Application of concise machine learning to construct accurate and interpretable EHR computable phenotypes

William La Cava[1]*, Paul C Lee[2]*, Imran Ajmal[2], Xiruo Ding[2], Priyanka Solanki[2], Jordana B Cohen[1], Jason H Moore[1], Daniel S Herman[2]

*[1]Department of Biostatistics, Epidemiology, and Informatics and [2]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, USA. *equal contribution*

ABSTRACT

Objective: Electronic health records (EHRs) can improve patient care by enabling systematic identification of patients for targeted decision support. But, this requires scalable learning of computable phenotypes. To this end, we developed the feature engineering automation tool (FEAT) and assessed it in targeting screening for the under-diagnosed, under-treated disease primary aldosteronism.

Materials and Methods: We selected 1199 subjects receiving longitudinal care in one health system between 2007 and 2017 and classified them for hypertension (N=608), hypertension with unexplained hypokalemia (N=172), and apparent treatment-resistant hypertension (N=176) by chart review. We derived 331 features from EHR encounters, diagnoses, laboratories, medications, vitals, and notes. We modified FEAT to encourage model parsimony and compared its models' performance and interpretability to that of expert-curated heuristics and conventional machine learning.

Results: FEAT models trained to replicate expert-curated heuristics had higher AUPRC scores than all other models ($p < 0.001$) except random forests and were smaller than all other models ($p < 1e-6$) except decision trees. FEAT models trained to predict chart review phenotypes exhibited similar AUPRC scores to penalized logistic regression while being substantially simpler than all other models ($p < 1e-6$). For treatment-resistant hypertension, FEAT learned a six-feature, clinically intuitive model that demonstrated an adjusted PPV of 0.73 and sensitivity of 0.54 in testing.

Discussion: FEAT learns computable phenotypes that approach the performance of expert-curated heuristics and conventional machine learning without sacrificing interpretability.

Conclusion: By constructing accurate and interpretable computable phenotypes at scale, FEAT has the potential to facilitate widespread, systematic clinical decision support.

Correspondence:

Daniel S. Herman, 3400 Spruce Street, Founders 7.103, Philadelphia, PA 19104

Daniel.herman2@pennmedicine.upenn.edu

215-662-6575

Keywords: phenotyping, interpretability, machine learning, feature engineering

Word count: ~4500

INTRODUCTION

The adoption of electronic health records (EHRs) is transforming medicine by aiding clinical decision making and facilitating translational research.[1,2] In order to leverage EHR data, researchers must often define rules or algorithms known as *computable phenotypes* that transparently identify patient cohorts with certain characteristics or phenotypes of interest.[3–5] While there have been significant advances in creating and standardizing computable phenotypes for various conditions, developing accurate computable phenotypes remains a time-consuming and challenging process due to the heterogeneity, imprecision, and high dimensionality of EHR data.[1,2,6–8]

Various rule-based and machine learning (ML) approaches have been developed for generating computable phenotypes.[7] Due to the challenges of learning from messy, high-dimensional, mixed-type data that constitutes EHRs, many recent studies have focused on training large, complex models using ensemble models or deep learning.[9–13] Many algorithms employed in these studies, e.g. random forests and neural networks, can perform very well in classification but often lack *interpretability*, a subjective concept that can be thought of as the extent to which a model can be understood and/or its behavior interpreted by a user.[14–18] Many have noted interpretability as a key concern in EHR-based ML models[19], particularly in biomedical applications.[20]

Despite its subjectivity, there are several reasons why interpretable phenotypes are preferable to black-box ML models.[21,22] Concise models are easier to understand and apply to existing decision-making frameworks, thus allowing clinicians to corroborate predictions. When a model's decision-making process is understood, clinicians can verify or second-guess predictions, which should lead to trust and an overall higher quality of clinical decision making. As the FDA's proposed regulatory framework for the evaluation of automated clinical decision support systems requires clinicians to "independently review the basis for such

recommendations," interpretability will be an important factor in determining the regulatory requirements for future ML deployments.[23] In addition, simpler, transparent models may be more easily adjusted as clinical practices change or models are applied in new practice settings.

In this paper, we have improved and then applied the feature engineering automation tool (FEAT) to generate computable phenotypes that are both accurate and interpretable.[24–26] FEAT constructs fully interpretable feature representations, encoded as networks, in tandem with fitting a classification model. The representations are evolved using a population-based Pareto optimization algorithm that jointly optimizes model discrimination and complexity.[27,28] This approach to model training significantly reduces the feature space while achieving full model *simulatability*, i.e. the ability of a human user to follow the full decision process of the model.

We have applied FEAT to learn a computable phenotype for identifying patients that should be screened for primary aldosteronism (PA), the most frequent cause of secondary hypertension.[29] Epidemiological studies suggest that PA affects ~1% of US adults, but recent literature demonstrates it is under-screened and under-diagnosed.[30–34] Thus, identifying patients who should be screened for PA could improve their care. Using FEAT, we have developed computable phenotypes to identify patients for whom clinical specialty guidelines recommend PA screening[30]. We observe that phenotypes constructed using FEAT are on average significantly less complex than those learned by conventional ML methods but achieve similar discriminative performance.

In the following section, we discuss related work on interpretability in more detail to motivate our interest in developing and applying FEAT to learning computable phenotypes. We then describe our data collection, the FEAT method, and the experimental design. Finally, we present performance comparisons and interpret the models produced by FEAT and other ML methods for the three phenotypes of interest, ending with a discussion of the implications and future work that follows from our analysis.

BACKGROUND AND SIGNIFICANCE

There are two overarching approaches to interpretable modeling. The first is to apply a post-hoc analysis tool to a black box model that determines empirically which factors are relevant to the model's predictions. Examples of post-hoc methods include permutation importance[35,36], LIME[37], and SHAP[38]. SHAP values in particular can be very useful for describing how a black-box model behaves under specific input conditions[39]. However, these approaches do not describe the *mechanism* by which factors result in the predictions. Furthermore, since these importance scores do not describe the behavior of the model over all input conditions, it is challenging to predict model behavior as inputs change.

The second approach to interpretable modeling is to focus on learning concise models that are self-explanatory. As Lundberg et al. put it, "the best explanation of a simple model is the model itself."[38] The most commonly used method in this category is logistic regression, often employed with regularization approaches, such as the lasso and ridge regression.[23,24] Decision trees and Bayesian rule lists can generate interpretable models when constrained to small tree depths and low rule count, respectively.[20,25] Yet these approaches are limited in that smaller models may not represent complex data trends, and larger models are often uninterpretable.[17,26] In regularized regression and pruned decision trees, the trade-off between simplicity and explanatory power is left to be tuned by the user. More sophisticated strategies can characterize the trade-off space between model complexity and model accuracy, such as Pareto optimization with symbolic regression.[28] *Symbolic regression* is a method of learning the functional form and parameters of a model using a randomized, heuristic search process such as evolutionary computation.[40] *Pareto optimization* refers to a multi-objective optimization process in which preference relations between models are determined by their closeness to the "Pareto front", which is a set of points that represent the best observed trade-offs between objectives. Symbolic regression with Pareto optimization has been used to develop simple models in other

domains, such as physics[41], biology[42], fluid dynamics[43], and wind energy.[44] To our knowledge, this is the first work to explore the application of symbolic regression with Pareto optimization to EHR phenotyping.

