not Resuscitate (see Table S4 for the full feature lists).

Discussion

Machine learning models are increasingly regarded as foundational to any precision medicine strategy. The assessment of model accuracy and utility in a healthcare environment is challenged by limited data availability, concerns about breach of protected health information confidentiality, and lack of technical infrastructure, domain expertise and process to systematically manage model evaluations. We implemented an architecture for an unbiased and transparent assessment of methods that overcomes these limitations, and in doing so were able to improve existing methodology. We demonstrated how community challenges can provide an inclusive and rigorous environment for hosting a machine learning clinical trial.

Using the MTD framework, 25 international teams submitted machine learning models to a private clinical dataset that otherwise would have remained inaccessible to these researchers. This was enabled by leveraging a common data model, in this case OMOP, a synthetic dataset for technical development and validation, a cloud environment hosting the synthetic data for pipeline and execution evaluation, standard containerization software, and a secure environment hosting the private clinical data. Docker images and descriptions for all (untrained) models have been made available.

Assessing a wide variety of methods from teams allowed us to evaluate the best approaches and assess intermethod variability when holding the evaluation data constant. Interestingly, even though models were trained and evaluated on the same data, there was variance in model accuracy across different demographic groups. White, Black, and other racial groups showed differences across models, with some models scoring higher on Black patients than White patients and vice versa, while Asian patients were consistently more accurate across nearly all models (Figure 3). This may have to do with the cause of death, as prevalence of different causes of death may vary between different populations. Unfortunately, we did not have access to cause of death data at the time of this analysis. With the exception of the 0-17 age group, method accuracy was inversely correlated with age (Figure S7). One hypothesis for this trend is that younger patients who pass away in 180 days and are coming into the hospital are more likely to have extreme conditions with a higher risk of death,

