



**QUEEN'S
UNIVERSITY
BELFAST**

Machine Learning-Based Risk Assessment for Cancer Therapy-Related Cardiac Dysfunction in 4300 Longitudinal Oncology Patients

Zhou, Y., Hou, Y., Hussain, M., Brown, S-A., Budd, T., Tang, W. H. W., Abraham, J., Xu, B., Shah, C., Moudgil, R., Popovic, Z., Cho, L., Kanj, M., Watson, C., Griffin, B., Chung, M. K., Kapadia, S., Svensson, L., Collier, P., & Cheng, F. (2020). Machine Learning-Based Risk Assessment for Cancer Therapy-Related Cardiac Dysfunction in 4300 Longitudinal Oncology Patients. *Journal of the American Heart Association*, 9(23), [e019628]. <https://doi.org/10.1161/JAHA.120.019628>

Published in:

Journal of the American Heart Association

Document Version:

Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:

[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights

Copyright 2020 the authors.

This is an open access Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits use, distribution and reproduction for non-commercial purposes, provided the author and source are cited.

General rights










Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

ORIGINAL RESEARCH

Machine Learning–Based Risk Assessment for Cancer Therapy–Related Cardiac Dysfunction in 4300 Longitudinal Oncology Patients

Yadi Zhou, PhD*; Yuan Hou, PhD*; Muzna Hussain, MD*; Sherry-Ann Brown , MD, PhD; Thomas Budd, MD; W. H. Wilson Tang , MD; Jame Abraham, MD; Bo Xu , MD; Chirag Shah, MD; Rohit Moudgil, MD, PhD; Zoran Popovic , MD; Leslie Cho, MD; Mohamed Kanj, MD; Chris Watson, PhD; Brian Griffin, MD; Mina K. Chung , MD; Samir Kapadia , MD; Lars Svensson , MD, PhD; Patrick Collier , MD, PhD; Feixiong Cheng , PhD

BACKGROUND: The growing awareness of cardiovascular toxicity from cancer therapies has led to the emerging field of cardio-oncology, which centers on preventing, detecting, and treating patients with cardiac dysfunction before, during, or after cancer treatment. Early detection and prevention of cancer therapy–related cardiac dysfunction (CTRCD) play important roles in precision cardio-oncology.

METHODS AND RESULTS: This retrospective study included 4309 cancer patients between 1997 and 2018 whose laboratory tests and cardiovascular echocardiographic variables were collected from the Cleveland Clinic institutional electronic medical record database (Epic Systems). Among these patients, 1560 (36%) were diagnosed with at least 1 type of CTRCD, and 838 (19%) developed CTRCD after cancer therapy (de novo). We posited that machine learning algorithms can be implemented to predict CTRCDs in cancer patients according to clinically relevant variables. Classification models were trained and evaluated for 6 types of cardiovascular outcomes, including coronary artery disease (area under the receiver operating characteristic curve [AUROC], 0.821; 95% CI, 0.815–0.826), atrial fibrillation (AUROC, 0.787; 95% CI, 0.782–0.792), heart failure (AUROC, 0.882; 95% CI, 0.878–0.887), stroke (AUROC, 0.660; 95% CI, 0.650–0.670), myocardial infarction (AUROC, 0.807; 95% CI, 0.799–0.816), and de novo CTRCD (AUROC, 0.802; 95% CI, 0.797–0.807). Model generalizability was further confirmed using time-split data. Model inspection revealed several clinically relevant variables significantly associated with CTRCDs, including age, hypertension, glucose levels, left ventricular ejection fraction, creatinine, and aspartate aminotransferase levels.

CONCLUSIONS: This study suggests that machine learning approaches offer powerful tools for cardiac risk stratification in oncology patients by utilizing large-scale, longitudinal patient data from healthcare systems.

Key Words: anthracycline therapy ■ cancer therapy–related cardiac dysfunction ■ cardio-oncology ■ cardiotoxicity ■ echocardiography ■ machine learning

Cardiovascular disease (CVD) is the leading cause of death and the second leading cause of morbidity in cancer survivors after recurrent malignancy

in the United States.¹ Comorbidity between CVD and cancer suggests underlying shared disease pathogenesis, which can be both genetic and environmental.

Correspondence to: Patrick Collier, MD, PhD, FESC, Sydell and Arnold Miller Family Heart and Vascular Institute, Cleveland Clinic, Cleveland, OH. E-mail: collep@ccf.org or Feixiong Cheng, PhD, Lerner Research Institute, Cleveland Clinic, Cleveland, OH. E-mail: chengf@ccf.org

Supplementary Material for this article is available at <https://www.ahajournals.org/doi/suppl/10.1161/JAHA.120.019628>

*Dr Zhou, Dr Hou, and Dr Hussain contributed equally to this work.

For Sources of Funding and Disclosures, see page 12.

© 2020 The Authors. Published on behalf of the American Heart Association, Inc., by Wiley. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- This study presents the first, large-scale machine learning–based approach to evaluate complications between cancer therapies and cardiovascular diseases using cardiovascular echocardiographic and laboratory test variables from over 4300 longitudinal cancer patients.
- We developed machine learning models with high performance and verified the generalizability using time-split data to simulate real-world scenarios and found that combining both laboratory test and echocardiographic variables resulted in the highest performance.
- We identified and validated multiple clinically relevant variables associated with cancer therapy–related cardiac dysfunction using learned weight analysis of the optimal machine learning models.

What Are the Clinical Implications?

- We demonstrate the potential clinical implication of using a machine learning method to predict 6 types of cancer therapy–related cardiac dysfunction, including heart failure, atrial fibrillation, coronary artery disease, myocardial infarction, stroke, and de novo cancer therapy–related cardiac dysfunction.
- These machine learning models offer potential tools for risk assessment of cancer therapy–related cardiac dysfunction in cardio-oncology clinical practices.

Nonstandard Abbreviations and Acronyms

AUPR	area under the precision-recall curve
AUROC	area under the receiver operating characteristic curve
CTRCD	cancer therapy–related cardiac dysfunction
GB	gradient tree boosting
LR	logistic regression
ML	machine learning
RF	random forest
SMOTE	Synthetic Minority Oversampling Technique
SVM	support vector machine

One critical issue regarding environmental factors is that CVD can be associated with various treatments for cancer itself. First recognized in the 1960s,² cancer

therapy–related cardiac dysfunction (CTRCD) has been increasingly diagnosed and investigated.^{3–8} For example, a growing number of cancer survivors (>5 million) are at risk for cardiotoxicity caused by anthracycline therapy years or even decades prior for various types of cancer.⁹

Through the success of basic and translational research, cancer survivors have become one of the largest growing subsets of patients in the US health-care system.¹⁰ Currently, there are over 16.9 million cancer survivors in the United States. This number is projected to reach more than 22.1 million by 2030.¹¹ Increasing numbers of oncology patients are facing CTRCD risks as cancer survival improves. The growing awareness of cardiovascular toxicity by cancer treatment has led to the emerging field of cardio-oncology, which centers on preventing, detecting, and treating patients with cardiovascular toxicity from cancer treatment. However, precise prediction and prevention of cardiovascular toxicity in individual cancer patients or survivors has proven elusive. Further, while basic and translational research studies continue, experimental assays in animal models are limited by significant functional disparities between animal and human cardiomyocytes. Development of novel methodologies or tools, such as computational approaches, would offer unique opportunities for cardio-oncology by utilizing the accumulated longitudinal clinical data available from healthcare systems.

In recent years, machine learning (ML) has been increasingly used for cardiovascular studies, such as for the prediction of drug-induced cardiovascular complications,^{12,13} cardiac resynchronization therapy response prediction,¹⁴ risk assessment of cardiovascular events after acute myocardial infarction (MI),^{15,16} and claims data–based mortality risk predictions.¹⁷ As more longitudinal clinical data are accumulated for oncology patients, ML presents a great opportunity to use these data to build predictive models in clinical practices.^{18,19}

In this study, we hypothesized that supervised ML models could accurately predict the risk for developing several cardiovascular outcomes in cancer patients. Specifically, we applied ML models to the prediction of 6 types of cardiac outcomes, namely heart failure (HF), atrial fibrillation (AF), coronary artery disease (CAD), MI, stroke, and de novo CTRCD. We also determined several clinically relevant variables associated with these outcomes.

METHODS

All data used in this study are available from the corresponding author on reasonable request and the approval of the institutional review board. The code

can be found at <https://github.com/ChengF-Lab/CO-ML>.

Study Design

Figure 1 shows the overview of the study design. We integrated both cardiovascular echocardiographic and laboratory testing variables from over 4300 longitudinal cancer patients. We developed and evaluated ML models to assist in the risk assessment of CTRCDs. We systematically tested 5 classification methods: *k*-nearest neighbors, logistic regression (LR), support vector machine (SVM), random forest (RF), and gradient tree boosting (GB). For the feature sets, we tested: (1) laboratory tests only, (2) echocardiography only, and (3) laboratory tests and echocardiography combined.

The generalizability of these models was verified by time-based data split. We also interrogated the final models to uncover clinically relevant variables associated with CTRCDs using learned weight analysis.

