

AutoPrognosis 2.0: Democratizing Diagnostic and Prognostic Modeling in Healthcare with Automated Machine Learning

Fergus Imrie^a, Bogdan Cebere^b, Eoin F. McKinney^c, Mihaela van der Schaar^{b,d}

^a*Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA*

^b*Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK*

^c*Department of Medicine, University of Cambridge, Cambridge, UK*

^d*The Alan Turing Institute, London, UK*

Abstract

Diagnostic and prognostic models are increasingly important in medicine and inform many clinical decisions. Recently, machine learning approaches have shown improvement over conventional modeling techniques by better capturing complex interactions between patient covariates in a data-driven manner. However, the use of machine learning introduces a number of technical and practical challenges that have thus far restricted widespread adoption of such techniques in clinical settings. To address these challenges and empower healthcare professionals, we present a machine learning framework, AutoPrognosis 2.0, to develop diagnostic and prognostic models. AutoPrognosis leverages state-of-the-art advances in automated machine learning to develop optimized machine learning pipelines, incorporates model explainability tools, and enables deployment of clinical demonstrators, *without* requiring significant technical expertise. Our framework eliminates the major technical obstacles to predictive modeling with machine learning that currently impede clinical adoption. To demonstrate AutoPrognosis 2.0, we provide an illustrative application where we construct a prognostic risk score for diabetes using the UK Biobank, a prospective study of 502,467 individuals. The models produced by our automated framework achieve greater discrimination for diabetes than expert clinical risk scores. Our risk score has been

implemented as a web-based decision support tool¹ and can be publicly accessed by patients and clinicians worldwide. In addition, AutoPrognosis 2.0 is provided as an open-source python package. By open-sourcing our framework as a tool for the community, clinicians and other medical practitioners will be able to readily develop new risk scores, personalized diagnostics, and prognostics using modern machine learning techniques.

Software: <https://github.com/vanderschaarlab/AutoPrognosis>

1. Introduction

Machine learning (ML) systems have the potential to revolutionize medicine and become core clinical tools [1]. However, there are a diverse set of challenges that must be overcome prior to routine and widespread ML adoption [2, 3]. In particular, there are substantial technical challenges in developing, understanding, and deploying ML systems which currently render them largely inaccessible for medical practitioners [3, 4, 5, 6].

In an attempt to address this, we previously developed AutoPrognosis, an automated machine learning (AutoML) framework to train predictive models [7]. This framework has since been applied to derive prognostic models for cardiovascular disease [8], cystic fibrosis [9], and breast cancer [10], among a number of other indications [11, 12, 13, 14, 15, 16]. However, our initial approach had significant limitations from both algorithmic and usability perspectives.

Consequently, in this work, we describe AutoPrognosis 2.0, which addresses all major obstacles limiting the development, interpretation and deployment of ML methods in medicine and represents a step-change in diagnostic and prognostic modeling. In particular, we believe this is the world’s first method that can simultaneously: (1) solve classification, regression, and time-to-event problems; (2) optimize ML pipelines, determine the most appropriate models, and automatically tune hyperparameters; (3) identify key variables and novel risk factors, enabling clinicians to select different numbers of variables and understand the value of information; (4) provide a diverse range of model explanations, including feature-based, example-based, and

¹<https://autoprognois-biobank-diabetes.streamlitapp.com/>

closed-form risk equations; and (5) produce web-based applications, allowing models to be readily shared with the clinical community.

In this paper, we outline the major challenges facing clinical development and translation of diagnostic and prognostic modeling. We then describe our approach, AutoPrognosis 2.0, and detail how it addresses each challenge. Finally, we demonstrate the application of AutoPrognosis 2.0 in an illustrative scenario: prognostic risk prediction of diabetes using a cohort of 502,467 individuals from UK Biobank. However, we emphasize that AutoPrognosis can be applied to construct diagnostic and prognostic models for *any* disease or clinical outcome, and is explicitly designed to make model building accessibly by non-ML experts. We have open-sourced AutoPrognosis 2.0 as a tool for the community, allowing clinicians or non-expert users to adopt the automated framework to robustly and reproducibly develop optimized personalized diagnostics, prognostics, and risk scores using modern machine learning techniques.

2. Challenges in Diagnostic and Prognostic Modeling

There are numerous obstacles to developing and deploying diagnostic and prognostic models that currently prevent healthcare professionals from capitalizing on recent algorithmic advances [1]. Our work seeks to empower clinicians, medical researchers, epidemiologists, and biostatisticians through an accessible, automated framework capable of identifying optimal solutions to all major obstacles limiting ML model building with minimal need for technical expertise. We begin by describing the seven major challenges faced by these communities and how they are addressed by AutoPrognosis 2.0.

Challenge 1. Developing powerful ML pipelines

Developing performant ML models remains complex and typically involves significant time and effort, even for expert ML practitioners. Indeed, some estimates suggest over 95% of work is expended on software technicals, leaving less than 5% for addressing the medical or scientific problem at hand [17]. This is further complicated by the myriad of choices that must be made when developing a new predictive model for diagnosis or prognosis, such as: what imputation strategy should be used; how should the data be preprocessed; what (ML) model is best suited for the specific task; what configuration of hyperparameters should be used. These decisions affect each other, thus cannot be made in isolation; further, the optimal choices not only

Challenge 1. Developing powerful ML pipelines
AutoPrognosis uses AutoML to automate pipeline configuration, performing missing value imputation, feature processing, model selection, and hyperparameter optimization.
Challenge 2. Understanding the value of ML and when it is necessary
AutoPrognosis compares a range of ML methods to traditional approaches and automatically identifies what approach is best.
Challenge 3. Determining the value of information
AutoPrognosis can quantify the value of including additional predictors, enabling systematic identification of optimal variables.
Challenge 4. Understanding and debugging ML models
AutoPrognosis incorporates seven state-of-the-art interpretability methods, allowing models to be understood and debugged as they are generated.
Challenge 5. Making ML models accessible and usable
AutoPrognosis provides a platform to share model outputs by automating the creation of web-based applications.
Challenge 6. Deciding when and if to update clinical models
AutoPrognosis can quantify the benefit of additional data or new predictive variables, and automatically determine the optimal system for the new dataset.
Challenge 7. Transparent reproducibility
AutoPrognosis provides a standardized, publicly available framework, facilitating reproducibility.

Table 1: Major challenges facing clinical development of diagnostic and prognostic models and how these are addressed by AutoPrognosis. See Section 2 for more detail.

vary between applications, but also can change over time as more data is collected and clinical practice changes [18].