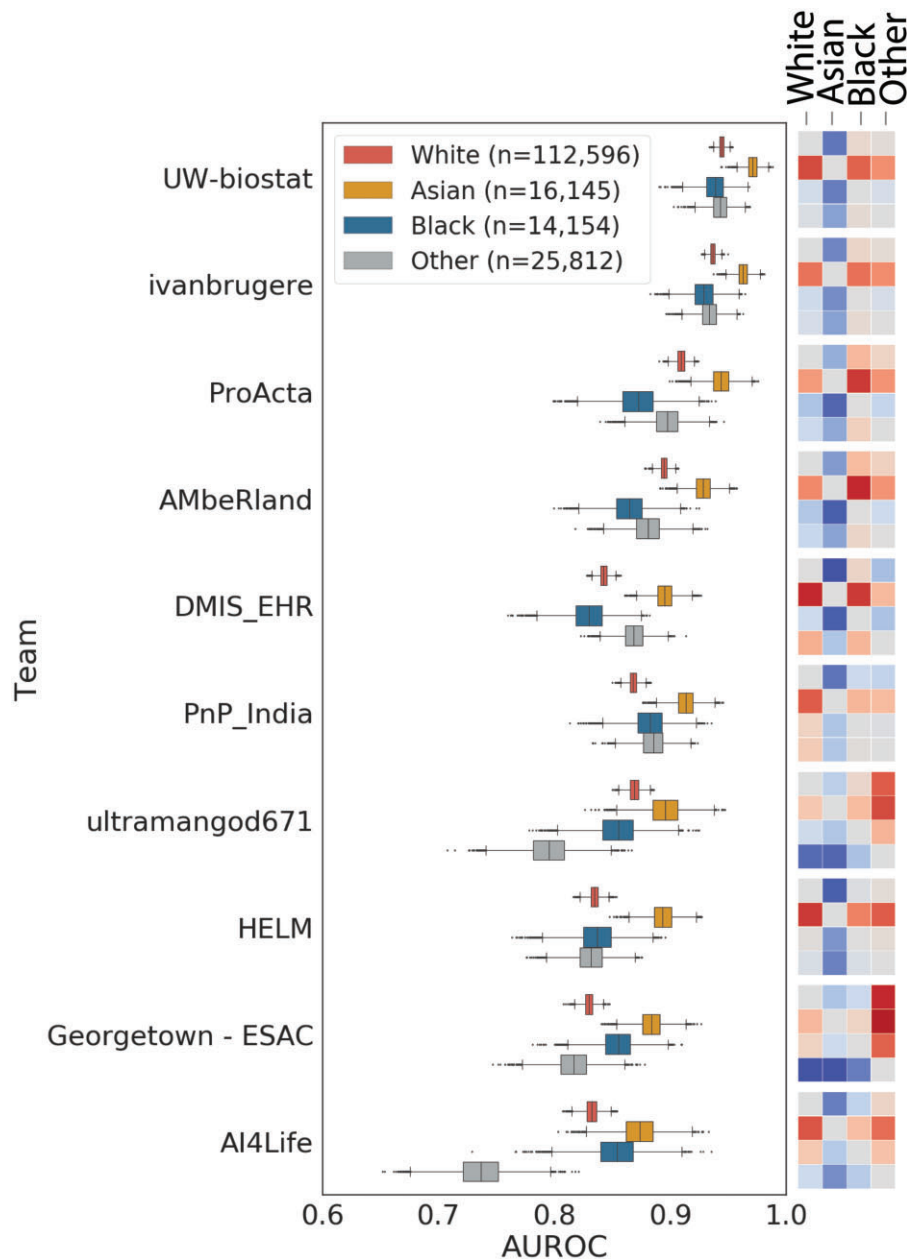


Figure 3. Bootstrapped distributions ($n = 10\,000$) of the top 10 model AUROCs broken down by race. Model predictions were randomly sampled with replacement and scored against the benchmark gold standard. Box-plot center lines represent the median AUROC, box limits represent the upper and lower quartiles, whiskers represent the $1.5\times$ interquartile ranges, and the points represent the outliers. Comparisons were made between each category of race and Bayes values calculated to assess the level of evidence for the model having a higher accuracy on racial category compared to another category. The heat maps represent the log of the calculated Bayes factors when comparing racial groups within each model. The darker the red, the stronger the evidence for the racial category being higher than the comparison category. Bayes factor values range from 10 000 to 0.0001. The darker the blue, the stronger the evidence for the racial category being lower than the comparison category. The color scale is normalized across all comparisons. Raw Bayes factor values can be found in [Table S2](#).

while older patients are simply more likely to have diseases and health problems in general, making it more difficult to predict risk of death. Models also had varied accuracies from the last visit type, highlighting the need to develop context-specific clinical prediction algorithms (Figure S9). In other cases, models were aligned in their bias, universally scoring higher on females than males (Figure S5). We found no meaningful difference in model accuracy between Hispanic and Non-Hispanic ethnicity (Figure S5). This evaluation provided UW Medicine with insights into statistical heterogeneity of

these methods since each method is trained on the same data, solving the same problem and was validated on the same data.

Evaluating models in a pseudoprospective manner allowed us to assess how models would perform over time in the UW environment. We found that most models decreased in performance in the validation phase when compared to the leaderboard phase (Table 2, Figure 2). This is in line with the literature, as previous studies have shown that the utility of clinical data can have a half-life of as little as 3 months.⁴ As

Table 3. The top 10 weighted features as reported from the top 4 performing teams.

Rank	UW-biostat	IvanBrugere	ProActa	AMBeRand
1	Age	Not for resuscitation	Age	Age
2	Average pulse	Temperature	Creatinine in serum	Not for resuscitation
3	Average diastolic blood pressure	Albumin in plasma	Heart rate	Antineoplastic chemotherapy regimen
4	Average systolic blood pressure	History of clinical finding in subject	Inpatient visit	Lactate dehydrogenase (LD), (LDH)
5	Latest systolic blood pressure value	Natriuretic peptide B [Mass/volume] in serum or plasma	Blood typing, serologic; ABO	Administration of antineoplastic agent
6	Year of last visit	Racial variable (White)	Palliative care	Patient encounter procedure
7	Latest pulse measurement	Antineoplastic chemotherapy regimen	Albumin in plasma	Secondary malignant neoplasm of lung
8	Latest diastolic blood pressure value	Protein in plasma	Hematocrit of blood by automated count	Disorder of lung
9	Latest glucose measurement	Heart rate	Neutrophils/100 leukocytes in blood by automated count	Dexamethasone
10	Indicator: unknown conditions	Essential hypertension	Cholecalciferol	Bacterial culture

Features were ordered by their model weight and assigned a rank out of all available features. Feature names were either reported by the teams or were mapped using the OMOP concept table from the reported concept ids.