Study Population and Data Preparation

This study was reviewed and approved by the institutional review board and the patients gave informed consent. We extracted the clinical data of over 4600 oncology patients receiving cancer therapies from our institutional electronic medical health record database. All adult patients with cancer referred to the cardio-oncology service at the Cleveland Clinic from 1997 to 2018 were included. Five outcomes, including HF, AF, CAD, MI, and stroke, were extracted using *International Classification*

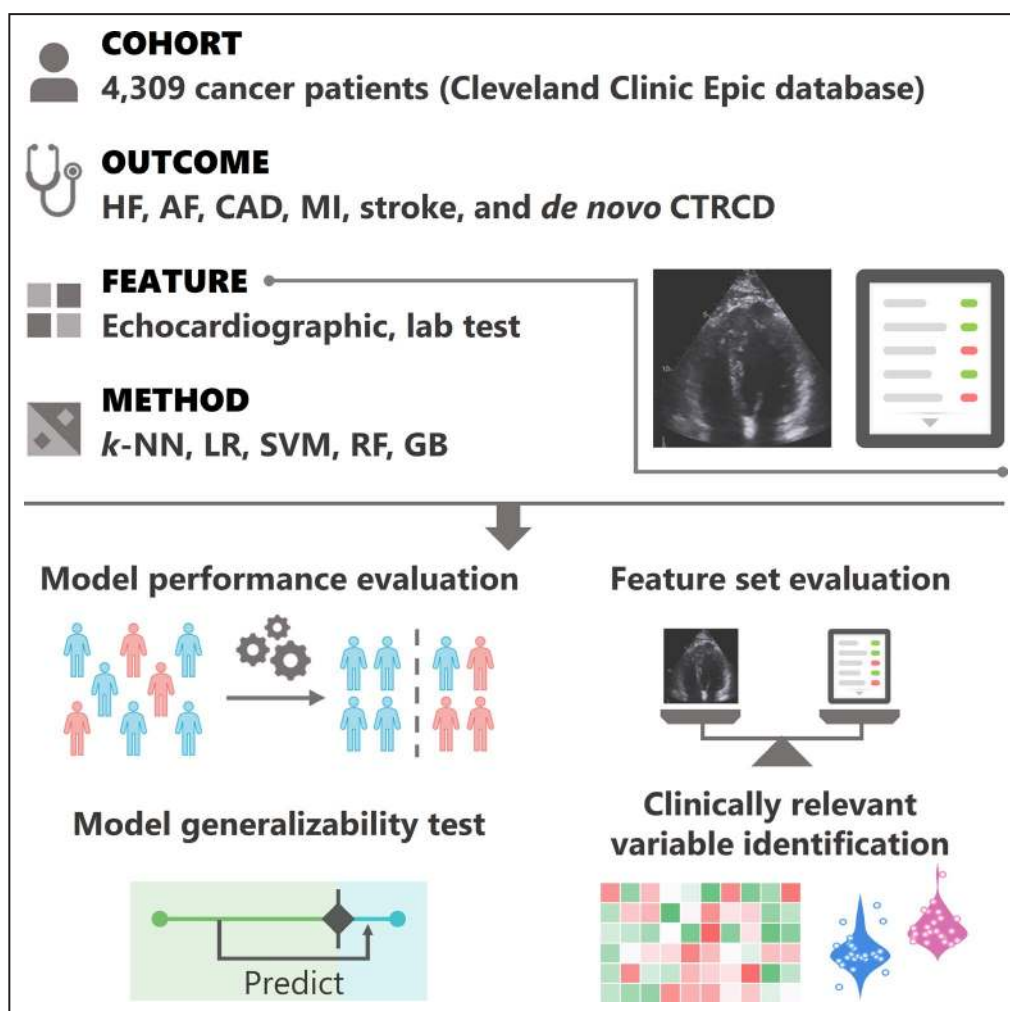


Figure 1. Overview of the study design.

We integrated both cardiovascular echocardiographic and laboratory testing variables from over 4300 longitudinal cancer patients for the prediction of 6 outcomes, including heart failure (HF), atrial fibrillation (AF), coronary artery disease (CAD), myocardial infarction (MI), stroke, and *de novo* cancer therapy-related cardiac dysfunction (CTRCD). We systematically tested 5 classification methods: *k*-nearest neighbors (*k*-NN), logistic regression (LR), support vector machine (SVM), random forest (RF), and gradient tree boosting (GB). For the feature sets, we tested laboratory test variables only, echocardiographic variables only, and laboratory test and echocardiographic variables combined.

of Diseases, Ninth and Tenth Revision (ICD-9, ICD-10), diagnosis codes and were manually checked by looking at patient charts on EPIC for accuracy (Epic Systems Corporation). Both inpatient and outpatient codes were included in this study. An additional outcome, de novo CTRCD, was also examined in this study. According to the diagnosis date of these 5 cardiac events, we identified the cardiac events that were diagnosed before cancer therapy as preexisting cardiac events and those after as de novo CTRCD. All variables were collected per patient based on the entirety of all available data. All patients had 2 sets of clinical variables: laboratory tests and echocardiographic variables. Laboratory test results included variables such as estimated glomerular filtration rate, glycated hemoglobin, glucose, calcium, total protein, and many others. Echocardiographic data included variables such as left ventricular ejection fraction, left ventricular end-systolic volume index, and left ventricular end-diastolic volume index. Since available echocardiographic data were longitudinal, we extracted several features for each echocardiographic variable: maximum of all follow-ups, minimum of all follow-ups, slope of all follow-ups, maximum increase within 3 months, and maximum decrease within 3 months (see Table S1 for a list of the variables). Finally, clinical variables were used as features to build ML models among 6 types of cardiovascular outcomes. After removing patients with >6 missing variables, the final data set contained 4309 patients (see Table for the characteristics of the cohort).

Classifier Development and Evaluation

Our first goal was to identify the optimal classification method and feature set combination. To do this, we systematically tested all of the combinations of 5 classification methods and 3 feature sets. For each outcome, we adopted a training-validation test procedure, repeated 100 times. In each iteration, all patients were randomly split into training set (81%), validation set (9%), or test set (10%). The training and validation sets were used in a grid search (Table S2) to identify the optimal hyperparameters for each classification method and feature set combination. Then, these 2 sets were merged and trained with the optimal hyperparameters to build the final model, which was evaluated using the test set. See Figure S1 for the detailed workflow of method and feature selection. All classification models were trained using the Python package scikit-learn.²⁰ We tested the effect of balancing the data sets using Synthetic Minority Oversampling Technique (SMOTE) implemented in the Python package imbalanced-learn.²¹

To test the generalizability of our ML models, we adopted a time-based data split strategy to simulate real-world scenarios, in which models used to predict new patients (external validation set) are built on data from the past. Specifically, we selected January 1, 2017 (2017.1.1) as the cutoff time point, as it produced subsequent test

Table 1. Characteristics of the Entire Cardio-Oncology Cohort

Variables	Cohort (N=4309)
Basic characteristics	
Age, y	61.1±13.7*
Sex	
Female	2552 (59) [†]
Male	1757 (41)
Body mass index, kg/m ²	28.3±7.3
Tobacco use	2162 (50)
Alcohol use	1995 (48)
Family history	1548 (36)
Comorbidity characteristics	
Hypertension	2450 (57)
Hyperlipidemia	1877 (44)
Diabetes mellitus	974 (23)
Chest pain	1724 (40)
Shortness of breath	1523 (35)
Fatigue	2202 (51)
Cardiac outcomes	
CTRCD	1560 (36)
HF	596 (14)
AF	653 (15)
CAD	673 (16)
MI	193 (4)
Stroke	275 (6)
Preexisting CVD	722 (17)
de novo CTRCD	838 (19)
Cancer therapy	
Chemotherapy	4011 (93)
Radiation	1969 (46)
Chemotherapy and radiation	1780 (41)
Anthracycline	1764 (41)
Cyclophosphamide	1567 (36)
Trastuzumab	822 (19)

AF indicates atrial fibrillation; CAD, coronary artery disease; CTRCD, cancer therapy–related cardiac function; CVD, cardiovascular disease; HF, heart failure; and MI, myocardial infarction.

*Continuous variables are reported as mean±SD.

[†]Categorical variables are reported as number (percentage).

sets with reasonable sizes. Patients who received cancer therapies before 2017.1.1 were used as the training set, and those who received cancer therapies after 2017.1.1 were used as the test set. The detailed workflow of this strategy is provided in Figure S2.

Model Criteria to Determine Predictive Variables

Next, we sought to understand which clinically relevant variables were significantly associated with CTRCD and further contributed to the high performance of ML models. We examined the weights of the 100 final LR

models for each outcome. LR learns a weight for each feature, and the prediction is the summation of all of the products of the weight and feature pairs squashed using a sigmoid function. We identified the clinically relevant variables based on 2 criteria: (1) the absolute coefficient of variation (the ratio of SD and mean) was low to ensure small fluctuation of the weight in the 100 repeats; (2) the absolute associated weight compared with the extremum weight for that outcome was high (relative weight). We used 0.5 and 0.3 as the 2 cutoffs:

$$|\text{Coefficient of variation}| < 0.5$$

$$\text{Relative weight} = \frac{|w_i|}{\max_{j \in T, \text{sgn}(w_j) = \text{sgn}(w_i)} |w_j|} > 0.3$$

where T denotes the feature set, w_i denotes the learned weight for feature i , and sgn is the sign function.

To verify the clinically relevant variables uncovered by examining the LR weights, we tested the hazard ratios (95% CIs) of the clinically relevant variables for the de novo CTRCD. The Wald χ^2 test was used to evaluate the variables with statistically significant coefficients. In addition, the log-rank test was used for global significance evaluation. The hazard analyses were performed with the survival (v2.44-1.1) and survminer (v0.4.6) packages on R 3.6.1.

Statistical Analysis

To evaluate the performance of ML models, we used 2 metrics: area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR). AUROC and AUPR were computed using the *metrics.roc_auc_score* and *metrics.average_precision_score* functions from the scikit-learn Python package. For the comparison of the performances of the laboratory test and echocardiographic feature sets, we applied a 2-sided paired sample t test using the AUROCs of the test sets from 100 iterations. $P < 0.05$ was considered statistically significant. The t test was performed using the *stats.ttest_rel* function from the SciPy Python package.²² We applied χ^2 test for the categorical variables to verify their associations with the outcomes. Kolmogorov-Smirnov test was used for the continuous variables. These 2 statistical analyses were performed by *stats.chi2_contingency* and *stats.ks_2samp* from the SciPy Python package.