Few resources are available to help empirically define optimal computational pipelines. AutoPrognosis 2.0 addresses this by incorporating an AutoML approach within a standardized framework, automating the process of pipeline configuration. AutoPrognosis navigates a broad algorithmic search space in an efficient fashion, systematically performing missing value imputation, feature processing, model selection, and hyperparameter optimization in an unbiased manner without the need for human intervention or expert insight. This avoids arbitrary parameter selection and ensures standardization of pipelines, facilitating both reproducibility and optimized model performance. Critically, this democratizes the model building step, eliminating the requirement for expert ML knowledge and making cutting-edge methodology accessible to all, freeing healthcare domain experts to define and address the core clinical problems.

Challenge 2. Understanding the value of ML and when it is necessary

Traditional approaches, such as linear regression and Cox proportional hazard models [19], are widely used and accepted across healthcare. Before replacing these established methods, it is vital to understand whether ML is valuable for a given problem and quantify the benefit of ML systems. Indeed, there is no “free lunch” and we should not expect ML to always outperform existing approaches. Several recent examples exist that present settings where comparatively “simple” approaches outperformed ML [20, 21]. AutoPrognosis 2.0 can be used to compare a range of ML methods to traditional approaches at minimal technical cost to the user. Furthermore, since these solutions are included in the algorithmic search space, AutoPrognosis will automatically identify whether such approaches are indeed best or if more complex ML models are required.

Challenge 3. Determining the value of information

Selecting which variables to include in a predictive model represents a key decision that not only impacts model performance but also the ease of subsequent clinical use since any feature used will need to be collected in an ongoing manner to use such systems. Thus, understanding the *value* of an individual variable and the information it provides is critical. Often, this is assessed by univariate statistical analysis or other selection methods such as forward selection or backwards elimination [22]. AutoPrognosis 2.0 provides

methods to test and quantify the value of including additional predictors, allowing systematic identification of optimal variables in an informed manner.

Challenge 4. Understanding and debugging ML models

A predictive clinical model must be more than just accurate, it must be interpretable. Without a transparent understanding of *how* a model makes predictions it may act in unintended and undesirable ways, for example learning incorrect or aberrant features unique to the training data [23, 24]. This debugging step is critical for building model trust [5] and cannot be achieved without interpretation of the training features or cases that support model accuracy. It is clear that clinical deployment of an interpretable model is supported by the additional trust gained by understanding the models performance [25].

Furthermore, a clear understanding of computational models is now a requirement for deployment in healthcare systems globally: in the United States, the FDA demands “transparency about the function and modifications of medical devices” as a key safety aspect [26], while Article 22 of GDPR legislation in the EU requires that “meaningful information about the logic involved” be provided [27]. To achieve this transparency, interpretable outputs of a specific form are typically required. For example, the American Joint Committee on Cancer requires explicit risk equations [28]. The ‘black-box’ nature of many ML methods means that they remain inherently uninterpretable and require specialized methods to unravel the underlying rationale for predictions. In AutoPrognosis 2.0, we have incorporated seven state-of-the-art interpretability methods allowing researchers to understand and debug ML models as they are generated.

Challenge 5. Making ML models accessible and usable

Predictive models need to be accessible to be used in clinical practice. This step often limits adoption, since bespoke deployment can result in significant costs and reliance on technical expertise. While full clinical deployment may require additional systems (e.g. due to regulatory requirements), a standardized, user-friendly solution to rapidly visualize and share models is also a necessary part of both debugging and confirming clinical acceptance. AutoPrognosis 2.0 provides a platform to share model outputs by automating the creation of web-based applications, allowing clinicians to explore predictions in diverse scenarios.

Challenge 6. Deciding when and if to update clinical models

Over time, more data is collected, new variables are measured, and even clinical practice changes. For the former, existing clinical predictive models might benefit from additional data or features, while in the latter case, model performance may degrade [18]. However, deciding whether to update a clinical model is not a decision to be made lightly, since beyond model building, further regulatory approval might be necessary and the updated model will need to be redeployed. AutoPrognosis can help answer this difficult question by quantifying the benefit of additional data and new predictive variables, while also automatically determining the optimal system configurations for the new dataset, which may have changed.

Challenge 7. Transparent reproducibility

Reproducibility is a fundamental requirement for the acceptance and adoption of any predictive model. While transparently reproducing a model’s output on a given dataset is conceptually simple, several factors can confound this necessary step. Serial data releases, code updates and even inherent properties of ML algorithms (for example, stochastic descent methods can give different answers even when run repeatedly on the same data) can conspire to make ML model building less reproducible than it should be [29]. These issues demonstrably obstruct translation of clinical prediction and erode trust in ML approaches [30, 31]. AutoPrognosis 2.0 addresses this major challenge by providing a standardized, publicly available framework to train predictive models, allowing straightforward demonstration of reproducibility on source data.

3. AutoPrognosis 2.0

AutoPrognosis 2.0 is an algorithmic framework and software package that allows healthcare professionals to leverage ML to develop diagnostic and prognostic models. Our framework employs automated machine learning [32] to tackle the challenges faced by clinical users. By automating the optimization of ML pipelines involving data processing, model development, and model training, we reduce the burden on technical experts and turn deriving ML models from an art to a science, democratizing machine learning and opening the field to non-ML domain experts, such as clinicians. We believe that AutoPrognosis 2.0 represents a step-change in algorithmic and software

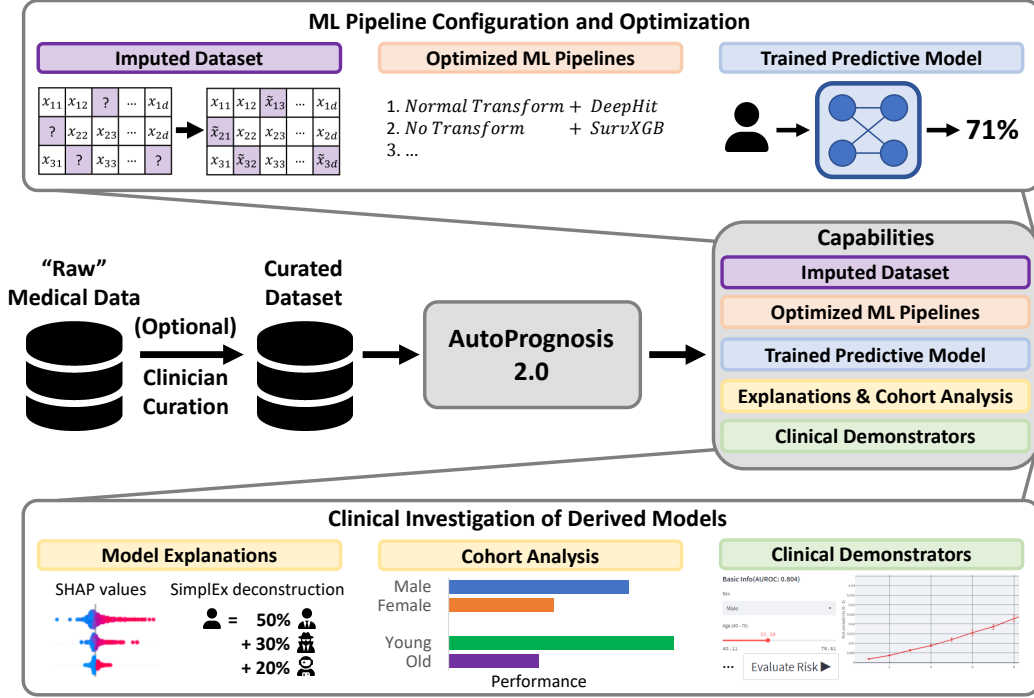


Figure 1: Overview of the AutoPrognosis 2.0 framework. AutoPrognosis takes either raw or curated medical datasets and provides an imputed dataset, a report detailing the optimized machine learning pipelines, a diagnostic or prognostic model, explanations, and a web-based interface for clinicians to interact with and use the derived model.