an example from the UW data, 19.5% of patients with a condition record in 2018 had the concept code for “malignant tumor of prostate” while in 2019, only 1.5% of patients with a condition record had that code. Comparing results from the postchallenge resplit data to the validation phase final results, the majority of models performed better when the training data was longitudinally closer to the test data (Table 2). UW-biostat explicitly down-weighted older data and relied more heavily on the most recent 6 months of data in the training dataset resulting in their model having the lowest decrease in performance of any model. By combining this technique with their “roll up” feature engineering, UW-biostat’s model proved to be the most robust against longitudinal changes, covariate shift, and cohort changes. In contrast, some models dropped by a significant margin. For instance, AI4Life’s model was among the most accurate during the leaderboard phase, but dropped to tenth in the validation phase (Table 2). While it is difficult to completely account for their drop, one possible explanation is their overall lower accuracy on first time visiting patients (Figure S8) combined with the increase of first-time visitors in the validation data (leaderboard phase data—13.8% compared to the validation phase data—19.7%). AI4Life’s model had a high score on the resplit data, indicating that their model was susceptible to covariate drift as well. Evaluating models prospectively or pseudoprospectively brings us closer to understanding how the models will perform in a live clinical setting.

Study limitations

There were several notable limitations of this study. First, mortality prediction is not immune to censoring, and is susceptible to an open world limitation as some patients may die out of state or outside UW clinical care without the ability to map their death to UW clinical records.³⁴ Improper record keeping, mismatched information, and delayed record entry to the state death records can limit the capture and sensitivity of the state death record mapping process, leading to improper true negative attribution (ie, patients are designated as still alive but have actually passed away). Second, all-cause mortality is not a clinically actionable question, as these models are not specific enough for clinical action. Future EHR

challenges will focus on clinically actionable prediction questions. Third, we set a model runtime limit of 10 hours to limit the burden to the University of Washington secure servers; however, this also limited the types of models participants were able to build and excluded models like deep learning. However, this limit forced participants to carefully consider the efficiency of their algorithms. Fourth, while we did prospectively evaluate models on a future holdout set to control for overfitting, evaluation on data from one site does not fully assess model generalizability. For future assessments, we hope to partner with other hospitals to externally validate models. Finally, we acknowledge that the framework described in this article does not fully address all the complexities associated with AI implementation within a healthcare system. Additional considerations not addressed here include: clinical utility, model interpretability, integration into clinical workflows and decision-making processes, and clinical adoption. Nonetheless, a systematic, rigorous assessment of models’ performance over time is an integral step, and the framework described in this manuscript is likely to accelerate the deployment of AI into clinical practice for the benefit of patients.

Conclusion

Machine learning promises to enhance patient care and improve health outcomes; however, if not properly vetted and evaluated, risks and negative effects may be introduced. These risks include breach of privacy in the development and assessment of methods, inaccuracy or methodological bias when deployed, and the gradual loss of accuracy over time as data and business practices change. This study highlights these challenges by showing that while highly accurate methods are possible, even methods from world-class scientists have considerable variability and that variability (such as differences in accuracy based on race or gender) may not be detectable from high-level measures such as population-level AUCs or accuracy. Our framework enables this assessment and also brings the community challenge culture to private datasets, in this case data that is subject to the HIPAA privacy rule. Further, machine learning methods may be able to address some causes of treatment disparities but may cause others for patients without rich longitudinal data, patients of certain

races, gender, or age. Based on these results, we believe that multisite standardized architecture and independent oversight is required to truly assess new methods.

Author contributions

T.B. contributed to the writing of the manuscript, the design and execution of the study, and the analysis and interpretation of the data. T.S. contributed to the writing of the manuscript and the design and execution of the study. Y.Y. contributed to the writing of the manuscript, the design of the study, and the analysis and interpretation of the data. T.Y. contributed to the design and execution of the study. J.P. contributed to the design and execution of the study. Ji.G., G.C., L.C., Z.N., I.B., R.R., Al.P., Au.P., Y.C., S.L., J.C., I.L., S.K., and J.K. contributed by submitting mortality prediction models during the challenge, by adapting their models to output feature information during the collaboration phase, and by revising the manuscript. S.D.M. contributed to the design of the study, the writing of the manuscript, and the acquisition and interpretation of the data. Ju.G. contributed to the design of the study, the interpretation of data, and the writing of the manuscript.

Supplementary material

[Supplementary material](#) is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the Clinical and Translational Science Awards Program National Center for Data to Health funding by the National Center for Advancing Translational Sciences at the National Institutes of Health (grant numbers U24TR002306 and UL1 TR002319). Any opinions expressed in this document are those of the Center for Data to Health community and the Institute for Translational Health Sciences and do not necessarily reflect the views of the National Center for Advancing Translational Sciences, team members, or affiliated organizations and institutions. T.B., Y.Y., S.D.M., Ju.G., T.Y., T.S., and J.P. were supported by grant number U24TR002306. T.B., J.P., and S.D.M. were supported by grant number UL1 TR002319.

Conflict of interest

None declared.