RESULTS

Overview of the Classifier Performance

In this study, we built a large, longitudinal cardio-oncology cohort with 4309 oncology patients collected

from our institutional electronic medical record database (Table). The median age was 61.1 years (interquartile range [IQR], 53.8–70.5 years) for the overall population. Six types of cardiac events, including HF (n=596), AF (n=653), CAD (n=673), MI (n=193), stroke (n=275), and de novo CTRCD (n=838) were evaluated. In total, 1560 (36%) of patients had at least 1 type of diagnosed cardiac events, among which 722 (17%) patients had preexisting cardiac events/disease before cancer therapy, while 838 (19%) patients developed de novo CTRCD afterward. Among all of the patients, 4011 (93%) were treated with chemotherapy and 1969 (46%) were treated with radiation. For chemotherapy, 1764 (41%) patients were treated with anthracycline drugs (including doxorubicin, idarubicin, daunorubicin, and epirubicin), 1567 (36%) were treated with cyclophosphamide, and 822 (19%) patients were treated with trastuzumab. A list of all therapies can be found in Table S3. Two sets of clinical variables—laboratory tests (such as estimated glomerular filtration rate, glycated hemoglobin, glucose, calcium, and total protein) and echocardiographic variables (such as left ventricular ejection fraction, left ventricular end-diastolic volume index, and left ventricular end-systolic volume index)—were used to build the ML models. Table S1 lists all of the variables used in this study.

We conducted a systematic evaluation of 5 ML algorithms (k -nearest neighbors, LR, SVM, RF, and GB) and 3 feature sets (laboratory tests only, echocardiography only, or both combined). The average performance and SD for each outcome based on the 100 iterations are listed in Table S4 (AUROC) and Table S5 (AUPR). LR, RF, and GB achieved the first-tier performance, followed by SVM, then k -nearest neighbors. Although LR, RF, and GB performed similarly, LR achieved the highest AUROCs among 5 outcomes and comparable AUROC for HF which GB achieved the highest AUROC. LR was selected as the optimal classification method for all further analyses.

Figure 2 shows the overall performance for LR models. The AUROCs were 0.882 (95% CI, 0.878–0.887) for HF, 0.787 (95% CI, 0.782–0.792) for AF, 0.821 (95% CI, 0.815–0.826) for CAD, 0.807 (95% CI, 0.799–0.816) for MI, 0.660 (95% CI, 0.650–0.670) for stroke, and 0.802 (95% CI, 0.797–0.807) for de novo CTRCD. All AUPRs were at least 2-fold of their respective baselines of random classifiers. Precision-recall curve showed the trade-off between precision and recall, which, in this case, means the fraction of patients actually developed the disease in the patients who were predicted to have disease (precision) and their fraction in all of the patients who developed the disease (recall). In the case of a random classifier, the prediction error made by the classifier is consistent (a horizontal line in the precision-recall plot), thus leading to a baseline AUPR

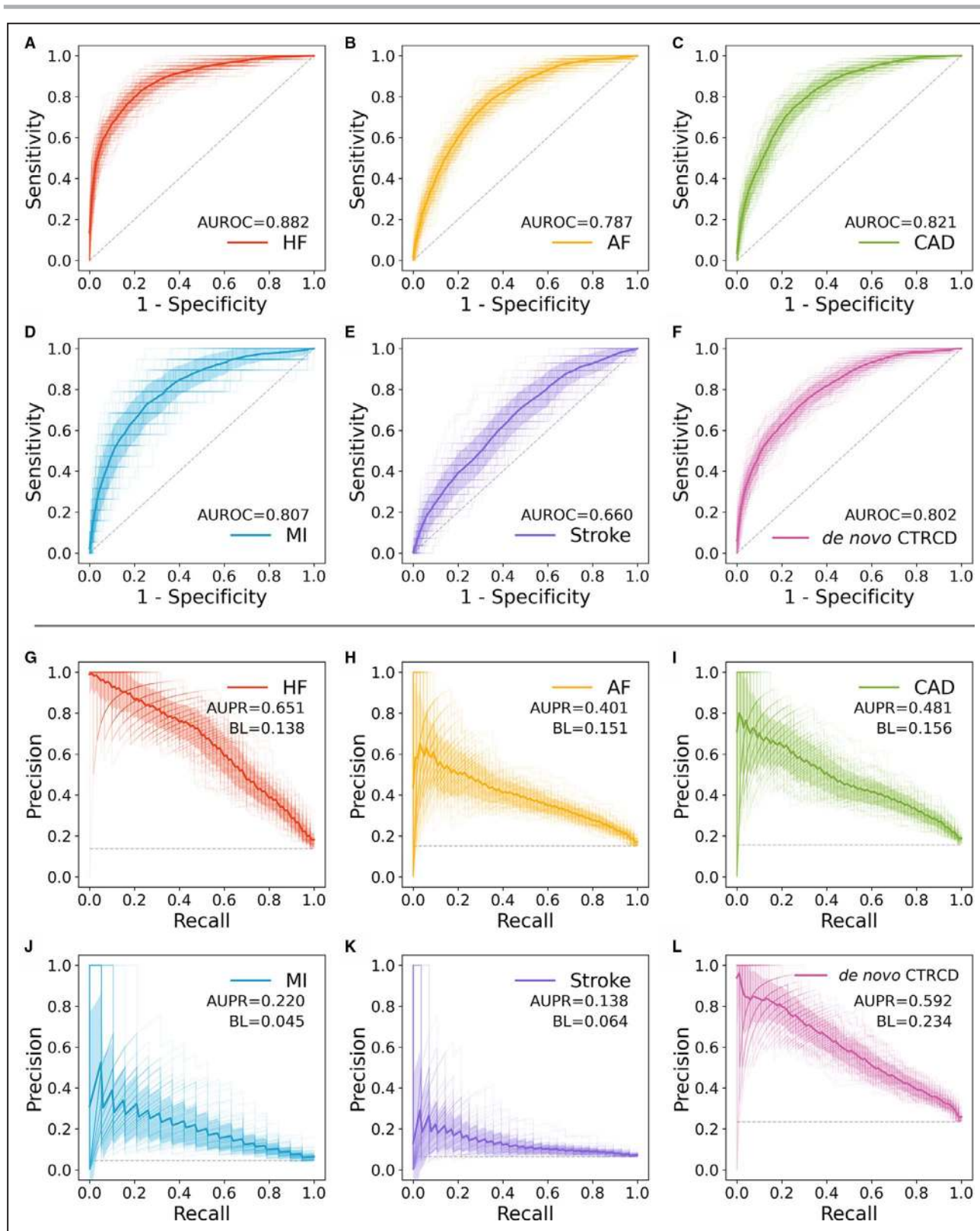


Figure 2. Performances for the 6 outcomes in receiver operating characteristic (A through F) and precision-recall (G through L) curves using logistic regression and the combined feature set.

For each subplot, light-colored lines correspond to the 100 iterations; the saturated-colored line is the average of the 100 iterations; background indicates mean±SD; the grey dotted line indicates the baseline of a random classifier. The area under the receiver operating characteristic curves (AUROCs) and area under the precision-recall curves (AUPRs) shown are the averages. AF indicates atrial fibrillation; CAD, coronary artery disease; CTRCD, cancer therapy-related cardiac dysfunction; HF, heart failure; and MI, myocardial infarction.

Downloaded from <http://ahajournals.org> by on December 1, 2020

that is the percentage of patients with the outcomes in the cohort. The AUPRs compared with their respective baselines were 0.651 (95% CI, 0.641–0.661) versus 0.138 for HF, 0.401 (95% CI, 0.392–0.411) versus 0.151 for AF, 0.481 (95% CI, 0.469–0.492) versus 0.156 for CAD, 0.220 (95% CI, 0.206–0.234) versus 0.045 for MI, 0.138 (95% CI, 0.131–0.146) versus 0.064 for stroke, and 0.592 (95% CI, 0.583–0.601) versus 0.234 for de novo CTRCD.

Combining Echocardiographic and Laboratory Test Variables Showed the Best Performance

Next, we wanted to find out the complementary effect of different feature sets on the model performance. Based on the 100 iterations, we found that while echocardiographic or laboratory test variables alone were predictive, inclusion of both types of data synergistically improved performance of the models (Figure 3

and Figure S3). Moreover, we showed that laboratory test and echocardiographic features performed differently among the outcomes (2-sided paired *t* test). Echocardiographic features outperformed laboratory test for HF (0.854 versus 0.729, $P < 0.001$), MI (0.766 versus 0.746, $P = 0.003$), and de novo CTRCD (0.742 versus 0.733, $P = 0.04$). Laboratory test outperformed echocardiographic features for AF (0.760 versus 0.700, $P < 0.001$), CAD (0.797 versus 0.702, $P < 0.001$), and stroke (0.656 versus 0.617, $P < 0.001$). In summary, combining both echocardiographic and laboratory test variables showed the best performance.