capabilities and can unlock the potential of ML in healthcare for clinical researchers *without* the requirement for extensive technical capabilities.

AutoPrognosis 2.0 empowers healthcare professionals with the following capabilities:

1. Build highly performant ML pipelines for classification, regression and time-to-event analysis, optimized specifically for the data at hand.
2. Understand when ML provides benefits over traditional regression models, and thus when ML is valuable.
3. Enable principled selection of variables and allow users to understand the value of information.
4. Explain and debug ML models using diverse interpretability methods.
5. Update systems whenever the available data changes to ensure the best possible clinical models.
6. Provide confidence in the reproducibility of models.

Overview

After a clinician has determined an appropriate cohort of patients and an outcome of interest, the AutoPrognosis framework handles all steps in the computational pipeline: missing data imputation, feature processing, model selection and fitting, model interpretability or explanations, and production of clinical demonstrators. Together, we believe AutoPrognosis significantly reduces the technical expertise necessary to derive powerful prognostic models, empowering clinical users and democratizing machine learning in health-care. An overview of AutoPrognosis 2.0 is provided in Figure 1. Below, we provide a summary of each of the core components of AutoPrognosis.

Missing data imputation

Medical datasets are often incomplete; however, most models require complete data as input, thus imputation is a necessary first step. There are many different imputation methods available, ranging from traditional statistical approaches such as mean imputation to well known alternatives such as MICE [33] and MissForest [34]. We include eight common imputation algorithms in AutoPrognosis for users to select if they desire a specific imputation method.

In addition, we also include a state-of-the-art AutoML approach for imputation, HyperImpute [35]. HyperImpute is a generalized iterative imputation algorithm that automatically configures feature-wise imputation models. HyperImpute inherits the usual properties of classical iterative imputation algorithms [33, 36, 37] while benefiting from an automated model selection and hyperparameter optimization procedure that allows the most appropriate model to be chosen for each feature. HyperImpute optimizes over five classes of model, with a total of 29 configurable hyperparameters. For additional details, we refer to the recent technical report detailing HyperImpute [35]. HyperImpute is the recommended imputation strategy in AutoPrognosis, unless a specific method is preferred by the user. Alternatively, the imputation step can be jointly optimized as part of a larger pipeline.

Developing optimized ML pipelines

After imputation, we construct ML pipelines consisting of feature processing, model selection, and model fitting. Given an objective function, these steps are jointly optimized using AutoML. There are several possible choices for the pipeline search algorithm, such as Bayesian optimization [7, 38] or bandit-based approaches [39]. A key difference in this work is the extension

Pipeline Stage	Algorithm (No. Hyperparameters Optimized by AutoPrognosis)				
Imputation	HyperImpute	Mean (0)	Median (0)	Most-Frequent (0)	MissForest (2)
	(M)ICE (0)	SoftImpute (2)	EM (1)	Sinkhorn (6)	None (0)
Dimensionality Reduction	Fast ICA (1)	Feat. Agg. (1)	Gauss. Rand. Proj. (1)	PCA (1)	Var. Thresh. (0)
Feature Scaling	L2 Norm. (0)	Max (0)	MinMax (0)	Normal Trans. (0)	Quant. Trans. (0)
	Unif. Trans. (0)	None (0)			
Classification	ADABOOST (3)	Bagging (4)	Bernoulli NB (1)	CatBoost (2)	Decision Tree (1)
	ExtraTree (1)	Gauss. NB (0)	Grad. Boost. (3)	Hist. Grad. Boost. (2)	KNN (4)
	LDA (0)	Light GBM (6)	Linear SVM (1)	Log. Reg. (4)	Multi. NB (1)
	Neural Net. (6)	Perceptron (2)	QDA (0)	Random Forest (5)	Ridge Class. (1)
	TabNet (8)	XGBoost (11)			
Regression	Bayesian RR (1)	CatBoost (2)	Linear (0)	MLP (0)	Neural Net. (6)
	TabNet (8)	XGBoost (2)			
Survival Analysis	Cox PH (2)	CoxNet (6)	DeepHit (7)	LogLogistic AFT (1)	LogNorm. AFT (2)
	Surv. XGB (4)	Weibull AFT (2)			
Interpretability	INVASe	KernelSHAP	LIME	Effect Size	Shap Permutation
	SimplEx	Symb. Persuit			

Table 2: List of algorithms currently included in AutoPrognosis 2.0, grouped by pipeline stage. Numbers in brackets correspond to the number of hyperparameters optimized over by AutoPrognosis. AutoPrognosis is readily extendable to additional methods, algorithms, and hyperparameters.

of such approaches beyond hyperparameter optimization, the typical use of AutoML, to accommodate more general configuration spaces that encompass ML pipelines. AutoPrognosis is flexible to the choice of AutoML search algorithm and can be extended as new approaches are developed. Currently, our default approach is based on Bayesian optimization. In Table 2, we provide a list of the algorithms currently implemented in AutoPrognosis 2.0, together with the number of hyperparameters optimized over for each method. We emphasize the extendability of our approach to new methods, algorithms, and hyperparameters.

Feature processing. While imputation ensures data is complete, preprocessing datasets is a common requirement for many ML estimators. In particular, feature scaling to normalize the range or the shape of features can significantly affect performance [40]. AutoPrognosis can optimize over five dimensionality reduction and six feature scaling algorithms.

Model selection and fitting. Next, a model and hyperparameters must be selected. This is a key step as suboptimal choice of model or hyperparameters can significantly affect the performance of the resulting ML system. AutoPrognosis contains 22 classification algorithms, seven regression algorithms, and seven methods for survival analysis. Together with a range of hyperparameters, this defines a broad algorithmic search space. While navigating

this space manually by hand is extremely challenging, AutoPrognosis learns relationships between different settings to efficiently arrive at an optimized solution. Finally, AutoPrognosis combines the best performing models into a single ensemble using the posterior belief of the AutoML algorithm.