MATERIALS AND METHODS

In the first part of this section, we describe the development of the cohort used to generate the computable phenotypes. In the second part, we describe the FEAT algorithm. We introduce the methodological changes made with the intention of promoting conciseness in the resulting models. In the final part, we describe the empirical studies that validated our methodological updates to FEAT as well as our study of computable phenotypes targeting patients that meet guideline-based criteria for being screened for secondary hypertension.

Benchmark Data

To benchmark changes to FEAT, we applied variant methods to 20 classification tasks in the Penn Machine Learning Benchmark (PMLB)[45], described in the Supplemental Material.

Patients

We studied 1200 patients receiving longitudinal primary care in the University of Pennsylvania Healthcare System (UPHS). Patients included had at least five outpatient visits in at least three separate years between 2007 and 2017, at least two encounters at one of 40 primary care practice sites, and were 18 years or older in 2018. A set of 1000 random patients from this cohort were divided into 800 for model training and 200 for model testing. One subject in the random training set was excluded because of a mid-study change in enterprise master patient index (EMPI) identifier.

For each subject, a study physician (I.A.) reviewed clinical charts and classified patients with three phenotypes of increasing complexity for hypertension related to screening guidelines for PA: (A) hypertension, (B) hypertension with unexplained hypokalemia (HTN-hk), and (C)

apparent treatment-resistant hypertension (aTRH). Classification was based on JNC7 Guidelines on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure[46]. Unclear cases were further reviewed by an additional study physician (D.S.H. or J.C.). See Supplemental Material for further details.

Preliminary expert-curated and final expert-curated heuristics for aTRH and HTN-hk (see below) were used to identify an additional 50 subjects each for model training and model testing, respectively. This yielded a total of 899 subjects for the training set and 300 subjects in the testing set. This study protocol was reviewed and approved by University of Pennsylvania IRB (#827260).

Clinical Data

We extracted 331 features from EHR clinical data repository Penn Data Store and EPIC Clarity reporting database. Demographic and encounter features included age, race, sex, categorized distance from zip code 19104, weight, BMI, blood pressures, and number of elevated blood pressures. Longitudinal features were aggregated as min, max, median, standard deviation, and skewness. The 34 most common laboratory test results (complete metabolic panel, complete blood count with differential, lipids, TSH, and hemoglobin A1c) with < 33% missingness were summarized as min, max, median, 1st quartile, and 3rd quartile. Diagnosis codes for hypertension, associated comorbidities, and indications for anti-hypertensives were aggregated and summarized as median per year and sum. Medication prescriptions were summarized as the number of days prescribed for each antihypertensive class; count of encounters while prescribed 1, 2, 3, or 4 or more medications were summarized as sum, median, standard deviation, and skewness, as well as the sum of elevated blood pressures at those encounters. Regular expressions were applied to clinical notes to identify mentions of ‘hypertension’ and variants thereof, summarized as counts. Features with values outside of physiologically

reasonable ranges, less than 5% non-zero counts, or variance less than 0.05 were excluded. Missing values were median-imputed. The full data dictionary is available as Supplemental Data.

Construction of expert-curated heuristics

Next, computable phenotypes (heuristics) were manually curated for the three target phenotypes by expert review of EHR data and several iterations of proposing, applying, and evaluating the heuristics. Heuristics were initially developed from a set of random training patients. A preliminary set of heuristics for HTN-hk and aTRH were used to identify 50 patients, and iteratively evaluated and updated. Thus, final heuristics were developed from the entire set of 799 random and 100 targeted training patients. They were then used to flag an additional 100 patients for the held-out testing set.

The heuristic designed for hypertension queried for a history of two or more diagnosis codes for hypertension (ICD-9: 401.*, 405.*; ICD-10: I10.*, I15.*). For HTN-hk, we labeled patients with at least two diagnosis codes for hypokalemia (ICD-9: 276.8; ICD-10: E87.6); or at least two outpatient encounters with low potassium measurement (< 3.6 mmol/L); or at least two prescriptions for an oral potassium supplement. For aTRH, we labeled patients (1) with documentation of at least 2 out of 5 consecutive outpatient encounters with elevated blood pressure (systolic blood pressure >= 140 mmHg or diastolic blood pressure >= 90 mmHg) while on antihypertensive medications from 3 distinct classes for at least 30 days prior to the elevated blood pressures or (2) prescribed four or more antihypertensive drug classes for at least 30 days. Exclusion criteria for aTRH included patients with diagnosis codes for heart failure or transplant (ICD-9: 428.*, V42.1; ICD-10: 150.*, Z94.1) or moderate to severe chronic kidney disease (estimated Glomerular Filtration Rate (MDRD) < 45 mL/min/1.73 m^2) prior to meeting the above criteria.

Feature Engineering Automation Tool (FEAT)

We adapted a recent method for learning informative feature representations called FEAT for automated clinical phenotyping (https://lacava.github.io/feat).[24–26] For this task, we are interested in learning a classification model from a set of *N* paired samples, $\{(\mathrm{y_i}, \mathbf{x_i}), \mathrm{i} = 1, \ldots, \mathrm{N}\}$, with binary labels $y \in \{0,1\}$ and attributes $\boldsymbol{x} \in \boldsymbol{R}^d$. FEAT attempts to learn a set of features for a logistic regression model of the form

$$logit(P(\hat{y}(\boldsymbol{x}) = 1)) = \boldsymbol{\beta}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{\phi}(\boldsymbol{x})$ is a *p*-dimensional vector of transformations of $\boldsymbol{x}$ learned from FEAT's optimization process. The coefficients $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]$ are associated with each of these transformed features. An overview of the FEAT algorithm is given in Fig. 1.

For the purposes of this study, we restricted the transformation operators to boolean functions: <, >, AND, OR, NOT. This means that FEAT only searches the space of representations consisting of these operators and the input features. We include operators that use Gini impurity to choose the split threshold for each feature in an equivalent way to classification and regression trees. Note that because the optimization process includes mutation to or insertion of new input features, it allows for non-greedy search to occur to find the best fit for the problem at hand, in contrast to decision trees.

To encourage model parsimony, we modified FEAT in two distinct ways. First, to handle high-dimensional data, rather than fitting a multivariate linear model to all the data at the start of optimization, we sampled the input data based on univariate logistic regression coefficients. Second, we added a post-run simplification procedure to shrink the final feature representation without significantly altering its behavior. This post-run simplification procedure consists of 1) explicitly removing redundant serial logical operators, 2) adaptively pruning highly correlated components of representations, and 3) applying random deletion mutations to the features in a

hill-climbing fashion. For a detailed description of these changes and a benchmark validation of their effectiveness, see the Supplemental Material.

Comparator Methods

To assess how FEAT compares to conventional ML models, we tested five supervised classifiers: LASSO-penalized logistic regression (L1 LR), ridge-penalized logistic regression (LR L2), decision tree (DT), random forest (RF), and Gaussian Naïve Bayes (GNB). Hyperparameters for each of the models were optimized using 5-fold nested cross-validation. All of the comparator methods were implemented using Scikit-learn[47]. We report the mean test area under the precision-recall curve (AUPRC) and area under the receiver-operating curve (AUROC) for all experiments. AUPRC is calculated as average precision (see *sklearn.metrics.average_precision_score*, scikit-learn version 0.23.2). In addition to comparing discrimination performance, we compare the size of the final models. For the tree-based methods (FEAT, decision tree, and random forest), we define the size of the models to be the total number of nodes in the trees. For the linear methods and GNB, we define the size to be the number of predictors with non-zero coefficients in the final model. Model performance and model sizes were compared using paired Wilcoxon rank-sum tests. Model thresholds were selected in training set to achieve a positive-predictive value (PPV) in the longitudinal, primary care cohort of 0.70. Study code is available in this repository: https://bitbucket.org/hermanlab/ehr_feat/.