Generalizability of the Models

An important aspect of ML models is real-world generalizability. The patients were further split by dates—those with cancer therapy start dates before 2017.1.1 (see Methods) as the training set and those with start dates after 2017.1.1 as the test set. The results show

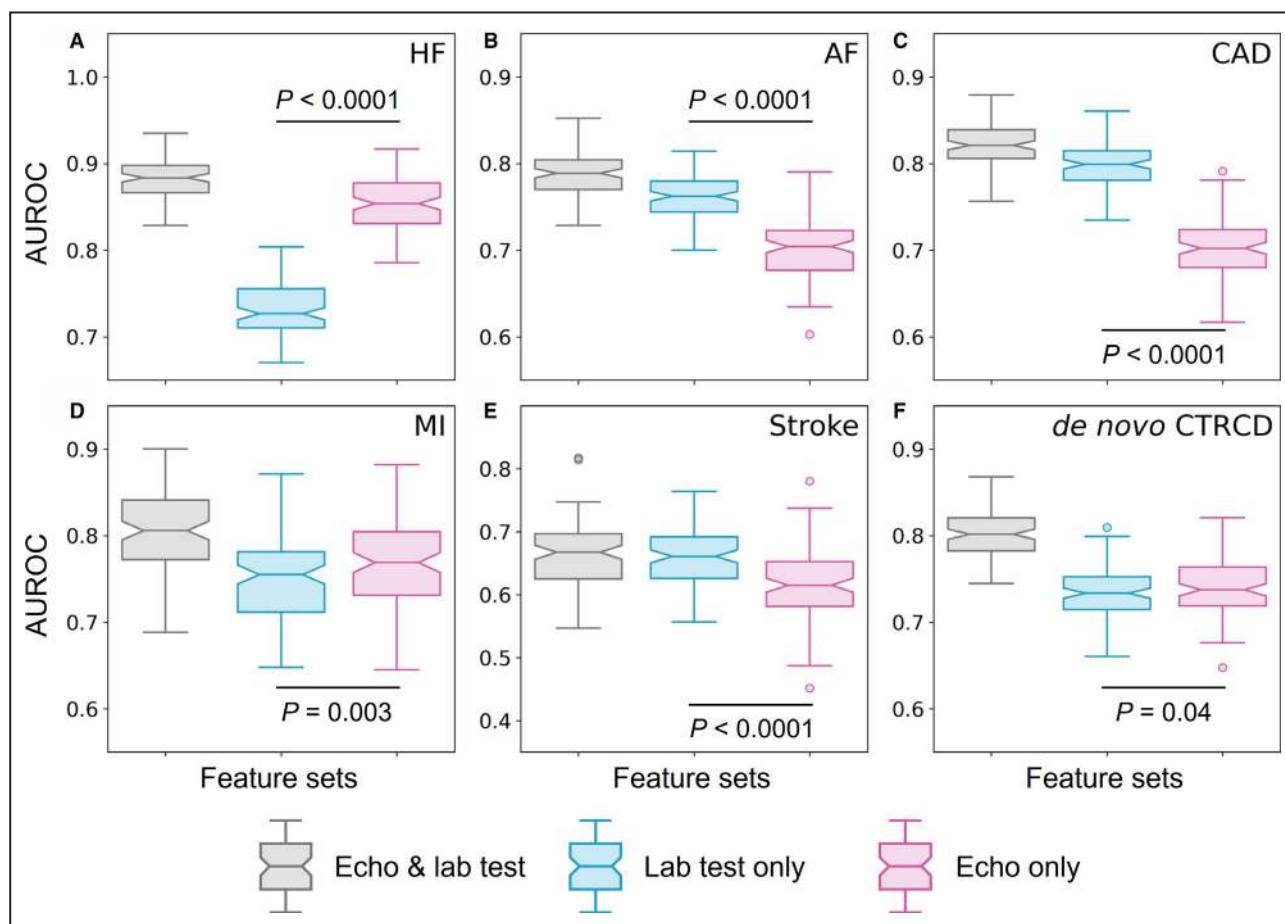


Figure 3. Comparison of the performances of laboratory test and echocardiographic feature sets.

A through F, When using the combined feature set, the models outperformed those that used either feature set individually. **A, D,** and **F,** Echocardiographic features showed significantly better performances for heart failure (HF), myocardial infarction (MI), and de novo cancer therapy–related cardiac dysfunction (CTRCD) than laboratory test. **B, C,** and **E,** Laboratory test features significantly outperformed echocardiographic features for atrial fibrillation (AF), coronary artery disease (CAD), and stroke. *P* values were calculated using 2-sided paired sample *t* test. AUROC indicates area under the receiver operating characteristic curve; and AUPR, area under the precision-recall curve.

that for all 6 outcomes, the AUROCs ranged from 0.913 for HF to 0.656 for MI (Figure 4 and Table S6). All AUPRs were higher than their corresponding baselines as well (Figure S4), indicating high generalizability of ML models in the prediction of CTRCD for new patients in real-world clinical practices.

Clinical Interpretability of the Models

We next interrogated what the LR models learned from the data to determine associations between clinical variables and the CTRCD outcomes. We examined the model weights of the 100 final models for each outcome. Using the mean and SD of the weight, we derived 2 metrics, coefficient of variation and relative weight (see Methods), to identify the features that have stable and relatively large absolute weights throughout the 100 iterations. Figure 5A shows the 23 variables that were predictive of at least 1 cardiovascular outcome; the actual values of the weights in the LR models can be found in Table S7. Age was most predictive

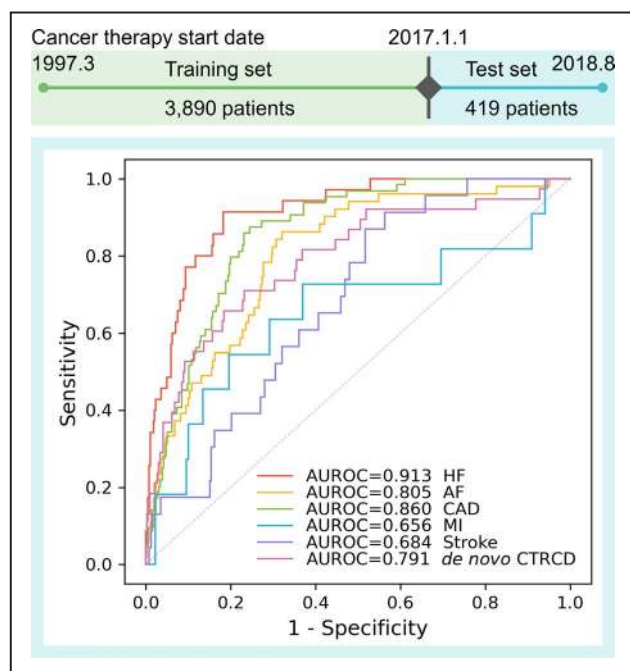


Figure 4. Evaluation of the model generalizability using time-split data.

The receiver operating characteristic curve for each outcome is shown. Dotted line indicates the theoretical baseline performance of a random classifier. Patients were split by the date January 1, 2017. Patients who received cancer therapies before this date were used for model training, and patients who received cancer therapies after this date comprised the test sets. Logistic regression and the combined feature set were used. All models achieved moderate to high performances, suggesting a high generalizability of the models. AF indicates atrial fibrillation; AUROC, area under the receiver operating characteristic curve; CAD, coronary artery disease; CTRCD, cancer therapy-related cardiac dysfunction; HF, heart failure; and MI, myocardial infarction.

for all 6 outcomes, followed by hypertension and left ventricular ejection fraction, which were also predictive for the 6 outcomes. The predictive variables for each outcome can be found in Table S8. Using Cox proportional hazards model analysis for de novo CTRCD, left ventricular ejection fraction, hazard ratio, and risk factors such as sex, age, and hypertension, were verified as predictive (Figure 5B). The distributions of the 23 variables among the patients further illustrated the clinical relevancy of the variables uncovered by LR model weight analysis (Figure 5C and 5D, Figures S5 and S6).

Impact of Cancer Treatment Types on the Models

We next examined whether cancer treatment information can affect the model performances by conducting 2 separate experiments.

In the first experiment, we pursued to find out whether our models could be applied to patients with specific types of cancer treatments. We generated 5 subpopulations (Table) based on whether the patients were treated with the following cancer therapies respectively: (1) chemotherapy, (2) radiation therapy, (3) chemotherapy and radiation therapy, (4) anthracycline, and (5) trastuzumab. We found high AUROCs in the prediction of de novo CTRCD among different types of cancer therapies as well (Figure 6). Specifically, the AUROCs were 0.779 (95% CI, 0.771–0.787) for anthracycline and 0.764 (95% CI, 0.746–0.783) for trastuzumab.

In the second experiment, we examined whether cancer therapy information used as features can improve model performances. We included 4 additional categorical features: the usage of chemotherapy, radiation, anthracycline, or trastuzumab. We found that incorporating treatment information had a marginal improvement on the model performances (AUROC: 0.805 versus 0.802; $P > 0.1$, *t* test) (Figure S7).

DISCUSSION

In this study, we built predictive ML models for cardiac risk assessment among 6 types of cardiovascular outcomes, including HF, AF, CAD, MI, stroke, and de novo CTRCD. Based on 100 model iterations, all outcomes received relatively high or high AUROC, ranging from 0.882 for HF to 0.660 for stroke (Figure 2). In addition, models built using time-split data demonstrated a high generalizability of our models for potential clinical implementation (Figure 4).

By comparing the model performances using different feature sets, we found that both laboratory test variables and echocardiographic variables contributed to the overall high performance. When laboratory test data were used alone, all outcomes still