Model explanations

Predictive models alone are not sufficient and more must be done to engender model trust from both clinical users [5] and regulatory bodies [26, 27, 28]. Consequently, AutoPrognosis contains a suite of methods for explaining ML models. We have included feature-based interpretability methods, such as SHAP [41], that allow us to understand the importance of individual features, as well as an example-based interpretability method, Simplex [42], that explains the model output for a particular sample with examples of similar instances, similar to case-based reasoning. Furthermore, sometimes outputs of a specific form are required, such as explicit risk equations [28]. We have therefore included the ability to convert optimized models into transparent risk equations using symbolic regression [43].

Demonstrators

In order for risk scores to be useful, they need to be readily available to clinical practitioners. To facilitate this, AutoPrognosis allows interactive demonstrators to be produced for clinical use. We build our clinical demonstrators on top of the open-source Streamlit package [44]. Compared to traditional solutions, these require almost no technical capabilities to set up, and the standardized nature simplifies adoption for end-users.

4. Illustrative application of AutoPrognosis 2.0

In this section, we show how AutoPrognosis 2.0 can be applied to address the challenges described in Section 2. We demonstrate the application of AutoPrognosis 2.0 using an illustrative scenario: prognostic risk prediction of developing diabetes using a cohort of 502,467 individuals from UK Biobank. Our goal is *not* to develop the best model for diabetes risk prediction possible, but instead to exemplify how our tool can be used.

In our use-scenario, we show that the model derived with AutoPrognosis outperforms risk models currently used in clinical practice and quantify the benefit of ML methods over Cox proportional hazard models. In addition, we show how the model interpretability components of AutoPrognosis can be

used to understand the drivers of predictions and identify novel risk factors not incorporated into previous risk scores. Finally, we use AutoPrognosis to share the diabetes risk score as a web-based decision support tool which can be publicly accessed by patients and clinicians worldwide.²

While we illustrate risk prediction of developing diabetes using a cohort from UK Biobank, AutoPrognosis can be applied to construct diagnostic and prognostic models for any disease or clinical outcome. Furthermore, AutoPrognosis is applicable to classification and regression tasks, in addition to survival analysis.

4.1. Designing experiments

Selecting which dataset to use AutoPrognosis can be used with data from many different origins, such as biobanks [8], registries [9, 10], and private hospital data [13]. Here, we use the UK Biobank, due to its availability and popularity as a resource for healthcare researchers. UK Biobank enrolled half a million participants from 22 assessment centers across England, Wales, and Scotland between 2006 and 2010 [45], with follow-up data collected from hospital records [46]. From UK Biobank, we extracted a cohort of participants who were 40 years of age or older with no diagnosis or history of diabetes at baseline; the primary outcome was diagnosis of diabetes within a 10 year horizon. We selected diabetes as our outcome of interest due to its global prevalence and role as a risk factor for a multitude of other indications [47].

Selecting variables Variables can be selected for inclusion in a study in a myriad of ways. Often, healthcare professionals will select a subset of exploratory features that are of particular interest to them. This could be due to supporting medical literature, to explore a hypothesis, or based on features included in existing risk scores. Alternatively, we can always chose to initially include all available variables. Here, we selected an initial set of 109 exploratory features based on their general clinical availability, discussions with clinicians, and features used by existing risk scores. We purposefully selected almost an order of magnitude increase compared to existing risk scores to illustrate how AutoPrognosis can be used in such a scenario.

Selecting benchmarks Often, existing risk scores will exist for the outcome of interest; this is certainly true for diabetes, where several risk scores

²<https://autoprognosis-biobank-diabetes.streamlitapp.com/>

that estimate the probability of developing diabetes are currently used in clinical practice. Therefore, we use the following as baseline risk scores:

- **ADA:** The American Diabetes Association (ADA) risk score [48] is a points-based score employing six features, namely age, sex, family history of diabetes, history of hypertension, obesity, and physical activity.
- **FINRISK:** A risk score for diabetes was derived from FINRISK, a large population survey in Finland, based on age, body mass index (BMI), waist circumference, history of antihypertensive drug treatment and high blood glucose, physical activity, and daily consumption of fruits, berries, or vegetables [49].
- **DiabetesUK:** The risk score from Diabetes UK uses seven features: gender, age, ethnicity, family history, waist size, BMI, and high blood pressure requiring treatment.
- **QDiabetes:** Finally, QDiabetes [50] consists of three separate models depending on the clinical information available and stage of risk screening. Model A uses 16 non-laboratory features that do not require a blood test and is intended primarily as an initial screening tool. Models B and C include the same variables as Model A together with fasting blood glucose and hemoglobin A1c (HbA1c), respectively, with the aim of refining risk assessment following a blood test.

In addition to the baseline risk scores, a comparison with traditional modeling approaches can be made using AutoPrognosis. We demonstrate this by fitting Cox proportional hazard (Cox PH) [19] models using the same features as each of the baseline risk scores. These models can be thought of as variants of the respective risk scores calibrated to the specific dataset.

4.2. Using AutoPrognosis 2.0 to address the challenges of diagnostic and prognostic modeling

Through the lens of our example (diabetes risk prediction), we demonstrate how AutoPrognosis 2.0 can be used to address the challenges of diagnostic and prognostic modeling introduced in Section 2.

Challenge 1. Developing powerful ML pipelines

We begin by using AutoPrognosis to derive a clinical risk score for diabetes. We evaluate the performance of the models using concordance index

Table 3: Diabetes risk prediction results. The risk scores automatically derived by AutoPrognosis outperform the existing risk scores and Cox PH models retrained on the same features. Mean performance reported with 95% confidence interval.

Method	C-index \uparrow	Brier score \downarrow	AUROC \uparrow
ADA	0.696 ± 0.015	0.011 ± 0.000	0.697 ± 0.018
FINRISK	0.728 ± 0.029	0.019 ± 0.000	0.729 ± 0.020
DIABETESUK	0.759 ± 0.013	0.016 ± 0.000	0.759 ± 0.019
QDIABETES MODEL A	0.794 ± 0.022	0.008 ± 0.000	0.795 ± 0.017
QDIABETES MODEL B	0.788 ± 0.019	0.015 ± 0.000	0.788 ± 0.013
QDIABETES MODEL C	0.839 ± 0.021	0.005 ± 0.000	0.840 ± 0.010
Cox PH (ADA)	0.774 ± 0.027	0.002 ± 0.000	0.774 ± 0.020
Cox PH (FINRISK)	0.786 ± 0.023	0.002 ± 0.000	0.786 ± 0.026
Cox PH (DIABETESUK)	0.794 ± 0.023	0.002 ± 0.000	0.794 ± 0.022
Cox PH (QDIABETES C)	0.858 ± 0.007	0.002 ± 0.000	0.860 ± 0.018
AutoPrognosis 2.0	0.888 ± 0.007	0.002 ± 0.000	0.888 ± 0.012
AUTOPROGNOSIS (19 FEAT.)	0.870 ± 0.011	0.002 ± 0.000	0.867 ± 0.020