RESULTS

Development of automated phenotyping method

To automatically construct computable phenotypes whose outputs are directly interpretable by clinical practitioners, we extended FEAT to better implement boolean logic, added procedures to

encourage model parsimony, and developed approaches for improving robustness. To evaluate these modifications, we applied them to a set of benchmark datasets[45] that were similar in shape to our EHR dataset. We found that restricting operators and simplifying models did not significantly impair classification performance but did decrease the size of resulting models considerably (Supp. Fig. 1; p = 7.2e-9). Detailed results are available in the Supplemental Material.

Recreating manually curated computable phenotypes

We next applied our optimized FEAT method to learn models to recapitulate the expert-curated heuristics for hypertension, HTN-hk, and aTRH from a training set of 899 subjects. For each heuristic, we ran 50 trials of 5-fold cross-validation on shuffled training datasets and seeds and averaged test scores across folds (Fig. 2, top row; Table 1). Across all three heuristics, FEAT models achieved higher AUPRC scores ($p < 0.001$; Supplemental Fig. 4) than all other models except RF. FEAT models were significantly smaller than all other models ($p < 1\text{x}10^{-6}$) except decision trees.

Automated learning of computable phenotypes

Next, we compared the performance of models trained to predict the chart-review phenotypes (Fig. 2, bottom; Table 1), which were present in 423 (47%), 93 (10%), and 103 (11%) subjects, respectively. Across all phenotypes, FEAT models achieved AUPRC scores that were higher than GNB, LR L2, and DT models ($p < 0.001$; Supplemental Fig. 4), comparable to LR L1 models ($p > 0.99$), and lower than RF models ($p < 1e\text{-}6$). These relationships were relatively consistent across outcomes, except that FEAT models appeared to also outperform LR L1 for HTN-hk. FEAT models were significantly smaller than all other models ($p < 1e\text{-}6$); on average, models were approximately 1800 times smaller than RF models and 2.9 times smaller than LR

L1 models. We next explored the trade-off between model performance and complexity for heuristic and chart-review trained models (Fig. 3). The FEAT models clustered near the high-performance, low-complexity region of this tradeoff space (top left), indicating that they learned a relatively efficient trade-off between these two objectives.

For the most complex phenotype, aTRH, FEAT models achieved a median AUPRC score of 0.69 (interquartile range [IQR]: 0.05) at a median size of 9.8 (IQR: 1.8). These models showed reasonable discrimination across all potential decision thresholds, as depicted by PRC and ROC curves (Fig. 4). Of note, the expert-curated heuristic demonstrated superior discrimination to all ML models at its single operating point.

Assessment of Model Generalization and Clinical Utility

Next, we applied the methods refined by cross-validation to learn models from the entire training set and assessed their performance on a test set of 300 subjects, including 185 (61%), 79 (26%), and 73 (24%) subjects for each chart-review phenotype. Model performance and size (Table 2) were relatively consistent with cross-validation estimates. Most appeared to have slightly better discrimination, likely due in part to the great enrichment for cases in the testing cohort. For hypertension, HTN-hk, and aTRH, the FEAT models demonstrated AUPRC scores of 0.99, 0.96, and 0.80, and AUROC scores of 0.99, 0.98, and 0.94, respectively. For HTN-hk, these FEAT models improved upon the AUPRC of the expert-curated heuristics by 18%, and on the other two heuristics, FEAT models and expert heuristics performed similarly (within 2% of each other).

To further evaluate the utility of the resulting models, we selected an interpretive threshold from training data expected to yield a PPV of 0.7 in our target population of patients receiving longitudinal, primary care. For aTRH, we assumed a case prevalence of 7.5% based on the frequency observed in our training set and meta-analyses[48]. This resulted in the selection

of a threshold of 0.596, which demonstrated a sensitivity of 0.62 in training. Among the 200 randomly drawn test patients, this FEAT model yielded an adjusted PPV of 0.73 and sensitivity of 0.54. In comparison, the heuristic showed an adjusted PPV of 0.87 and sensitivity of 0.92. Among the 100 test patients flagged by the heuristics, the final FEAT model had a PPV of 0.87, compared to a PPV 0.83 for the expert heuristic.

Model Interpretability

Finally, we evaluated the relative interpretability of the resulting models, focusing on the models for predicting aTRH. The final FEAT model was concise and interpretable (Fig. 5). The FEAT model assigns risk according to the following factors, in order of absolute coefficient magnitudes: first, a history of more than one encounter while prescribed three or more anti-hypertensive medications ($\beta = 1.33$); second, a mean systolic blood pressure above 128.6 mmHg ($\beta = 0.95$); third, a history of low variability (standard deviation) in the number of encounters while prescribed two anti-hypertensive medications per year ($\beta = -0.52$); fourth, a history of a median of 1.25 or more encounters per year while on four or more hypertension medications ($\beta = 0.49$); fifth, more than 40 mentions of hypertension in the patient notes ($\beta = 0.42$); and sixth, a maximum total calcium greater than 10.1 mg/dL ($\beta = 0.40$). To investigate the factors underlying the maximum calcium feature, we explored its associations. We found that subjects with aTRH were in fact more likely to have elevated maximum calcium (OR=4.4, $p=4x10^{-9}$) and that these elevations were in turn associated with days prescribed thiazide diuretics (OR=1.5 per SD, $p=3x10^{-6}$) and beta-blockers (OR=1.4, $p=2x10^{-4}$).

None of the other derived models can be described in such compact, clear language. So, to compare and contrast FEAT with other methods, we calculated SHAP values[38] for the test subjects. SHAP values summarize the impact of input variables on model outputs by generating an additive feature attribution model. Positive and negative SHAP values indicate an

increase and decrease in predictions, respectively. The summary plots of SHAP values (Fig. 6A, 6C) depict the distribution of SHAP values relative to the magnitude of each input variable, with each dot representing a single test subject. The decision plots of SHAP values (Fig. 6B, D) illustrate how each feature contributes to predictions for several individual subjects.

The FEAT summary plot (Fig. 6C) reflects the simplicity of the FEAT model. For the five dichotomized features, each patient's prediction is either increased or decreased a fixed increment. The one continuous feature affects each patient distinctly, but the effect has a clear directionality, i.e. high variability in the number of encounters on two anti-hypertensive medications decreases the prediction. These simple effects translate into clear, largely self-explanatory interpretations for individual subjects as to "why" the model is calling them positive or negative (Fig. 6D). The positive-slope increases in output show that most patients predicted to be positive have 2 or more encounters on 3 anti-hypertensive medications. They also either have high mean systolic blood pressure and more than 40 mentions of hypertension in notes or multiple encounters per year while prescribed 4 or more anti-hypertensive medications.