Data availability

Due to its sensitive nature and to comply with institutional policy around HIPAA protected data privacy, investigators who wish to access the raw, row-level data will need to directly collaborate with the University of Washington. An Institutional Review Board approval letter and a Data Use Agreement for external collaborators will be necessary. Data access requests can be made through this web page: <https://www.iths.org/investigators/services/bmi/clinical-data-extraction/>. We have included the aggregate demographic and mortality status distributions of the patient cohorts (Table 1). Figure 2 visualizes Table 2. The bootstrapped distributions used to generate the box plots in Figure 3 are included as

Source Data, and the raw Bayes factors used to generate the heatmaps in Figure 3 are provided in Table S2.

Code availability

The individual docker models evaluated in this challenge are available on Synapse at <https://doi.org/10.7303/syn26433349>. You must have a Synapse account and agree to the Terms of Use in order to download the docker images. The code used to generate the figures and tables can be found at <https://github.com/Sage-Bionetworks-Challenges/EHR-DREAM-Challenge-Patient-Mortality-Prediction>. Figure S10b was generated using the VEDx tool (<https://github.com/UWMooneyLab/VEDx>).

Authors of the Patient Mortality Prediction DREAM Challenge Consortium

Aaron Lee¹⁴, Ali Salehzadeh-Yazdi¹⁵, Alidivinas Prusokas¹⁰, Anand Basu¹⁶, Anas Belouali¹⁷, Ann-Kristin Becker¹⁸, Ariel Israel¹⁹, Augustinas Prusokas²⁰, B. Winter²¹, Carlos Vega Moreno²², Christoph Kurz^{23,24}, Dagmar Waltemath²¹, Darius Schweinoch¹⁸, Enrico Glaab²², Gang Luo²⁵, Guanhua Chen⁵, Helena U. Zacharias²⁶, Hezhe Qiao²⁷, Inggeol Lee¹², Ivan Brugere⁸, Jaewoo Kang¹², Jifan Gao²⁸, Julia Truthmann²¹, JunSeok Choe¹², Kari A. Stephens²⁹, Lars Kaderali¹⁸, Lav R. Varshney^{30,31}, Marcus Vollmer^{18,32}, Maria-Theodora Pandi³³, Martin L. Gunn³⁴, Meliha Yetisgen²⁵, Neetika Nath³⁵, Noah Hammarlund²⁵, Oliver Müller-Stricker¹⁸, Panagiotis Togiass³⁶, Patrick J. Heagerty³⁷, Peter Muir^{16,38}, Peter Banda²², Renata Retkute⁹, Ron Henkel²¹, Sagar Madgi³⁹, Samir Gupta¹⁷, Sanghoon Lee¹², Sean Mooney², Shabeeb Kannattikuni¹⁷, Shamim Sarhadi⁴⁰, Shikhar Omar³⁹, Shuo Wang⁴¹, Soumyabrata Ghosh²², Stefan Neumann³⁵, Stefan Simm¹⁸, Subha Madhavan⁴¹, Sunkyu Kim⁴², Thomas Von Yu¹, Venkata Satagopam²², Vikas Pejaver²⁵, Yachee Gupta²¹, Yonghwa Choi¹², Zofia Nawalany⁶, Łukasz Charzewski^{6,7}

¹⁴Department of Ophthalmology, University of Washington, ¹⁵Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany, ¹⁶ESAC Inc., Rockville, United States, ¹⁷Innovation Center for Biomedical Informatics, Georgetown University, Washington DC, United States, ¹⁸Institute of Bioinformatics, University Medicine Greifswald, Greifswald, Germany, ¹⁹Department of Research and Data, Division of Planning and Strategy, Clalit Health Services, Tel-Aviv, Israel, ²⁰Department of Life Sciences, Imperial College London, London, UK, ²¹Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany, ²²Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, ²³Helmholtz Zentrum München, Institute of Health Economics and Health Care Management, Neuherberg, Germany, ²⁴Munich School of Management and Munich Center of Health Sciences, Ludwig-Maximilians-Universität München, Munich, Germany, ²⁵Department of Biomedical Informatics and Medical Education, University of Washington, ²⁶Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany, ²⁷Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China, ²⁸School of Medicine and Public Health, University of Wisconsin-Madison, Madison, United States, ²⁹Department of