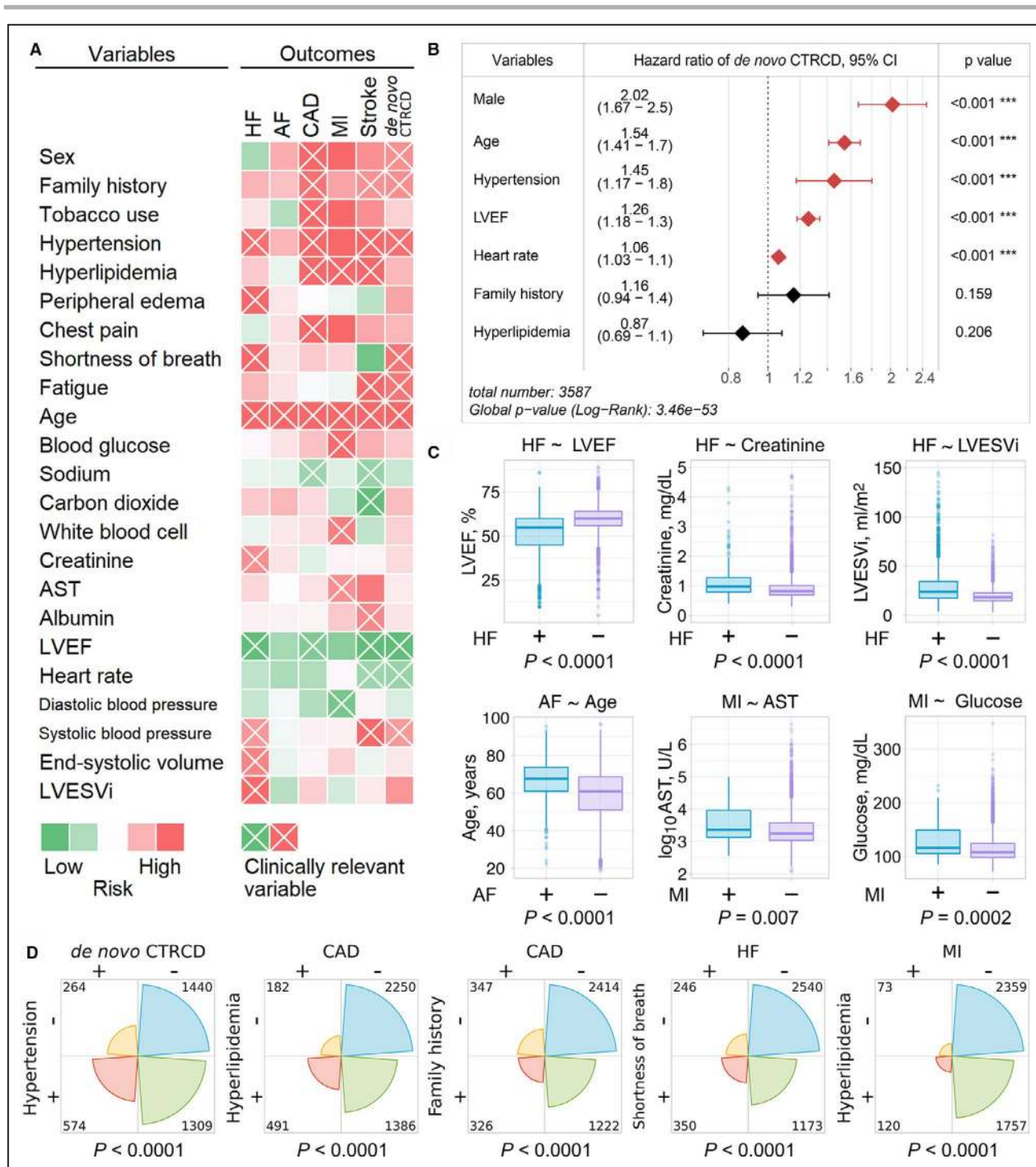


Figure 5. Clinically relevant variables uncovered by weight examination of the final logistic regression models. **A**, Twenty-three predictive variables for at least 1 outcome (marked by an “X” in the grid). Color gradient indicates that, as the value of the variable increases, the risk for the outcome increases (red) or decreases (green). **B**, Cox proportional hazards model analysis was performed for *de novo* cancer therapy-related cardiac dysfunction (CTRCD), which verified the clinically relevant variables using the machine learning method. **C**, Distributions of 6 continuous variables by the outcomes (*P* values were computed by Kolmogorov-Smirnov test). **D**, Distributions of 5 categorical variables (*P* values were computed by χ^2 test). +/- indicates whether the patients have the symptoms (row) or the outcomes (column). AF indicates atrial fibrillation; AST, aspartate aminotransferase; CAD, coronary artery disease; HF, heart failure; LVEF, left ventricular ejection fraction; LVESVi, left ventricular end-systolic volume index; and MI, myocardial infarction.

achieved moderate to high AUROCs (Figure 3), with 5 of the AUROCs >0.7 and 1 at 0.66. In addition, by comparing the performances of laboratory test and

echocardiographic feature sets, we found that for HF, MI, and *de novo* CTRCD, echocardiographic features significantly outperformed the laboratory test.

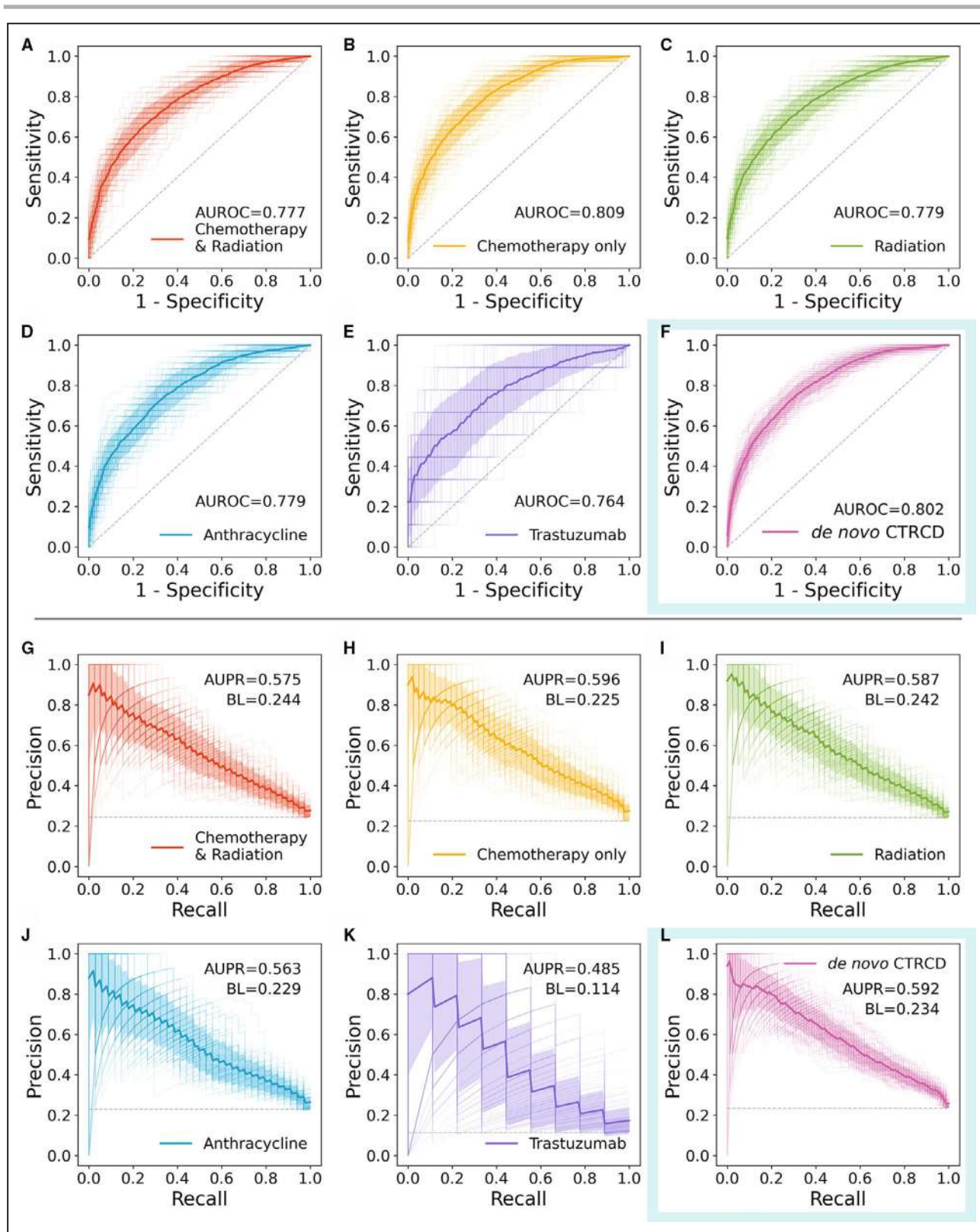


Figure 6. Performances for de novo cancer therapy-related cardiac dysfunction for patients with different cancer therapies. A through F, Receiver operating characteristic curves. G through L, Precision-recall curves. F and L, The model performances using all of the patients with de novo cancer therapy-related cardiac dysfunction (CTRCD) as comparison. For each subplot, light-colored lines correspond to the 100 iterations; saturated-colored line is the average of the 100 iterations; background indicates mean±SD; grey dotted line indicates the baseline of a random classifier. The area under the receiver operating characteristic curves (AUROCs) and area under the precision-recall curves (AUPRs) shown are the averages.

Downloaded from <http://ahajournals.org> by on December 1, 2020

For AF, CAD, and stroke, laboratory test performed better than echocardiographic features (Figure 3 and Figure S3). These highly predictive models offer potential approaches for cardio-oncology clinical practice. Oncologists referred these patients to the cardio-oncology services based on professional assessment of clinical factors such as cardiac symptoms, preexisting cardiac diseases, or cardiovascular risk factors. The models trained on laboratory test data could assist in the decision of referring, with or without incorporation of echocardiographic data.

To understand which specific variables contributed to model performance, we examined the learned weights for the features (Figure 5, Figures S5 and S6). We found that increased creatinine level was associated with high risk of cancer treatment–associated HF. In the general population, creatinine elevation in patients with HF is associated with increased mortality.²³ Creatinine is the metabolic product of creatine that is excreted in the urine.²⁴ An elevated glucose level is commonly found in patients with acute MI.²⁵ Studies have also shown that high glucose level is associated with high mortality risk in patients with MI.²⁶ Our results showed that a higher glucose level was associated with higher risks of cancer treatment–associated MI. Other risk factors, such as sex, hypertension, and age, were also verified. Men have a higher risk of heart disease than women.^{27–31} Age is a well-known CVD risk factor,^{32–34} and it was identified for all 6 outcomes. Hypertension is another strong risk factor for many types of CVDs.^{35–37} To summarize, by looking at the learned weights of the LR models, we uncovered the clinically relevant variables that were strong predictors for the CTRCD outcomes in the oncology cohort.

The skewness of the cardiovascular events in the data sets, especially in MI and stroke, could negatively affect the performances. Therefore, we tested this issue using SMOTE.³⁸ As shown in Figure S8, LR did not benefit from the resampling. The resampling marginally improved the performance of other methods for certain outcomes, such as *k*-nearest neighbors for MI, SVM for MI, and SVM for AF. However, the improved models still do not outperform LR. We also experimented with stacking the output of these models. We found that stacking LR, RF, and GB achieved a marginal improvement compared with using LR alone (Figure S9; HF, AF, and stroke). In summary, these observations suggest a low risk of data skewness in our current models, especially for LR models; yet, potential further improvements by combining the techniques such as stacking and resampling, and perhaps by a meta-classifier trained using the output of the models, are achievable in the future.