In contrast, the LR L1 (Fig. 6A,B) and RF (Supp. Fig. 2) summary and decision plots reflect much more complicated models, in which many features confer small contributions to the prediction scores. The summary plots show the modest effect of each of the 20 features with the highest model coefficients (LR L1) or mean absolute SHAP value. The decision plots demonstrate that each patient has a distinct "reason" for a positive or negative prediction, determined by a combination of many features. In addition, there is also considerable signal from the features not depicted, as evident in the variable intercepts between each patient and the bottom model output value x-axis. Notably, for the LR L1 model many of the features depicted (e.g. minimum HDL cholesterol) are not intuitively linked to the phenotype, likely due to feature co-linearity. To address this, we also calculated LR L1 SHAP values after adjusting for feature covariance (Supp. Fig 3). After adjustment, the top features (e.g. # enc 4+ meds,

median) now match clinical intuition. That said, the resulting plots still depict a more complicated relationship between features and SHAP values and the persistence of a large number of features with small individual effects. As a result, we cannot simply explain for most subjects "why" the LR L1 or RF models are predicting them as positive or negative. For the sake of comparison, similarly accounting for co-linearity in the FEAT model reinforces the explainability of its individual subject predictions (Supp. Fig. 3C,D).

Of note, the FEAT models' interpretability does have costs. For instance, some patients were classified as positive by the model but excluded by chart review because of heart failure or chronic kidney disease (Fig. 6D). In contrast, the LR L1 model appears to learn to lower predictions based on maximum creatinine or heart failure diagnosis codes (Supp. Fig. 3A). Such features were considered in FEAT training and were included in 4 of the 10 training iterations' final models, but these models not selected by our algorithm because of their overall higher complexity.

DISCUSSION & CONCLUSION

We developed a computational method to automate the construction of EHR computable phenotypes and applied that method to find patients that should be screened for the under-diagnosed, under-treated disease primary aldosteronism. Conventional approaches to manually build accurate computable phenotypes cannot scale to the expanse of potential clinical use cases. However, by embedding the design goals for such heuristics into ML approaches, it may be possible to automate their development. The expert design of computable phenotypes applies clinical knowledge in an intuitive manner. Our goal in applying FEAT to automatically create such heuristics is to generate a reasonable symbolic model that is highly accurate and interpretable by clinical practitioners.

We have compared FEAT's ability to learn computable phenotypes to that of expert heuristic curation and standard ML approaches. The models generated by FEAT are more concise and interpretable than other ML approaches that achieve similar levels of accuracy (i.e. LR and random forests). The FEAT models matched the discriminative performance of other models across the varied tasks, except for the random forest model of the most complicated phenotype, aTRH. In this case, the FEAT performed less well than RF yet was completely interpretable.

In comparison to expert-curated heuristics, the FEAT models showed better discrimination for two phenotypes but slightly worse discrimination for aTRH. This underperformance for aTRH was not unexpected for several reasons. First, the FEAT method was not empowered to learn temporal relationships between features that enabled the expert heuristic to achieve specificity, such as including a minimum time interval between meeting anti-hypertension medication criteria and assessment for persistently elevated blood pressure. We expect that future improvements to the feature representation learning method may enable the approach to natively identify such temporal relationships from longitudinal EHR data. Second, the comparison between FEAT and the expert heuristic was biased because the heuristic was used to identify most of the affected test subjects, likely inflating its observed performance. Even beyond classification performance, we believe that FEAT generated models are more adaptable to changing data compared to expert-curated heuristics.

The model that FEAT learned to identify patients with aTRH was both accurate and understandable. Its components matched those of the expert heuristic and are consistent with clinical intuition. The model demonstrated the power of combining complementary sources of information, including medication, vitals, laboratories, and concepts from notes. Finally, it learned an unexpected, but clinically intuitive valuable rule, maximum blood calcium > 10.1 mg/dL. Anti-hypertensive medications, particularly diuretics, can dysregulate calcium

homeostasis. We suspect this rule enabled the model to identify a few affected subjects on intensive anti-hypertensive regimens that were missed by the conventional rules interrogating medication and blood pressure.

There are several possible directions for further improving FEAT. For one, the ability of FEAT to recapitulate expert-curated heuristics suggests that simpler expert heuristics, such as anchor variables[49], may be leveraged as teachers in a semi-supervised approach. This could be implemented with multi-stage learning, first to predict heuristics and then to predict chart-review. Or, expert heuristics could be encoded as syntax trees and used to seed initial runs of FEAT. To improve the aTRH phenotype, FEAT transformations should include temporal reasoning. Another limitation of this work is the non-trivial, manual feature engineering upstream of FEAT. Future work could peel back this manual feature engineering by enabling FEAT to learn from longitudinal data. Although the search space would considerably increase, there is more opportunity to learn temporal relationships. And as we apply this tool to less engineered input features and harder problems, the search space will be very large. The approach would benefit considerably from learning on top of a framework that encodes expert clinical knowledge, such as ontologies and knowledge graphs[50]. The incorporation of expert knowledge would improve search efficiency and potentially performance, while maintaining interpretability.

In summary, FEAT can effectively learn highly accurate and interpretable computable phenotypes. Further refinements to the learning framework and process should eventually allow experts to review automated computable phenotypes, rather than manually design them. We believe such tools will enable widespread implementation of computable phenotype-triggered clinical decision support.

Acknowledgments

We would like to thank Debbie Cohen for helpful discussions about caring for patients with secondary hypertension.

Competing Interests

None declared

Funding

This work was supported by Grant 2019084 from the Doris Duke Charitable Foundation and the University of Pennsylvania. W. La Cava was supported by NIH grant K99 LM012926. J.H. Moore and W. La Cava were supported by NIH grant R01 LM010098.

Reference List

1. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc*. 2013;20(e2):e206-211. doi:10.1136/amiajnl-2013-002428

2. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2013;20(1):117-121. doi:10.1136/amiajnl-2012-001145

3. Mo H, Thompson WK, Rasmussen LV, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J Am Med Inform Assoc*. 2015;22(6):1220-1230. doi:10.1093/jamia/ocv112

4. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet*. 2010;86(4):560-572. doi:10.1016/j.ajhg.2010.03.003

5. Mosley JD, Driest SLV, Larkin EK, et al. Mechanistic Phenotypes: An Aggregative Phenotyping Strategy to Identify Disease Mechanisms Using GWAS Data. *PLOS ONE*. 2013;8(12):e81503. doi:10.1371/journal.pone.0081503

6. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011;4:13. doi:10.1186/1755-8794-4-13

7. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annual Review of Biomedical Data Science*. 2018;1(1):53-68. doi:10.1146/annurev-biodatasci-080917-013315

8. Conway M, Berg RL, Carrell D, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA Annu Symp Proc*. 2011;2011:274-283.

9. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*. 2016;6:26094. doi:10.1038/srep26094

10. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. $$\backslash$mathtt ${$Deepr$}$ $: A Convolutional Net for Medical Records. *IEEE journal of biomedical and health informatics*. 2017;21(1):22–30.

11. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*. 2018;22(5):1589-1604. doi:10.1109/JBHI.2017.2767063

12. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*. 2018;1(1):18. doi:10.1038/s41746-018-0029-1

13. Harutyunyan H, Khachatrian H, Kale DC, Steeg GV, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci Data*. 2019;6(1):96. doi:10.1038/s41597-019-0103-9

14. Liaw A, Wiener M. Classification and Regression by RandomForest. *Forest*. 2001;23.

15. Specht DF. Probabilistic neural networks. *Neural Networks*. 1990;3(1):109-118. doi:10.1016/0893-6080(90)90049-Q

16. Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Association for Computing Machinery; 2018:1–18. doi:10.1145/3173574.3174156

17. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Comput Surv*. 2018;51(5):93:1–93:42. doi:10.1145/3236009

18. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:170208608 [cs, stat]*. Published online March 2, 2017. Accessed June 15, 2020. http://arxiv.org/abs/1702.08608

19. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature reviews Genetics*. 2012;13(6):395.

20. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities And Obstacles For Deep Learning In Biology And Medicine. *bioRxiv*. Published online May 28, 2017:142760. doi:10.1101/142760

21. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *PNAS*. 2019;116(44):22071-22080. doi:10.1073/pnas.1900654116

22. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*. 2019;19(1):146. doi:10.1186/s12911-019-0874-0

23. Health C for D and R. Clinical Decision Support Software. U.S. Food and Drug Administration. Published May 6, 2020. Accessed June 15, 2020. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

24. La Cava W, Moore JH. Semantic variation operators for multidimensional genetic programming. *Proceedings of the Genetic and Evolutionary Computation Conference*. Published online July 13, 2019:1056-1064. doi:10.1145/3321707.3321776

25. La Cava W, Singh TR, Taggart J, Suri S, Moore JH. Learning concise representations for regression by evolving networks of trees. *arXiv:180700981 [cs]*. Published online July 3, 2018. Accessed July 9, 2019. http://arxiv.org/abs/1807.00981

26. La Cava W, Moore JH. Learning feature spaces for regression with genetic programming. *Genet Program Evolvable Mach*. Published online March 11, 2020. doi:10.1007/s10710-020-09383-4

27. La Cava W, Helmuth T, Spector L, Moore JH. A Probabilistic and Multi-Objective Analysis of Lexicase Selection and ε-Lexicase Selection. *Evolutionary Computation*. 2019;27(3):377-402. doi:10.1162/evco_a_00224

28. Smits GF, Kotanchek M. Pareto-Front Exploitation in Symbolic Regression. In: O'Reilly U-M, Yu T, Riolo R, Worzel B, eds. *Genetic Programming Theory and Practice II*. Genetic Programming. Springer US; 2005:283-299. doi:10.1007/0-387-23254-0_17

29. Thomas RM, Ruel E, Shantavasinkul PC, Corsino L. Endocrine hypertension: An overview on the current etiopathogenesis and management options. *World J Hypertens*. 2015;5(2):14-27. doi:10.5494/wjh.v5.i2.14

30. Funder JW, Carey RM, Mantero F, et al. The Management of Primary Aldosteronism: Case Detection, Diagnosis, and Treatment: An Endocrine Society Clinical Practice Guideline. *J Clin Endocrinol Metab*. 2016;101(5):1889-1916. doi:10.1210/jc.2015-4061

31. Käyser SC, Dekkers T, Groenewoud HJ, et al. Study Heterogeneity and Estimation of Prevalence of Primary Aldosteronism: A Systematic Review and Meta-Regression Analysis. *J Clin Endocrinol Metab*. 2016;101(7):2826-2835. doi:10.1210/jc.2016-1472

32. Hannemann A, Wallaschofski H. Prevalence of primary aldosteronism in patient's cohorts and in population-based studies--a review of the current literature. *Horm Metab Res*. 2012;44(3):157-162. doi:10.1055/s-0031-1295438

33. Monticone S, Burrello J, Tizzani D, et al. Prevalence and Clinical Manifestations of Primary Aldosteronism Encountered in Primary Care Practice. *Journal of the American College of Cardiology*. 2017;69(14):1811-1820. doi:10.1016/j.jacc.2017.01.052

34. Jaffe Gilad, Gray Zachary, Krishnan Gomathi, et al. Screening Rates for Primary Aldosteronism in Resistant Hypertension. *Hypertension*. 2020;75(3):650-659. doi:10.1161/HYPERTENSIONAHA.119.14359

35. Breiman L. Random forests. *Machine learning*. 2001;45(1):5–32.

36. La Cava W, Bauer CR, Moore JH, Pendergrass SA. Interpretation of machine learning predictions for patient outcomes in electronic health records. In: AMIA 2019 Annual Symposium. AMIA; 2019. https://arxiv.org/abs/1903.12074

37. Ribeiro MT, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 2016:1135–1144.

38. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.; 2017:4765–4774. Accessed November 22, 2019. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

39. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-760. doi:10.1038/s41551-018-0304-0

40. Koza JR. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press; 1992.

41. Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. *Science*. 2009;324(5923):81–85.

42. Schmidt MD, Vallabhajosyula RR, Jenkins JW, et al. Automated refinement and inference of analytical models for metabolic networks. *Physical Biology*. 2011;8(5):055011. doi:10.1088/1478-3975/8/5/055011

43. La Cava WG. Automatic Development and Adaptation of Concise Nonlinear Models for System Identification. Published online 2016. Accessed December 6, 2016. http://scholarworks.umass.edu/dissertations_2/731/

44. La Cava W, Danai K, Spector L, Fleming P, Wright A, Lackner M. Automatic identification of wind turbine models using evolutionary multiobjective optimization. *Renewable Energy*. Published online November 2015. doi:10.1016/j.renene.2015.09.068

45. Olson RS, La Cava W, Orzechowski P, Urbanowicz RJ, Moore JH. PMLB: A Large Benchmark Suite for Machine Learning Evaluation and Comparison. *BioData Mining*. Published online 2017. https://arxiv.org/abs/1703.00512

46. Chobanian Aram V., Bakris George L., Black Henry R., et al. Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*. 2003;42(6):1206-1252. doi:10.1161/01.HYP.0000107251.49515.c2

47. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12(Oct):2825–2830.

48. Noubiap JJ, Nansseu JR, Nyaga UF, Sime PS, Francis I, Bigna JJ. Global prevalence of resistant hypertension: a meta-analysis of data from 3.2 million patients. *Heart*. 2019;105(2):98-105. doi:10.1136/heartjnl-2018-313599

49. Zhang L, Ding X, Ma Y, et al. A maximum likelihood approach to electronic health record phenotyping using positive and unlabeled patients. *J Am Med Inform Assoc*. 2020;27(1):119-126. doi:10.1093/jamia/ocz170

50. Freedman HG, Williams H, Miller MA, Birtwell D, Mowery DL, Stoeckert CJ. A novel tool for standardizing clinical data in a semantically rich model. *Journal of Biomedical Informatics: X*. 2020;8:100086. doi:10.1016/j.yjbinx.2020.100086