(C-index) to assess model discrimination, Brier score to assess calibration, and the area under the receiver-operating curve (AUROC) to assess prediction accuracy. We perform imputation five times and then conduct 3-fold cross validation for each of the imputed datasets. As seen in Table 3, the risk score developed by AutoPrognosis significantly outperforms all baseline risk scores and Cox PH models (p-value < 0.001), achieving a C-index on the validation cohort of 0.888 (95% confidence interval: 0.881-0.895). This compares to 0.696 (0.681-0.711) for the ADA score, 0.728 (0.699-0.757) for FINRISK, 0.759 (0.746-0.772) for DiabetesUK, and 0.839 (0.818-0.860) for the best performing QDiabetes model (Model C). Cox PH models fit with the same risk factors as the clinical risk scores achieved improved performance (C-indices: 0.774, 0.786, 0.794, and 0.858, respectively), but exhibit lower performance than AutoPrognosis.

As an alternate way of understanding the clinical impact of our results, we performed decision curve analysis and calculated the clinical net benefit across a range of risk threshold probabilities. We compared the predicted risk by AutoPrognosis with the QDiabetes models, the best performing of

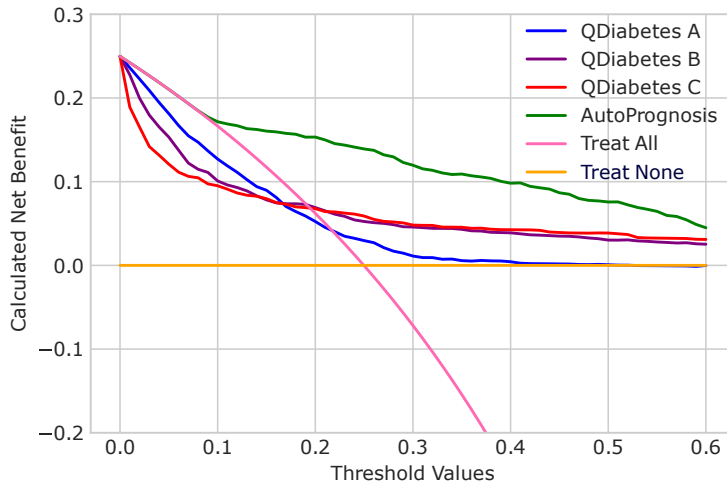


Figure 2: Decision curve analysis. AutoPrognosis exhibits higher net benefit at all decision thresholds compared to existing risk scores and baseline treatment plans.

the existing clinical risk scores, as well as baseline strategies to treat all patients (Treat All) or no-one (Treat None). Decision curve analysis further demonstrates the benefit of AutoPrognosis compared to existing risk scores for diabetes. At all decision thresholds, AutoPrognosis offers greater net benefit and is the only score to outperform “Treat All” between the 0.1 and 0.2 thresholds, and the only model to perform in-line with “Treat All” below a threshold of 0.1.

Challenge 2. Understanding when ML is necessary and its value

Table 3 demonstrates the benefit of AutoPrognosis compared to existing risk scores and Cox PH models retrained on the same features. We now directly compare AutoPrognosis to Cox PH models on the same training data to understand if ML is needed for this problem. In Table 4, we show the performance of AutoPrognosis and a Cox PH model using the full feature set considered. We see that while some of the benefit is due to the additional features, there remains value in the improved modeling approach, even for identical feature sets (p-value: 0.005).

Table 4: Quantifying the value of ML. The risk score automatically derived by AutoPrognosis significantly outperforms a Cox PH model trained on the same features (p-value: 0.005).

Method	C-index \uparrow
<i>All Variables</i>	
Cox PH	0.883 ± 0.010
AUTOPROGNOSIS	0.888 ± 0.007

Challenge 3. Determining the value of information

Understanding the predictive power of variables is key and often there is a trade-off (e.g. cost or time) in clinical practice to acquiring additional variables. We evaluate AutoPrognosis using different subsets of features. We selected features using the magnitude of the effect size. We measure the distributional shift for an increase in predicted risk using Cohen’s D [51], and select features with effect sizes exceeding the thresholds {0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. Even using only eight features, AutoPrognosis slightly outperforms the best performing existing risk score, QDiabetes Model C (Figure 3). As the number of features increases, performance rapidly increases until 35 features are used (effect size: 0.5). After this point, while there is some gain from additional features, it could be considered marginal given the number of additional features employed.

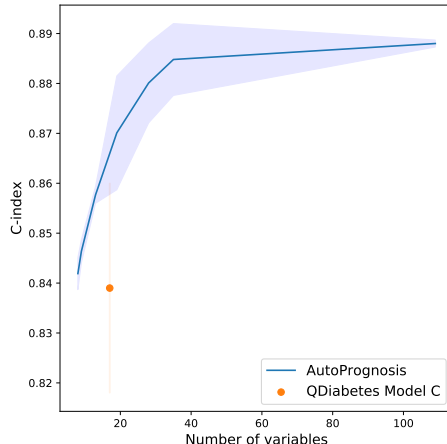


Figure 3: Value of information. We evaluate AutoPrognosis using different numbers of features, corresponding to different effect size thresholds. The feature efficiency is compared to QDiabetes Model C, the best performing existing risk score.

Challenge 4. Understanding and debugging ML models

Highly predictive models alone are insufficient and it is necessary to understand which features are important. We demonstrate how the interpretability methods incorporated in AutoPrognosis 2.0 can be used to understand how ML models make predictions and debug their behavior. We begin by examining the SHAP values [41] to explain the key contributors to model performance. Figure 4 shows the top 20 features. Encouragingly, these features are largely consistent with clinical knowledge, providing evidence that the model is acting in a desirable manner. Several of the top risk factors, such as HbA1c, waist size, and body mass index, were also included in previous risk scores. However, a number of additional features, including both laboratory and non-laboratory tests, were deemed important. A number of these features have been shown to be risk factors for diabetes (e.g. gamma-glutamyl transferase [52]), but have not been incorporated into other risk

Table 5: Performance of diabetes risk scores for subgroups defined by HbA1c.