Our future work includes several directions. First, we will continue to improve the models as more data

are gathered, since we noticed marginally increased model performances when the training sizes increased (Figure S10), suggesting the importance of large-scale cohorts for ML studies. Our models may also be improved with a more model-specific variable selection procedure to further reduce risk of “overfitting.” When we tested the effect of limiting variables to a certain period (ie, variable collected within 1, 5, and 10 years of the first diagnosis for the outcome), we found that the models performed similarly, although certain outcomes may be slightly improved (Figure S11). Second, we are actively incorporating imaging data^{39,40} directly using convolutional neural networks to improve performance of models further. Third, we plan to integrate ML-based risk assessment with online tools for use in clinical practice.

Limitations

We acknowledge several potential limitations in the current study. First, because of the retrospective nature of this study and potential risk of patient selection bias, the model performances may be overestimated for real-world uses, even though model generalizability was evaluated with time-split data as the external validation set. Although each *ICD-9/10* diagnosis code was manually reviewed by a physician for accuracy, potential errors of *ICD-9/10* codes may influence the performance of ML models. In addition, while our models can output a probability for each outcome, they have not been explicitly programmed to predict risk levels. This could be considered in the next iteration of the models, in which a system of risk-based tertiles or quartiles could potentially be implemented based on our data.^{41,42}

We did not include feature interactions as additional features for the modeling using LR. Some risk factors are known to interact with others, such as sex and diabetes mellitus.⁴³ A potential improvement would be to include these interactions as features. However, we should also note that it could introduce a large number of features and could potentially increase the risk of model overfitting. In addition, some of the classification methods we have evaluated had such capacity, but they did not outperform LR.

We were able to identify several clinically relevant variables that were stable strong predictors of the outcomes. However, this method could not reveal all of the factors. When 2 features are linearly related (multicollinearity), their final learned weights may fluctuate and will depend on the initial randomization of the weights. These features will have high absolute coefficient of variations, and their contributions to the observed outcomes cannot easily be inferred using this method.

Last, although we applied L2 regularization for the training of the LR models, the models could still

potentially overfit. To overcome this, we could filter the features to remove irrelevant ones, which could be performed through variance analysis, mutual information, and L1 regularization.

CONCLUSIONS

ML models were built for each of 6 CTRCD outcomes for the oncology population based on a systematic evaluation of 5 classification methods and 3 feature sets. These models showed moderate to high performances and real-world generalizability using time-split data. We found that laboratory test and echocardiographic variables were each associated with different outcomes. We uncovered several clinically relevant variables associated with CTRCD, offering potential predictive factors and biomarkers for cardio-oncology clinical practices. Future versions of our models can include risk stratification in tertiles or quartiles to help with clinical decision-making to impact patient outcomes. To this end, we are currently working on the development of free online outcomes and risk calculators that integrate our models for shared decision-making. Our findings suggest that ML tools hold promise for cardiac risk assessment for patients before, during, or after cancer treatments by integrating large-scale, longitudinal patient data from healthcare systems.

ARTICLE INFORMATION

Received March 12, 2020; accepted October 15, 2020.

Affiliations

From the Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH (Y.Z., Y.H., F.C.); Robert and Suzanne Tomsich Department of Cardiovascular Medicine, Sydell and Arnold Miller Family Heart and Vascular Institute, Cleveland Clinic, Cleveland, OH (M.H., W.H.T., B.X., R.M., Z.P., L.C., M.K., B.G., M.K.C., S.K., P.C.); School of Medicine, Dentistry and Biomedical Sciences, Wellcome-Wolfson Institute of Experimental Medicine, Queen's University, Belfast, United Kingdom (M.H., C.W.); Cardio-Oncology Program, Division of Cardiovascular Medicine, Medical College of Wisconsin, Milwaukee, WI (S.-A.B.); Department of Hematology/Medical Oncology, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH (T.B., J.A., F.C.); Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH (W.H.T., M.K.C., P.C.); Department of Radiation Oncology, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH (C.S.); Department of Cardiovascular Surgery, Cleveland Clinic, Cleveland, OH (L.S.); and Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH (F.C.).

Sources of Funding

This work was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award numbers K99 HL138272 and R00 HL138272 to F.C.

Disclosures

None.

Supplementary Material

Tables S1–S8
Figures S1–S11

REFERENCES

- Curtin SC. Trends in cancer and heart disease death rates among adults aged 45–64: United States, 1999–2017. *Natl Vital Stat Rep*. 2019;68:1–8.
- Tan C, Tasaka H, Yu KP, Murphy ML, Karnofsky DA. Daunomycin, an antitumor antibiotic, in the treatment of neoplastic disease. Clinical evaluation with special reference to childhood leukemia. *Cancer*. 1967;20:333–353.
- Steinherz LJ. Cardiac toxicity 4 to 20 years after completing anthracycline therapy. *JAMA*. 1991;266:1672–1677.
- Hancock SL, Hoppe RT, Tucker MA. Factors affecting late mortality from heart disease after treatment of Hodgkin's disease. *JAMA*. 1993;270:1949–1955.
- Fedele P, Orlando L, Schiavone P, Ciccarese M, Forcignanò RC, Calvani N, Marino A, Nacci A, Sponziello F, Mazzoni E, et al. Clinical outcomes and cardiac safety of continuous antiHer2 therapy in c-erbB2-positive metastatic breast cancer patients. *J Chemother*. 2013;25:369–375.
- Hahn VS, Lenihan DJ, Ky B. Cancer therapy-induced cardiotoxicity: basic mechanisms and potential cardioprotective therapies. *J Am Heart Assoc*. 2014;3:e000665. DOI: 10.1161/JAHA.113.000665
- Pituskin E, Mackey JR, Koshman S, Jassal D, Pitz M, Haykowsky MJ, Pagano JJ, Chow K, Thompson RB, Vos LJ, et al. Multidisciplinary approach to novel therapies in cardio-oncology research (MANTICORE 101-Breast): a randomized trial for the prevention of trastuzumab-associated cardiotoxicity. *J Clin Oncol*. 2017;35:870–877.
- Lee J, Hur H, Lee JW, Youn HJ, Han K, Kim NW, Jung SY, Kim Z, Kim KS, Lee MH, et al. Long-term risk of congestive heart failure in younger breast cancer survivors: a nationwide study by the SMARTSHIP group. *Cancer*. 2020;126:181–188.
- Brown SA, Sandhu N, Herrmann J. Systems biology approaches to adverse drug effects: the example of cardio-oncology. *Nat Rev Clin Oncol*. 2015;12:718–731.
- Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin*. 2019;69:7–34.
- Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, Jemal A, Kramer JL, Siegel RL. Cancer treatment and survivorship statistics, 2019. *CA Cancer J Clin*. 2019;69:363–385.
- Cai C, Fang J, Guo P, Wang Q, Hong H, Moslehi J, Cheng F. In silico pharmacoepidemiologic evaluation of drug-induced cardiovascular complications using combined classifiers. *J Chem Inf Model*. 2018;58:943–956.
- Cai C, Guo P, Zhou Y, Zhou J, Wang Q, Zhang F, Fang J, Cheng F. Deep learning-based prediction of drug-induced cardiotoxicity. *J Chem Inf Model*. 2019;59:1073–1084.
- Feeny AK, Rickard J, Patel D, Toro S, Trulock KM, Park CJ, Labarbera MA, Varma N, Niebauer MJ, Sinha S, et al. Machine learning prediction of response to cardiac resynchronization therapy: Improvement versus current guidelines. *Circ Arrhythm Electrophysiol*. 2019;12:e007316. DOI: 10.1161/CIRCEP.119.007316
- Wang Y, Li J, Zheng X, Jiang Z, Hu S, Wadhwa RK, Bai X, Lu J, Wang Q, Li Y, et al. Risk factors associated with major cardiovascular events 1 year after acute myocardial infarction. *JAMA Netw Open*. 2018;1:e181079.
- Xu B, Kocyigit D, Grimm R, Griffin BP, Cheng F. Applications of artificial intelligence in multimodality cardiovascular imaging: a state-of-the-art review. *Prog Cardiovasc Dis*. 2020;63:367–376.
- Krumholz HM, Coppi AC, Warner F, Triche EW, Li SX, Mahajan S, Li Y, Bernheim SM, Grady J, Dorsey K, et al. Comparative effectiveness of new approaches to improve mortality risk models from Medicare claims data. *JAMA Netw Open*. 2019;2:e197314.
- Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *NPJ Digit Med*. 2018;1:40.
- Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380:1347–1358.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
- Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res*. 2017;18:1–5.
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17:261–272.