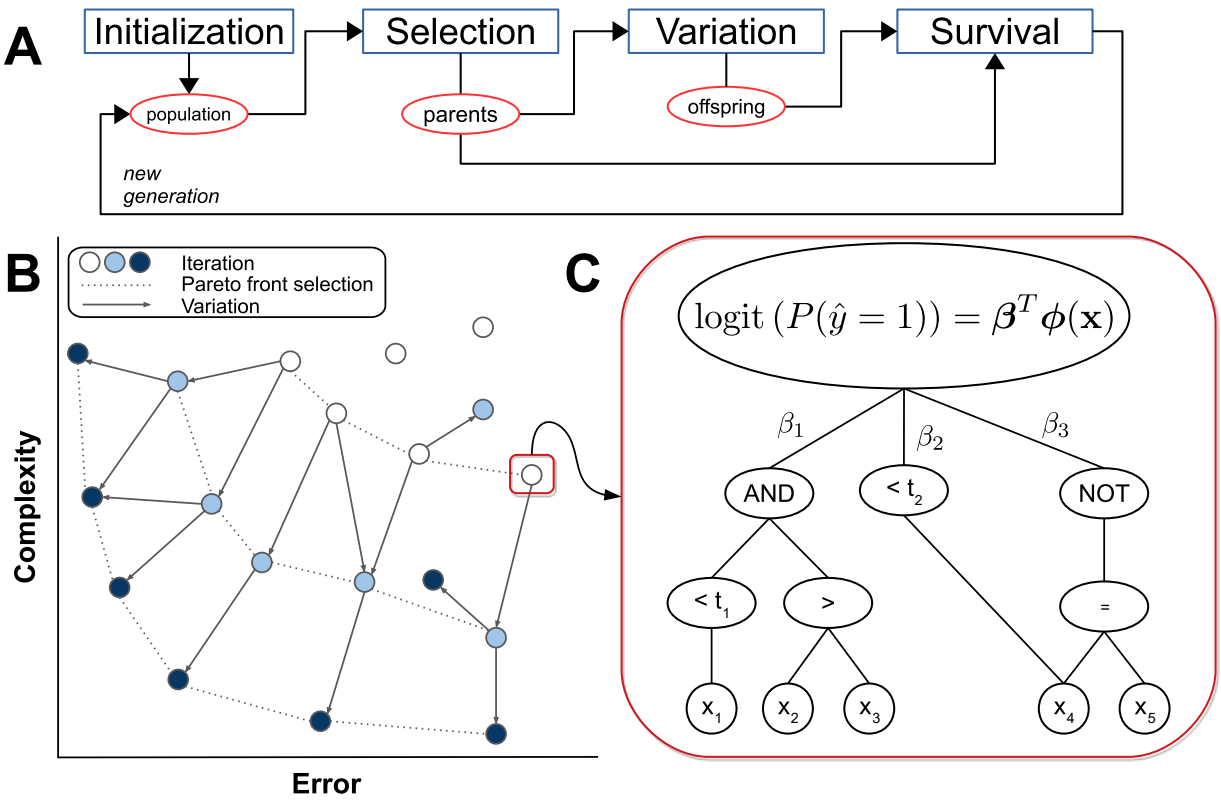

Figure 1: How FEAT works. (A) Steps in the genetic programming process. Candidate models are initialized in a population; the best models (parents) are selected via epsilon-lexicase selection; offspring are created by applying variation operations to the parents; and then parents and offspring compete in a survival step using NSGA-II [22]. The process then repeats. (B) The evaluation of a candidate models' complexity and performance in Pareto Optimization framework in the Survival step. (C) Example model in which input features are transformed by boolean functions with or without threshold operators.

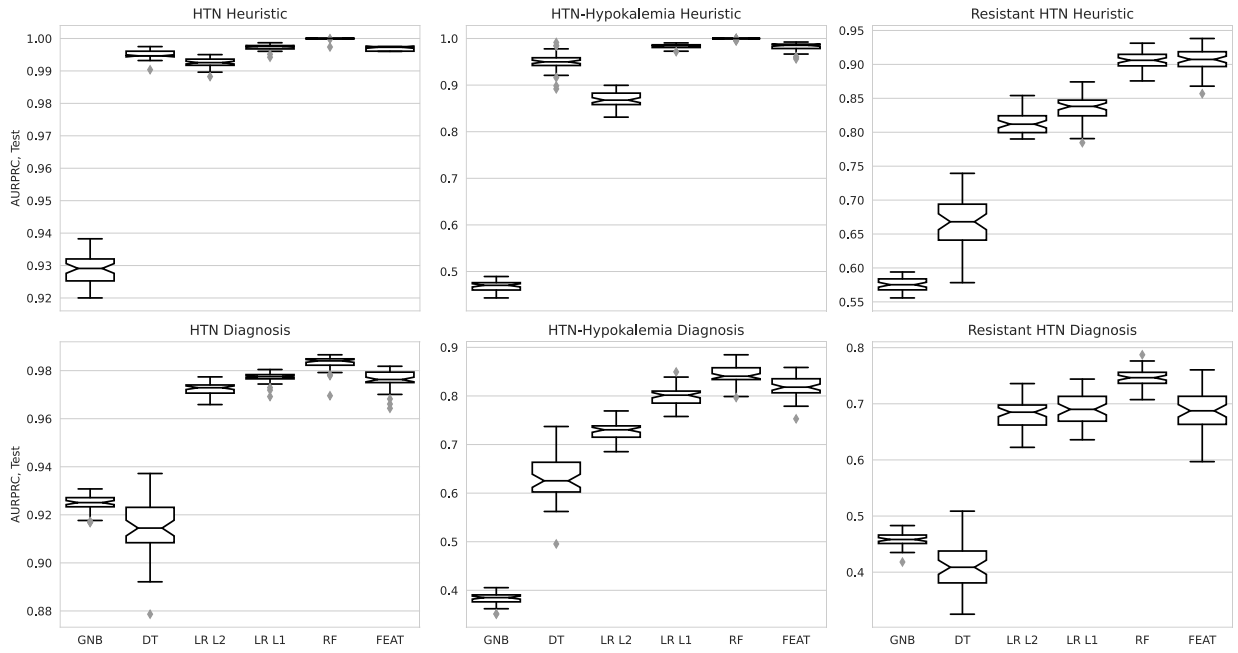

Figure 2. Estimating discriminatory power of methods by cross-validation. AUPRC scores for phenotyping models trained in 5-fold cross-validation over 50 iterations, each averaged across testing folds. Each subplot represents a different training outcome; heuristics are shown in the top row, and chart-review diagnoses are shown in the bottom row.

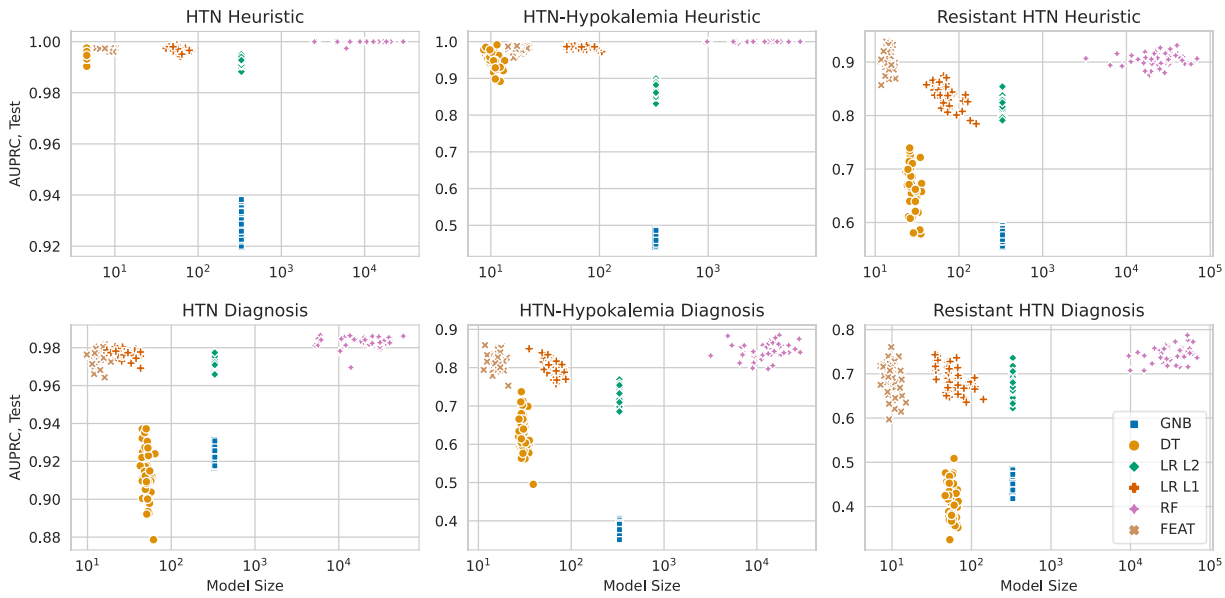

Figure 3. The tradeoff between model performance and complexity. Each point shows the cross-validation testing AUPRC (y-axis) and size (x-axis) for models trained in 50 repeat trials for each method. Each subplot represents a different expert-curated heuristic (top row) or chart review phenotype (bottom). The ideal model is discriminative and simple, meaning it is near the top left corner.