Method	C-index		AUROC	
	HbA1c < 4.69%	HbA1c \geq 4.69%	HbA1c < 4.69%	HbA1c \geq 4.69%
QDIABETES MODEL A	0.771 \pm 0.053	0.775 \pm 0.016	0.772 \pm 0.009	0.775 \pm 0.023
QDIABETES MODEL B	0.738 \pm 0.031	0.773 \pm 0.010	0.738 \pm 0.007	0.773 \pm 0.017
QDIABETES MODEL C	0.735 \pm 0.052	0.855 \pm 0.008	0.736 \pm 0.022	0.856 \pm 0.004
AUTOPROGNOSIS 2.0	0.818 \pm 0.047	0.889 \pm 0.011	0.807 \pm 0.013	0.896 \pm 0.009

scores. Of the existing risk factors, we find that HbA1c is significantly more important to the predictions of AutoPrognosis than blood glucose, which is consistent with our earlier experiments that showed QDiabetes Model C (which uses HbA1c) outperforms Model B (which uses blood glucose) on the UK Biobank population.

Finally, several features commonly incorporated in previous risk scores are notably missing: for example age and sex. One explanation could be that UK Biobank contains a limited age range (40-69 at enrollment), and thus the role of age could be reduced over that range. However, increasingly, younger individuals are being diagnosed with diabetes [53], which could also explain the omission of age as a key risk factor. In the case of sex, while it was once assumed that there were sex differences, diabetes is equally prevalent among men and women in most populations [54].

To illustrate debugging, we consider the development of diabetes in individuals with differing HbA1c levels. We divide the overall cohort into two approximately equal parts using the median HbA1c value of 4.69%. This equates to splitting the population into a low-normal subgroup and a high-normal and elevated subgroup [55].

We evaluated AutoPrognosis and the QDiabetes models on these two

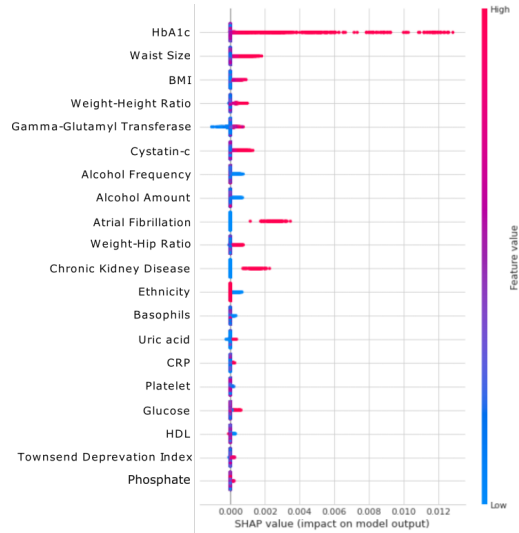


Figure 4: SHAP values for the most important features.

cohorts (Table 5). Despite displaying better performance across the entire dataset, QDiabetes Model C *underperforms* Model A for patients in the low-normal HbA1c cohort. Conversely, AutoPrognosis performs best for both subgroups, although predicting future risk of diabetes is more challenging for low-normal HbA1c patients, in line with the other models. This could suggest that QDiabetes Model C is overly reliant on HbA1c while AutoPrognosis has more accurately captured the risk factors for low HbA1c patients.

This raises the question of *why* AutoPrognosis is able to issue more accurate predictions for the low-normal HbA1c cohort, in particular given HbA1c is ranked as the most important feature globally (Figure 4). Table 6 shows the most important features (measured by risk effect size) for the two subgroups defined by HbA1c. While there is significant overlap, there are five unique features in the top 20 for each cohort. This type of analysis can help clinicians understand and debug the predictions of models not only for the entire population, but specific subgroups of interest.

Table 6: The most important features for AutoPrognosis measured by risk effect size (value in parenthesis) for the two cohorts defined by median HbA1c. Features in blue differ between the two cohorts.

HbA1c < 4.69%	HbA1c ≥ 4.69%
ATRIAL FIBRILLATION (3.0)	HbA1c (3.0)
WAIST SIZE (2.8)	GLUCOSE (2.5)
BODY MASS INDEX (2.7)	WEIGHT/HEIGHT RATIO (1.5)
WEIGHT/HEIGHT RATIO (2.7)	WAIST SIZE (1.5)
WEIGHT (2.7)	BODY MASS INDEX (1.4)
HIP SIZE (2.2)	WEIGHT (1.3)
WAIST/HIP RATIO (1.8)	WAIST/HIP RATIO (1.1)
CYSTATIN-C (1.6)	HIP SIZE (1.1)
KIDNEY DISEASE (1.5)	ALANINE TRANSAMINASE (0.87)
URIC ACID (1.3)	TRIGLYCERIDES (0.76)
ALANINE TRANSAMINASE (1.1)	GAMMA-GLUTAMYL TRANSFERASE (0.74)
ANTI-HYPERTENSIVE MEDICATION (1.1)	HDL (0.71)
HISTORY OF HYPERTENSION (0.99)	C-REACTIVE PROTEIN (0.70)
TRIGLYCERIDES (0.97)	CYSTATIN-C (0.68)
GAMMA-GLUTAMYL TRANSFERASE (0.96)	SEX HORMONE-BINDING GLOBULIN (0.67)

Challenge 5. Making ML models accessible and usable

Finally, we end our illustrative scenario with an example web-based demonstrator enabling the use of the risk model derived by AutoPrognosis. The web application can be accessed at <https://autoprognosis-biobank-diabetes>.

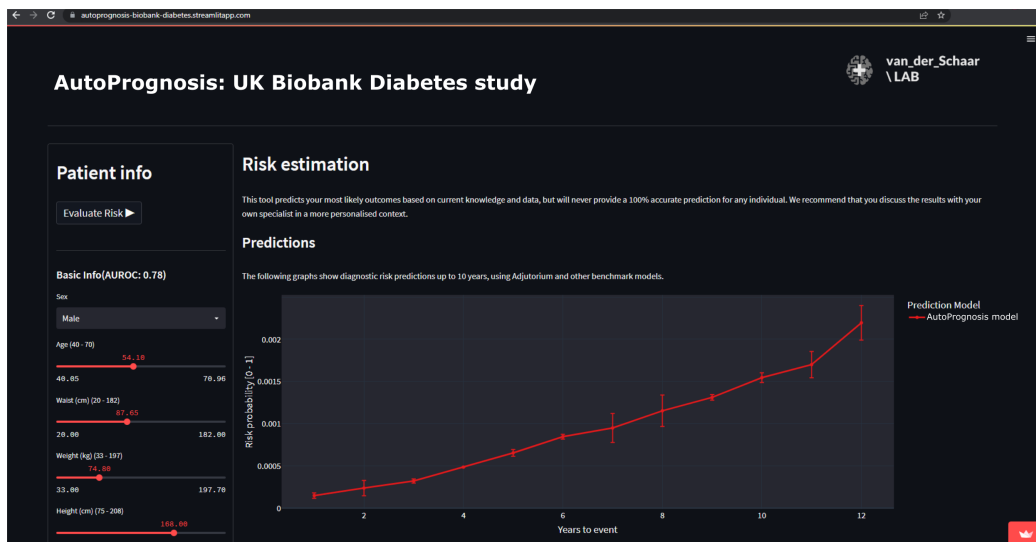


Figure 5: Screenshot of an example clinical demonstrator produced by AutoPrognosis.

streamlitapp.com/. A screenshot is provided in Figure 5.