23. Smith GL, Vaccarino V, Kosiborod M, Lichtman JH, Cheng S, Watnick SG, Krumholz HM. Worsening renal function: what is a clinically meaningful change in creatinine during hospitalization with heart failure? *J Card Fail.* 2003;9:13–25.
24. Perrone RD, Madias NE, Levey AS. Serum creatinine as an index of renal function: new insights into old concepts. *Clin Chem.* 1992;38:1933–1953.
25. Kosiborod M. Blood glucose and its prognostic implications in patients hospitalised with acute myocardial infarction. *Diabetes Vasc Dis Res.* 2008;5:269–275.
26. Ishihara M. Acute hyperglycemia in patients with acute myocardial infarction. *Circ J.* 2012;76:563–571.
27. Kannel WB, Hjortland MC, McNamara PM, Gordon T. Menopause and risk of cardiovascular disease: the Framingham study. *Ann Intern Med.* 1976;85:447–452.
28. Njølstad I, Arnesen E, Lund-Larsen PG. Smoking, serum lipids, blood pressure, and sex differences in myocardial infarction: a 12-year follow-up of the Finnmark study. *Circulation.* 1996;93:450–456.
29. Vitale C, Fini M, Speziale G, Chierchia S. Gender differences in the cardiovascular effects of sex hormones. *Fundam Clin Pharmacol.* 2010;24:675–685.
30. Dallongeville J, De Bacquer D, Heidrich J, De Backer G, Prugger C, Kotseva K, Montaye M, Amouyel P. Gender differences in the implementation of cardiovascular prevention measures after an acute coronary event. *Heart.* 2010;96:1744–1749.
31. De Smedt D, De Bacquer D, De Sutter J, Dallongeville J, Gevaert S, De Backer G, Bruthans J, Kotseva K, Reiner Ž, Tokgözoğlu L, et al. The gender gap in risk factor control: effects of age and education on the control of cardiovascular risk factors in male and female coronary patients. the EUROASPIRE IV study by the European Society of Cardiology. *Int J Cardiol.* 2016;209:284–290.
32. Castelli WP. Epidemiology of coronary heart disease: the Framingham study. *Am J Med.* 1984;76:4–12.
33. Rich-Edwards JW, Manson JE, Hennekens CH, Buring JE. The primary prevention of coronary heart disease in women. *N Engl J Med.* 1995;332:1758–1766.
34. Dhingra R, Vasan RS. Age as a risk factor. *Med Clin North Am.* 2012;96:87–91.
35. Wu CY, Hu HY, Chou YJ, Huang N, Chou YC, Li CP. High blood pressure and all-cause and cardiovascular disease mortalities in community-dwelling older adults. *Medicine (Baltimore).* 2015;94:e2160.
36. Stevens SL, Wood S, Koshiaris C, Law K, Glasziou P, Stevens RJ, McManus RJ. Blood pressure variability and cardiovascular disease: systematic review and meta-analysis. *BMJ.* 2016;354:i4098.
37. Kjeldsen SE. Hypertension and cardiovascular risk: general aspects. *Pharmacol Res.* 2018;129:95–99.
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321–357.
39. Henglin M, Stein G, Hushcha PV, Snoek J, Wiltchko AB, Cheng S. Machine learning approaches in cardiovascular imaging. *Circ Cardiovasc Imaging.* 2017;10:e005614.
40. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan CR, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation.* 2018;138:1623–1635.
41. Kang Y, Assuncao BL, Denduluri S, McCurdy S, Luger S, Lefebvre B, Carver J, Scherrer-Crosbie M. Symptomatic heart failure in acute leukemia patients treated with anthracyclines. *JACC CardioOncol.* 2019;1:208–217.
42. Abdel-Qadir H, Thavendiranathan P, Austin PC, Lee DS, Amir E, Tu JV, Fung K, Anderson GM. Development and validation of a multivariable prediction model for major adverse cardiovascular events after early stage breast cancer: a population-based cohort study. *Eur Heart J.* 2019;40:3913–3920.
43. Wakabayashi I. Gender differences in cardiovascular risk factors in patients with coronary artery disease and those with type 2 diabetes. *J Thorac Dis.* 2017;9:E503–E506.

SUPPLEMENTAL MATERIAL

Table S1. Clinical variables used in this study.

Lab test (including demographic)	Echocardiographic
Sex	LVEF (left ventricular ejection fraction)
Race	Heart rate
Family history	BSA (body surface area)
Tobacco use	SBP (systolic blood pressure)
Alcohol use	DBP (diastolic blood pressure)
Diabetes	EDV (end-diastolic volume)
Hypertension	ESV (end-systolic volume)
Hyperlipidemia	LVEDVi (left ventricular end-diastolic volume index)
Peripheral edema	LVESVi (left ventricular end-systolic volume index)
Orthopnea	
Chest pain	
Shortness of breath	
Fatigue	
Age	
BMI (body mass index)	
eGFR (estimated glomerular filtration rate)	
RBC (red blood cell)	
Hematocrit	
MCHC (mean corpuscular hemoglobin concentration)	
MCV (mean corpuscular volume)	
MCH (mean corpuscular hemoglobin)	
Blood glucose	
Calcium	
Total protein	
Sodium	
Potassium	
Chloride	
Carbon dioxide	
WBC (white blood cell)	
Platelet	
Creatinine	
ALT (alanine aminotransferase)	
AST (aspartate aminotransferase)	
Albumin	
ALP (alkaline phosphatase)	
Bilirubin	

Table S2. Classification methods evaluated and hyperparameters explored.

Classifier	Hyperparameter	Value
<i>k</i> -nearest neighbors (<i>k</i> -NN)	K	3, 5, 7
	metric	Euclidean, correlation
Logistic regression (LR)	C	0.01, 0.1, 1, 10, 100, 1000
Support vector machine (SVM)	C	0.1, 1, 10, 100
	gamma	1e-3, 1e-2, 1e-1
Random forest (RF)	max features	0.1, 0.2, 0.4, 0.8
	max depth	4, 8, 12
Gradient tree boosting (GB)	n estimators	100, 500
	max depth	2, 3, 4
	learning rate	0.01, 0.05, 0.1
	subsample	0.33, 0.66, 1

Table S3. A list of all therapies used in the entire cardio-oncology cohort.

Therapy	Patients	Therapy	Patients	Therapy	Patients	Therapy	Patients
Radiation	1969	Imatinib	89	Panobinostat	12	Elotuzumab	3
Cyclophosphamide	1567	Mercaptopurine	83	Regorafenib	12	Folinic Acid	3
Doxorubicin	1304	Dasatinib	80	Trimethoprim	12	Lomustine	3
Carboplatin	838	Vinorelbine	76	Mitomycin	11	Melphalan	3
Trastuzumab	822	Eribulin	75	Idelalisib	10	Oxaliplatin	3
Paclitaxel	684	Pazopanib	75	Methenamine	10	Pembrolizumab	3
Docetaxel	662	Sunitinib	73	Midostaurin	10	Trastuzumab emtansine	3
Cytarabine	645	Goserelin	72	Pertuzumab	10	Alectinib	2
Rituximab	632	Ixazomib	66	Prednisolone	10	Alitretinoin	2
Anastrozole	620	Leuprolide	66	Thioguanine	10	Arsenic	2
Vincristine	602	Sorafenib	64	Venetoclax	10	Axicabtagene ciloleucel	2
Etoposide	595	Palbociclib	63	Daratumumab	9	Flutamide	2
Methotrexate	506	Irinotecan	59	Olaparib	9	Ingenol	2
Busulfan	493	Methoxsalen	57	Pralatrexate	9	Lenvatinib	2
Lenalidomide	393	Fosfomycin	50	Aminolevulinic	8	Methenamine	2
Tamoxifen	385	Nilotinib	43	Gemcitabine	8	Mitotane	2
Plerixafor	371	Ruxolitinib	43	Temsirolimus	8	Palifosfamide	2
Bortezomib	356	Axitinib	42	Trifluridine	8	Ribociclib	2
Letrozole	302	Temozolomide	41	Chlorambucil	7	Sipuleucel-t	2
Daunorubicin	287	Tretinoin	36	Afatinib	6	Aldoxorubicin	1
Ifosfamide	281	Trametinib	31	Bexarotene	6	Atezolizumab	1
Fluorouracil	238	Erlotinib	30	Dactinomycin	6	BCG LIVE 81 MG INTRAVESICAL SUSPENSION	1
Hydroxyurea	226	Bicalutamide	28	Pegaspargase	6	Brentuximab	1
Cisplatin	222	Vorinostat	28	Procabazine	6	Cabazitaxel	1
Exemestane	220	Dabrafenib	27	Triptorelin	6	Enasidenib	1
Capecitabine	206	Neratinib	25	Vemurafenib	6	Fludarabine	1
Idarubicin	158	Topotecan	25	Asparaginase	5	Gefitinib	1
Leucovorin	153	Coenzyme M	24	Ceritinib	5	Histrelin	1
Carfilzomib	148	Ponatinib	24	Cobimetinib	5	Memantine	1
Mitoxantrone	147	Abiraterone	22	Niraparib	5	Nelarabine	1
Pomalidomide	136	Enzalutamide	20	Osimertinib	5	Nilutamide	1
Dacarbazine	129	Romidepsin	18	Pemetrexed	5	Octreotide	1
Everolimus	129	Alemtuzumab	15	Acalabrutinib	4	Omacetaxine	1
Bevacizumab	121	Cabozantinib	15	Decitabine	4	Peginterferon alfa- 2b	1
Vinblastine	119	Cetuximab	15	Ipilimumab	4	Pentostatin	1
Bleomycin	116	Crizotinib	15	Mechlorethamine	4	Thalidomide	1
Azacitidine	105	Epirubicin	15	Talimogene laherparepvec	4	Thiotepa	1
Ibrutinib	103	Bacillus calmette - guerin substrain tice live antigen	13	Abemaciclib	3	Vandetanib	1
Megestrol	93	Nivolumab	13	Aldesleukin	3		
Lapatinib	90	Bosutinib	12	Amifostine	3		

Table S4. Model performance in area under the receiver operating characteristic curve for all methods and feature sets combinations.