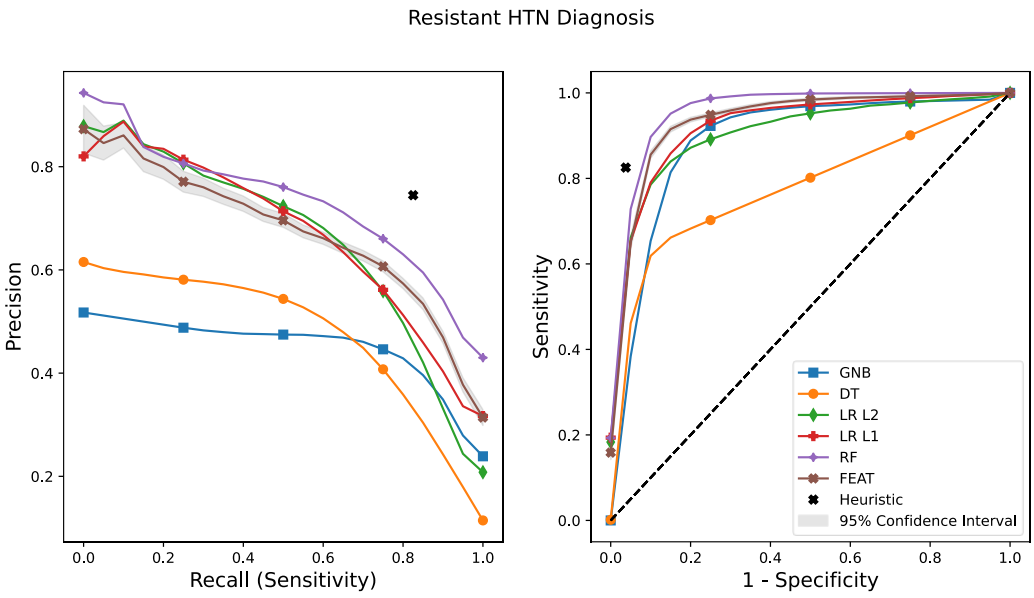

Figure 4. Model precision-recall and receiver-operating curves. Precision-recall curves (left) and receiver-operating curves (right) for phenotyping models trained to predict chart review classifications for aTRH. Values shown are means of test performance in 5-fold cross-validation iterated 50 times.

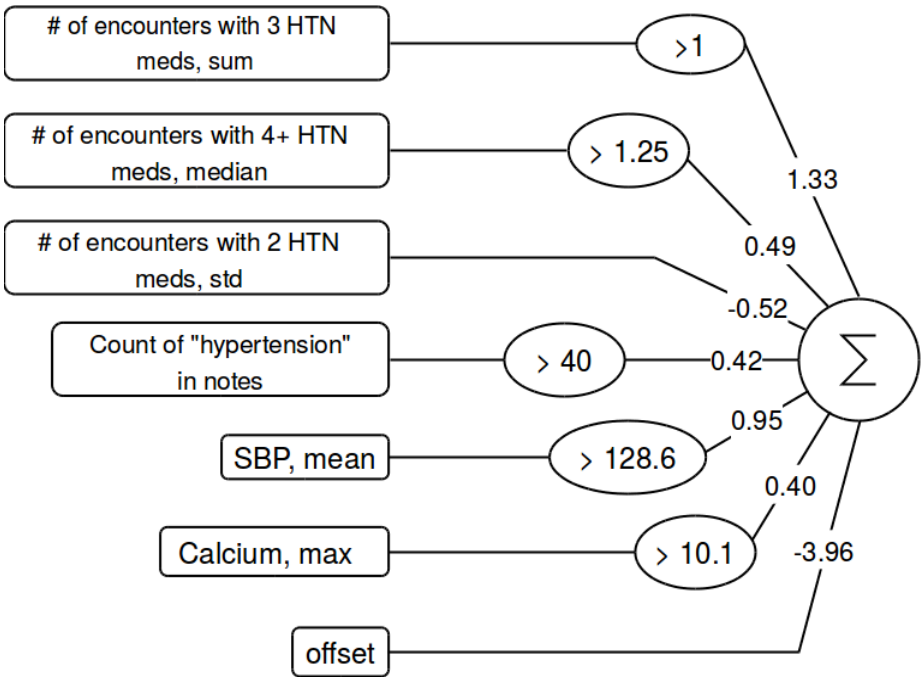

Figure 5. FEAT model trained to predict apparent treatment-resistant hypertension. The input features are shown on the left followed by the learned thresholds, the multiplication coefficients, and the summation. Note, the subsequent logit transformation and interpretive threshold is not depicted.