5. Using AutoPrognosis in Healthcare and Beyond

Advances in ML algorithms harbor the potential to transform healthcare; however, major challenges continue to limit their adoption in medicine. In this work, we define these challenges and describe the first integrated, automated framework for diagnostic and prognostic modeling, AutoPrognosis 2.0, that is designed explicitly to overcome each obstacle in a way that is accessible to non-expert users, democratizing model construction, understanding, debugging, and sharing.

While we have provided an illustrative example of how AutoPrognosis can be used, the key finding reported here is *not* the performance of a single illustrative model, but rather the way in which it was built. We believe AutoPrognosis 2.0 is a necessary development in the journey towards widespread adoption of ML systems in clinical practice and hope that researchers will engage with this tool. Rather than marginalizing healthcare experts, we believe AutoPrognosis places them at the center and empowers them to create new clinical tools. As part of this journey, we will continue to add new features and improve AutoPrognosis. Finally, while the focus and motivation for AutoPrognosis is medicine, it has not escaped our notice that AutoPrognosis

can be used to construct predictive models and risk scores for applications beyond healthcare.

References

- [1] E. J. Topol, High-performance medicine: The convergence of human and artificial intelligence, *Nature Medicine* 25 (1) (2019) 44–56. doi:10.1038/s41591-018-0300-7.
- [2] S. Gerke, T. Minssen, G. Cohen, Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare, in: A. Bohr, K. Memarzadeh (Eds.), *Artificial Intelligence in Healthcare*, Academic Press, 2020, pp. 295–336. doi:<https://doi.org/10.1016/B978-0-12-818438-7.00012-5>.
- [3] T. Q. Sun, R. Medaglia, Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare, *Government Information Quarterly* 36 (2) (2019) 368–383. doi:<https://doi.org/10.1016/j.giq.2018.09.008>.
- [4] K.-H. Yu, A. L. Beam, I. S. Kohane, Artificial intelligence in healthcare, *Nature Biomedical Engineering* 2 (10) (2018) 719–731. doi:10.1038/s41551-018-0305-z.
- [5] P. Rajpurkar, O. Chen, Emmaand Banerjee, E. J. Topol, AI in health and medicine, *Nature Medicine* 28 (1) (2022) 31–38. doi:10.1038/s41591-021-01614-0.
- [6] L. Petersson, I. Larsson, J. M. Nygren, P. Nilsen, M. Neher, J. E. Reed, D. Tyskbo, P. Svedberg, Challenges to implementing artificial intelligence in healthcare: A qualitative interview study with healthcare leaders in Sweden, *BMC Health Services Research* 22 (1) (2022) 850. doi:10.1186/s12913-022-08215-8.
- [7] A. Alaa, M. van der Schaar, AutoPrognosis: Automated clinical prognostic modeling via Bayesian optimization with structured kernel learning, In: *Proceedings of the 35th International Conference on Machine Learning* 80 (2018) 139–148.

- [8] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, M. van der Schaar, Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants, *PLOS ONE* 14 (5) (2019) 1–17. doi:10.1371/journal.pone.0213653.
- [9] A. M. Alaa, M. van der Schaar, Prognostication and risk factors for cystic fibrosis via automated machine learning, *Scientific Reports* 8 (1) (2018) 11242. doi:10.1038/s41598-018-29523-2.
- [10] A. M. Alaa, D. Gurdasani, A. L. Harris, J. Rashbass, M. van der Schaar, Machine learning to guide the use of adjuvant therapies for breast cancer, *Nature Machine Intelligence* 3 (8) (2021) 716–726. doi:10.1038/s42256-021-00353-8.
- [11] H. Rahbar, D. S. Hippe, A. Alaa, S. H. Cheeney, M. van der Schaar, S. C. Partridge, C. I. Lee, The value of patient and tumor factors in predicting preoperative breast MRI outcomes, *Radiology: Imaging Cancer* 2 (4) (2020) e190099. doi:10.1148/rycan.2020190099.
- [12] Z. Qian, A. M. Alaa, M. van der Schaar, CPAS: The UK’s national machine learning-based hospital capacity planning system for COVID-19, *Machine Learning* 110 (1) (2021) 15–35. doi:10.1007/s10994-020-05921-4.
- [13] A. A. Shah, S. K. Devana, C. Lee, R. Kianian, M. van der Schaar, N. F. SooHoo, Development of a novel, potentially universal machine learning algorithm for prediction of complications after total hip arthroplasty, *The Journal of Arthroplasty* 36 (5) (2021) 1655–1662.e1. doi:https://doi.org/10.1016/j.arth.2020.12.040.
- [14] S. K. Devana, A. A. Shah, C. Lee, A. R. Roney, M. van der Schaar, N. F. SooHoo, A novel, potentially universal machine learning algorithm to predict complications in total knee arthroplasty, *Arthroplasty Today* 10 (2021) 135–143. doi:https://doi.org/10.1016/j.artd.2021.06.020.
- [15] A. A. Shah, S. K. Devana, C. Lee, A. Bugarin, E. L. Lord, A. N. Shamie, D. Y. Park, M. van der Schaar, N. F. SooHoo, Machine learning-driven identification of novel patient factors for prediction of major complications after posterior cervical spinal fusion, *European Spine Journal* 31 (8) (2022) 1952–1959. doi:10.1007/s00586-021-06961-7.

- [16] A. A. Shah, S. K. Devana, C. Lee, A. Bugarin, M. K. Hong, A. Upfill-Brown, G. Blumstein, E. L. Lord, A. N. Shamie, M. van der Schaar, N. F. SooHoo, D. Y. Park, A risk calculator for the prediction of C5 nerve root palsy after instrumented cervical fusion, *World Neurosurgery* 166 (2022) e703–e710. doi:<https://doi.org/10.1016/j.wneu.2022.07.082>.
- [17] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, D. Dennison, Hidden technical debt in machine learning systems, *Advances in neural information processing systems* 28 (2015).
- [18] B. Nestor, M. McDermott, G. Chauhan, T. Naumann, M. C. Hughes, A. Goldenberg, M. Ghassemi, Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation, *Machine Learning for Health (ML4H) Workshop at NeurIPS* (2018).
- [19] D. R. Cox, Regression models and life-tables, *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (2) (1972) 187–202. doi:<https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
- [20] O. Akbilgic, R. L. Davis, The promise of machine learning: When will it be delivered?, *Journal of Cardiac Failure* 25 (6) (2019) 484–485. doi:<https://doi.org/10.1016/j.cardfail.2019.04.006>.
- [21] M.-A. Schulz, B. T. T. Yeo, J. T. Vogelstein, J. Mourao-Miranada, J. N. Kather, K. Kording, B. Richards, D. Bzdok, Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets, *Nature Communications* 11 (1) (2020) 4238. doi:[10.1038/s41467-020-18037-z](https://doi.org/10.1038/s41467-020-18037-z).
- [22] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *Journal of machine learning research* 3 (Mar) (2003) 1157–1182.
- [23] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligent models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015) 1721–1730.
- [24] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, H. A. Haenssle,

- Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition, *JAMA Dermatology* 155 (10) (2019) 1135–1141. doi:10.1001/jamadermatol.2019.1735.
- [25] C. H. Yoon, R. Torrance, N. Scheinerman, Machine learning in medicine: Should the pursuit of enhanced interpretability be abandoned?, *Journal of Medical Ethics* 48 (9) (2022) 581–585. doi:10.1136/medethics-2020-107102.
 - [26] Food and Drug Administration and others, Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) (2019).
 - [27] M. Mourby, K. Ó Cathaoir, C. B. Collin, Transparency of machine-learning in healthcare: The GDPR & European health law, *Computer Law & Security Review* 43 (2021) 105611.
 - [28] M. W. Kattan, K. R. Hess, M. B. Amin, Y. Lu, K. G. Moons, J. E. Gershenwald, P. A. Gimotty, J. H. Guinney, S. Halabi, A. J. Lazar, A. L. Mahar, T. Patel, D. J. Sargent, M. R. Weiser, C. Compton, members of the AJCC Precision Medicine Core, American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine, *CA: A Cancer Journal for Clinicians* 66 (5) (2016) 370–374. doi:https://doi.org/10.3322/caac.21339.
 - [29] A. L. Beam, A. K. Manrai, M. Ghassemi, Challenges to the reproducibility of machine learning models in health care, *JAMA* 323 (4) (2020) 305–306. doi:10.1001/jama.2019.20866.
 - [30] R. J. LeVeque, I. M. Mitchell, V. Stodden, Reproducible research for scientific computing: Tools and strategies for changing the culture, *Computing in Science & Engineering* 14 (4) (2012) 13–17. doi:10.1109/MCSE.2012.38.
 - [31] M. Milkowski, W. M. Hensel, M. Hohol, Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail, *Journal of Computational Neuroscience* 45 (3) (2018) 163–172. doi:10.1007/s10827-018-0702-z.

- [32] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, Efficient and robust automated machine learning, *Advances in Neural Information Processing Systems* 28 (2015) 2755–2763.
- [33] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, *Journal of Statistical Software* 45 (3) (2011) 1–67. doi:10.18637/jss.v045.i03.
- [34] D. J. Stekhoven, P. Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data, *Bioinformatics* 28 (1) (2011) 112–118. doi:10.1093/bioinformatics/btr597.
- [35] D. Jarrett, B. C. Ceber, T. Liu, A. Curth, M. van der Schaar, HyperImpute: Generalized iterative imputation with automatic model selection, In: *Proceedings of the 39th International Conference on Machine Learning* 162 (2022) 9916–9937.
- [36] J. Liu, A. Gelman, J. Hill, Y.-S. Su, J. Kropko, On the stationary distribution of iterative imputations, *Biometrika* 101 (1) (2013) 155–173. doi:10.1093/biomet/ast044.
- [37] S. Van Buuren, *Flexible imputation of missing data*, CRC press, 2018.
- [38] Z. Wang, C. Li, S. Jegelka, P. Kohli, Batched high-dimensional bayesian optimization via structural kernel learning, In: *Proceedings of the 34th International Conference on Machine Learning* (2017) 3656–3664.
- [39] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, A. Talwalkar, Hyperband: A novel bandit-based approach to hyperparameter optimization, *Journal of Machine Learning Research* 18 (185) (2018) 1–52.
- [40] S. F. Crone, S. Lessmann, R. Stahlbock, The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing, *European Journal of Operational Research* 173 (3) (2006) 781–800. doi:https://doi.org/10.1016/j.ejor.2005.07.023.
- [41] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems* 30 (2017).
- [42] J. Crabbe, Z. Qian, F. Imrie, M. van der Schaar, Explaining latent representations with a corpus of examples, *Advances in Neural Information Processing Systems* 34 (2021) 12154–12166.

- [43] J. Crabbe, Y. Zhang, W. Zame, M. van der Schaar, Learning outside the black-box: The pursuit of interpretable models, *Advances in Neural Information Processing Systems* 33 (2020) 17838–17849.
- [44] Streamlit, <https://streamlit.io/>.
- [45] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, R. Collins, UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age, *PLOS Medicine* 12 (3) (2015) 1–10. doi:10.1371/journal.pmed.1001779.
- [46] L. Adamska, N. Allen, R. Flaig, C. Sudlow, M. Lay, M. Landray, Challenges of linking to routine healthcare records in UK Biobank, *Trials* 16 (2) (2015) O68. doi:10.1186/1745-6215-16-S2-068.
- [47] W. H. Organization, et al., Global report on diabetes, World Health Organization, 2016.
- [48] H. Bang, A. M. Edwards, , A. S. Bomback, C. M. Ballantyne, D. Brillon, M. A. Callahan, S. M. Teutsch, A. I. Mushlin, L. M. Kern, Development and validation of a patient self-assessment score for diabetes risk, *Annals of Internal Medicine* 151 (11) (2009) 775–783. doi:10.7326/0003-4819-151-11-200912010-00005.
- [49] J. Lindström, J. Tuomilehto, The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk, *Diabetes Care* 26 (3) (2003) 725–731.
- [50] J. Hippisley-Cox, C. Coupland, Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study, *BMJ* 359 (2017). doi:10.1136/bmj.j5019.
- [51] J. Cohen, Statistical power analysis for the behavioral sciences, Routledge, 2013.
- [52] J. Nano, T. Muka, S. Ligthart, A. Hofman, S. Darwish Murad, H. L. Janssen, O. H. Franco, A. Dehghan, Gamma-glutamyltransferase levels, prediabetes and type 2 diabetes: A Mendelian randomization study, *International Journal of Epidemiology* 46 (5) (2017) 1400–1409. doi:10.1093/ije/dyx006.

- [53] International Diabetes Federation, IDF Diabetes Atlas, 6th edn., 2013.
- [54] E. A. Gale, K. M. Gillespie, Diabetes and gender, *Diabetologia* 44 (1) (2001) 3–15. doi:10.1007/s001250051573.
- [55] American Diabetes Association, <https://diabetes.org/diabetes/a1c/diagnosis>, Last accessed: 18th August 2022.