Psychiatry and Behavioral Sciences, University of Washington, ³⁰Salesforce Research, Palo Alto, United States, ³¹University of Illinois at Urbana-Champaign, Urbana, United States, ³²German Centre for Cardiovascular Research: DZHK, Greifswald, Germany, ³³School of Health Sciences, University of Patras, Rion, Greece, ³⁴Department of Radiology, University of Washington, ³⁵Department of Bioinformatics, University Medicine Greifswald, Greifswald, Germany, ³⁶School of Health Sciences, University of Patras, Rion, Greece, ³⁷Department of Biostatistics, University of Washington, ³⁸PJM Consulting LLC, San Diego, United States, ³⁹ZS Associates, Bengaluru, Karnataka, India, ⁴⁰Department of Medical Biotechnology, Tabriz University of Medical Sciences, Tabriz, Iran, ⁴¹Georgetown University, Washington DC, United States, ⁴²Department of Computer science, College of Informatics, Korea University, Seoul, South Korea

References

- Goldstein BA, Navar AM, Pencina MJ, Ioannidis, JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc.* 2017;24(1):198-208.
- Jauk S, Kramer D, Großauer B, et al. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. *J Am Med Inform Assoc.* 2020;27(9):1383-1392.
- Norel R, Rice JJ, Stolovitzky G. The self-assessment trap: can we all be better than average? *Mol Syst Biol.* 2011;7(1):537.
- Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int J Med Inform.* 2017;102:71-79.
- Hammarlund N. Racial treatment disparities after machine learning surgical risk-adjustment. *Health Serv Outcomes Res Method.* 2021;21(2):248-286.
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453.
- Saez-Rodriguez J, Costello JC, Friend SH, et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet.* 2016;17(8):470-486.
- Cai B, Li B, Kiga N, et al. Matching phenotypes to whole genomes: lessons learned from four iterations of the personal genome project community challenges. *Hum Mutat.* 2017;38(9):1266-1276.
- Daneshjou R, Wang Y, Bromberg Y, et al. Working toward precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Hum Mutat.* 2017;38(9):1182-1192.
- Andreoletti G, Pal LR, Moulton J, Brenner SE. Reports from the fifth edition of CAGI: the critical assessment of genome interpretation. *Hum Mutat.* 2019;40(9):1197-1201.
- Kryshtafovich A, Schwede T, Topf M, Fidelis K, Moulton J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins.* 2021;89(12):1607-1617.
- Zhou N, Jiang Y, Bergquist TR, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* 2019;20(1):244.
- Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. *Nat Biotechnol.* 2018;36(5):391-392.
- Bergquist T, Yan Y, Schaffter T, et al. Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction. *J Am Med Inform Assoc.* 2020;27(9):1393-1400. <https://doi.org/10.1093/jamia/ocaa083>
- Radijojac P, Clark WT, Oron TR, et al. A large-scale evaluation of computational protein function prediction. *Nat Methods.* 2013;10(3):221-227.
- Jiang Y, Oron TR, Clark WT, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.* 2016;17(1):184.
- Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol.* 2005;15(3):285-289.
- Weng SF, Vaz L, Qureshi N, Kai J. Prediction of premature all-cause mortality: a prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One.* 2019;14(3):e0214365.
- Fahey M, Rudd A, Béjot Y, Wolfe C, Douiri A. Development and validation of clinical prediction models for mortality, functional outcome and cognitive impairment after stroke: a study protocol. *BMJ Open.* 2017;7(8):e014607.
- Smolin B, Levy Y, Sabbach-Cohen E, Levi L, Mashiach T. Predicting mortality of elderly patients acutely admitted to the Department of Internal Medicine. *Int J Clin Pract.* 2015;69(4):501-508.
- Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* 2018;1:18.
- Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574-578.
- Enterprise Container Platform | Docker. Docker. <https://www.docker.com/>. Accessed August 8, 2023.
- Omberg L, Ellrott K, Yuan Y, et al. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet.* 2013;45(10):1121-1126.
- Lambert CG, Amritansh, Kumar P. Transforming the 2.33M-patient Medicare synthetic public use files to the OMOP CDMv5: ETL-CMS software and processed data available and feature-complete. 2016.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29-36.
- Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Guyon I, ed. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.; 2017:3146-3154.
- Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2016:785-794.
- Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. In: Bengio S, ed. *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc.; 2018:6638-6648.
- Ridgeway G. Generalized Boosted Models: A guide to the gbm package. <https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf>. Accessed August 9, 2023.
- Ghassemi M, Naumann T, Schulam P, et al. A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl Sci Proc.* 2020;2020:191-200.
- Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc.* 2017;24(6):1052-1061.
- Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc.* 2017;24(6):1052-1061.
- Dessimoz C, Škunca N, Thomas PD. CAFA and the open world of protein function predictions. *Trends Genet.* 2013;29(11):609-610.