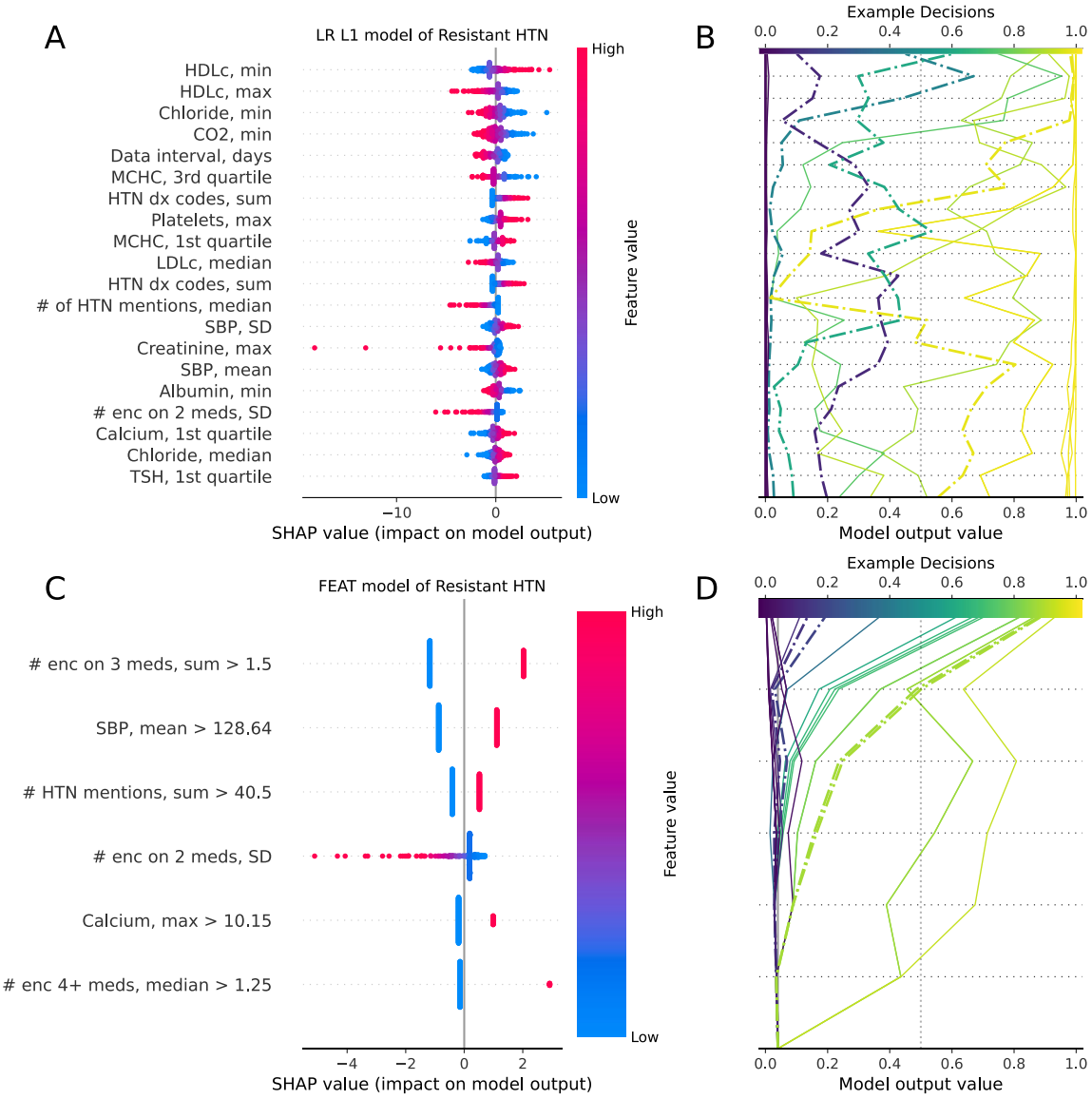

Figure 6. SHAP plots for explaining models. SHAP summary (A) and decision plots (B) for LR L1 and the summary (C) and decision (D) plots for FEAT. The summary plots (A,C) indicate the most important features, ranked by the mean absolute SHAP value for test data. The decision plots (B,D) show a sample of 10 positive and 10 negative point predictions by the models, with dash-dotted lines indicating misclassifications. The labels in the summary and decision plots identify which feature is responsible for the incremental change in the model score at each level.

Table 1. Computable phenotype discrimination and size for each target phenotype.

Phenotype	Method	Median CV AUPRC (IQR)	Median CV AUROC (IQR)	Median Size (IQR)
HTN Heuristic	GNB	0.93 (0.01)	0.96 (0.00)	331.00 (0.00)
	DT	0.99 (0.00)	1.00 (0.00)	4.60 (0.00)
	LR L1	1.00 (0.00)	1.00 (0.00)	52.30 (5.90)
	LR L2	0.99 (0.00)	0.99 (0.00)	330.80 (0.00)
	RF	1.00 (0.00)	1.00 (0.00)	8760.40 (6278.30)
	FEAT	1.00 (0.00)	1.00 (0.00)	8.20 (0.95)
HTN-Hypokalemia Heuristic	GNB	0.47 (0.02)	0.86 (0.00)	331.00 (0.00)
	DT	0.95 (0.02)	0.99 (0.01)	10.80 (1.60)
	LR L1	0.98 (0.01)	1.00 (0.00)	65.10 (18.05)
	LR L2	0.87 (0.02)	0.96 (0.01)	330.80 (0.00)
	RF	1.00 (0.00)	1.00 (0.00)	1855.80 (1640.80)
	FEAT	0.99 (0.01)	1.00 (0.00)	18.30 (2.55)
Resistant HTN Heuristic	GNB	0.58 (0.02)	0.92 (0.01)	331.00 (0.00)
	DT	0.67 (0.05)	0.89 (0.03)	27.00 (3.20)
	LR L1	0.84 (0.02)	0.95 (0.01)	70.50 (16.90)
	LR L2	0.81 (0.02)	0.94 (0.01)	330.80 (0.00)
	RF	0.91 (0.02)	0.99 (0.00)	27235.00 (13392.00)
	FEAT	0.91 (0.02)	0.98 (0.01)	14.20 (1.70)
HTN Diagnosis	GNB	0.93 (0.00)	0.95 (0.00)	331.00 (0.00)
	DT	0.91 (0.01)	0.94 (0.01)	51.00 (5.20)
	LR L1	0.98 (0.00)	0.98 (0.00)	23.20 (5.50)
	LR L2	0.97 (0.00)	0.98 (0.00)	330.80 (0.00)
	RF	0.98 (0.00)	0.99 (0.00)	20397.60 (19474.10)
	FEAT	0.98 (0.00)	0.98 (0.00)	13.90 (2.90)
HTN-Hypokalemia Diagnosis	GNB	0.38 (0.01)	0.86 (0.02)	331.00 (0.00)
	DT	0.63 (0.06)	0.88 (0.04)	30.20 (2.40)
	LR L1	0.80 (0.02)	0.94 (0.02)	68.10 (15.30)
	LR L2	0.73 (0.02)	0.91 (0.02)	330.80 (0.00)
	RF	0.84 (0.02)	0.98 (0.01)	13443.40 (8241.30)
	FEAT	0.82 (0.03)	0.97 (0.01)	16.20 (2.70)
Resistant HTN Diagnosis	GNB	0.46 (0.02)	0.90 (0.02)	331.00 (0.00)
	DT	0.41 (0.06)	0.78 (0.04)	56.80 (7.40)
	LR L1	0.69 (0.04)	0.93 (0.02)	52.90 (17.95)
	LR L2	0.69 (0.04)	0.91 (0.01)	330.80 (0.00)
	RF	0.75 (0.02)	0.96 (0.00)	38835.00 (20521.10)
	FEAT	0.69 (0.05)	0.94 (0.01)	9.80 (1.80)

Table 2: Final model performance on test set for each method, first when trained to predict heuristics, and then when trained to predict chart review phenotypes.

Phenotype	Method	Test AUPRC	Test AUROC	Size
HTN Heuristic	GNB	0.94	0.95	331
	DT	1.00	1.00	5
	LR L1	1.00	1.00	56
	LR L2	1.00	0.99	331
	RF	1.00	1.00	3106
	FEAT	1.00	1.00	12
HTN-Hypokalemia Heuristic	GNB	0.73	0.85	331
	DT	1.00	1.00	13
	LR L1	1.00	1.00	76
	LR L2	0.98	0.98	331
	RF	1.00	1.00	1106
	FEAT	1.00	1.00	24
Resistant HTN Heuristic	GNB	0.66	0.89	331
	DT	0.82	0.92	27
	LR L1	0.92	0.96	101
	LR L2	0.93	0.97	331
	RF	0.96	0.99	2648
	FEAT	0.94	0.98	16
HTN Diagnosis	GNB	0.96	0.96	331
	DT	0.97	0.97	43
	LR L1	1.00	0.99	32
	LR L2	0.99	0.98	331
	RF	1.00	0.99	67276
	FEAT	0.99	0.99	18
	Expert Heuristic	0.97	0.97	-
HTN-Hypokalemia Diagnosis	GNB	0.60	0.81	331
	DT	0.75	0.86	33
	LR L1	0.95	0.98	29
	LR L2	0.92	0.96	331
	RF	0.96	0.99	16256
	FEAT	0.96	0.98	8
	Expert Heuristic	0.81	0.96	-
Resistant HTN Diagnosis	GNB	0.57	0.86	331
	DT	0.23	0.47	67
	LR L1	0.74	0.87	130
	LR L2	0.78	0.91	331
	RF	0.89	0.96	119760
	FEAT	0.80	0.94	11
	Expert Heuristic	0.82	0.96	-