Outcome	k-NN	LR	SVM	RF	GB
	Feature - Combined				
HF	0.788 ± 0.035	0.882 ± 0.022	0.877 ± 0.025	0.876 ± 0.025	0.884 ± 0.022
AF	0.652 ± 0.039	0.787 ± 0.025	0.734 ± 0.031	0.772 ± 0.028	0.774 ± 0.028
CAD	0.675 ± 0.032	0.821 ± 0.027	0.795 ± 0.027	0.798 ± 0.029	0.809 ± 0.028
MI	0.597 ± 0.055	0.807 ± 0.045	0.702 ± 0.061	0.764 ± 0.050	0.756 ± 0.050
Stroke	0.531 ± 0.047	0.660 ± 0.052	0.527 ± 0.060	0.649 ± 0.052	0.643 ± 0.052
<i>de novo</i> CTRCD	0.681 ± 0.035	0.802 ± 0.027	0.793 ± 0.028	0.779 ± 0.027	0.789 ± 0.026
	Feature - Lab test				
HF	0.644 ± 0.035	0.729 ± 0.031	0.664 ± 0.044	0.726 ± 0.033	0.725 ± 0.034
AF	0.642 ± 0.030	0.760 ± 0.025	0.670 ± 0.038	0.747 ± 0.026	0.744 ± 0.026
CAD	0.660 ± 0.033	0.797 ± 0.027	0.720 ± 0.033	0.778 ± 0.026	0.781 ± 0.028
MI	0.594 ± 0.058	0.746 ± 0.045	0.645 ± 0.065	0.709 ± 0.047	0.710 ± 0.047
Stroke	0.538 ± 0.048	0.656 ± 0.048	0.572 ± 0.058	0.641 ± 0.049	0.641 ± 0.052
<i>de novo</i> CTRCD	0.643 ± 0.033	0.733 ± 0.030	0.683 ± 0.034	0.721 ± 0.031	0.721 ± 0.029
	Feature - Echo				
HF	0.755 ± 0.039	0.854 ± 0.030	0.843 ± 0.034	0.851 ± 0.028	0.858 ± 0.028
AF	0.583 ± 0.038	0.700 ± 0.031	0.644 ± 0.041	0.692 ± 0.036	0.695 ± 0.036
CAD	0.632 ± 0.037	0.702 ± 0.036	0.639 ± 0.038	0.703 ± 0.034	0.715 ± 0.033
MI	0.613 ± 0.055	0.766 ± 0.054	0.706 ± 0.073	0.745 ± 0.059	0.728 ± 0.060
Stroke	0.516 ± 0.043	0.617 ± 0.054	0.519 ± 0.052	0.589 ± 0.050	0.579 ± 0.052
<i>de novo</i> CTRCD	0.651 ± 0.034	0.742 ± 0.033	0.722 ± 0.033	0.728 ± 0.031	0.733 ± 0.029

HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction; CTRCD - cancer therapy-related cardiac dysfunction; k-NN - k-nearest neighbors; LR - logistic regression; SVM - support vector machine; RF - random forest; GB - gradient tree boosting.

Table S5. Model performance in area under the precision-recall curve for all methods and feature sets combinations.

Outcome	Baseline	k-NN	LR	SVM	RF	GB
		Feature - Combined				
HF	0.138	0.466 ± 0.055	0.651 ± 0.051	0.641 ± 0.056	0.632 ± 0.058	0.639 ± 0.054
AF	0.152	0.239 ± 0.034	0.401 ± 0.049	0.344 ± 0.051	0.385 ± 0.051	0.378 ± 0.051
CAD	0.156	0.284 ± 0.040	0.481 ± 0.057	0.432 ± 0.050	0.455 ± 0.054	0.471 ± 0.052
MI	0.045	0.078 ± 0.028	0.220 ± 0.069	0.130 ± 0.050	0.207 ± 0.069	0.174 ± 0.059
Stroke	0.064	0.082 ± 0.022	0.138 ± 0.037	0.093 ± 0.030	0.120 ± 0.036	0.126 ± 0.038
<i>de novo</i> CTRCD	0.234	0.402 ± 0.045	0.592 ± 0.048	0.580 ± 0.047	0.563 ± 0.046	0.572 ± 0.045
		Feature - Lab test				
HF	0.138	0.238 ± 0.033	0.345 ± 0.045	0.309 ± 0.052	0.343 ± 0.047	0.340 ± 0.053
AF	0.152	0.228 ± 0.024	0.352 ± 0.041	0.281 ± 0.041	0.326 ± 0.041	0.320 ± 0.037
CAD	0.156	0.247 ± 0.026	0.419 ± 0.051	0.342 ± 0.042	0.391 ± 0.046	0.393 ± 0.044
MI	0.045	0.072 ± 0.025	0.162 ± 0.055	0.092 ± 0.026	0.119 ± 0.033	0.119 ± 0.036
Stroke	0.064	0.078 ± 0.016	0.136 ± 0.039	0.098 ± 0.027	0.119 ± 0.025	0.125 ± 0.037
<i>de novo</i> CTRCD	0.234	0.332 ± 0.032	0.477 ± 0.048	0.431 ± 0.047	0.456 ± 0.043	0.452 ± 0.047
		Feature - Echo				
HF	0.138	0.446 ± 0.053	0.623 ± 0.056	0.610 ± 0.061	0.615 ± 0.060	0.621 ± 0.060
AF	0.152	0.195 ± 0.022	0.327 ± 0.045	0.263 ± 0.038	0.315 ± 0.043	0.316 ± 0.047
CAD	0.156	0.264 ± 0.038	0.342 ± 0.052	0.296 ± 0.048	0.363 ± 0.051	0.365 ± 0.051
MI	0.045	0.085 ± 0.031	0.187 ± 0.063	0.125 ± 0.040	0.196 ± 0.065	0.166 ± 0.063
Stroke	0.064	0.071 ± 0.008	0.109 ± 0.027	0.079 ± 0.014	0.103 ± 0.028	0.101 ± 0.028
<i>de novo</i> CTRCD	0.234	0.379 ± 0.038	0.527 ± 0.051	0.508 ± 0.052	0.519 ± 0.047	0.514 ± 0.046

HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction; CTRCD - cancer therapy-related cardiac dysfunction; k-NN - k-nearest neighbors; LR - logistic regression; SVM - support vector machine; RF - random forest; GB - gradient tree boosting.

Table S6. Test set performances by time-based data split.

Outcome	Number of cases vs. total patients*	AUROC	AUPR
HF	35 / 419 (0.084)	0.913	0.548
AF	51 / 419 (0.122)	0.805	0.379
CAD	64 / 419 (0.153)	0.860	0.494
MI	11 / 419 (0.026)	0.656	0.071
Stroke	23 / 419 (0.055)	0.684	0.120
<i>de novo</i> CTRCD	38 / 331 (0.115)	0.791	0.431

* Number of patients in the test sets. Values in the parentheses are the fractions of patients having the corresponding outcome, which are also the baselines for the AUPRs. HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction; CTRCD - cancer therapy-related cardiac dysfunction; AUROC - area under the receiver operating characteristic curve; AUPR - area under the precision-recall curve.

Table S7. Mean and standard deviation of the logistic regression weights of the clinically relevant variables from 100 iterations.

Variable	HF	AF	CAD	MI	Stroke	<i>de novo</i> CTRCD
Sex	-0.18 ± 0.12	0.20 ± 0.08	0.51 ± 0.20	0.39 ± 0.29	0.12 ± 0.10	0.18 ± 0.09
Family history	0.11 ± 0.04	0.15 ± 0.06	0.36 ± 0.11	0.13 ± 0.07	0.11 ± 0.05	0.19 ± 0.06
Tobacco use	0.04 ± 0.04	-0.14 ± 0.12	0.42 ± 0.21	0.25 ± 0.20	0.12 ± 0.09	0.08 ± 0.04
Hypertension	0.20 ± 0.08	0.18 ± 0.07	0.46 ± 0.13	0.28 ± 0.15	0.26 ± 0.12	0.24 ± 0.07
Hyperlipidemia	0.07 ± 0.03	-0.04 ± 0.06	0.77 ± 0.21	0.22 ± 0.10	0.21 ± 0.08	0.12 ± 0.02
Peripheral edema	0.37 ± 0.18	0.06 ± 0.03	0.01 ± 0.04	-0.03 ± 0.05	-0.04 ± 0.06	0.15 ± 0.04
Chest pain	-0.07 ± 0.10	0.08 ± 0.04	0.45 ± 0.13	0.36 ± 0.20	0.09 ± 0.04	0.12 ± 0.03
Shortness of breath	0.41 ± 0.15	0.06 ± 0.02	0.13 ± 0.03	0.06 ± 0.04	-0.09 ± 0.08	0.23 ± 0.04
Fatigue	0.10 ± 0.03	0.07 ± 0.03	-0.01 ± 0.06	-0.02 ± 0.06	0.30 ± 0.14	0.23 ± 0.05
Age	0.28 ± 0.04	0.60 ± 0.11	0.62 ± 0.09	0.23 ± 0.05	0.31 ± 0.06	0.50 ± 0.07
Blood glucose	0.01 ± 0.03	0.06 ± 0.02	0.16 ± 0.02	0.21 ± 0.03	0.08 ± 0.04	0.09 ± 0.02
Sodium	-0.04 ± 0.02	-0.04 ± 0.02	-0.17 ± 0.05	-0.05 ± 0.03	-0.06 ± 0.02	-0.07 ± 0.02
Carbon dioxide	0.07 ± 0.02	0.18 ± 0.02	0.09 ± 0.03	-0.09 ± 0.06	-0.10 ± 0.03	0.11 ± 0.02
White blood cell	-0.04 ± 0.02	0.05 ± 0.01	0.08 ± 0.02	0.18 ± 0.04	-0.04 ± 0.02	0.07 ± 0.02
Creatinine	0.16 ± 0.02	0.06 ± 0.03	-0.05 ± 0.06	0.01 ± 0.06	0.01 ± 0.03	0.06 ± 0.03
AST	0.05 ± 0.05	0.00 ± 0.02	0.07 ± 0.03	0.16 ± 0.07	0.14 ± 0.09	0.03 ± 0.02
Albumin	0.02 ± 0.02	0.03 ± 0.02	0.03 ± 0.01	0.07 ± 0.02	0.12 ± 0.02	0.04 ± 0.01
LVEF	-0.66 ± 0.05	-0.17 ± 0.04	-0.22 ± 0.05	-0.20 ± 0.11	-0.16 ± 0.03	-0.44 ± 0.03
Heart rate	-0.13 ± 0.04	-0.16 ± 0.06	-0.14 ± 0.03	0.01 ± 0.03	-0.06 ± 0.03	-0.14 ± 0.04
Diastolic blood pressure	-0.12 ± 0.02	-0.02 ± 0.02	-0.14 ± 0.04	-0.23 ± 0.09	0.01 ± 0.04	-0.05 ± 0.03
Systolic blood pressure	0.15 ± 0.04	-0.02 ± 0.03	0.03 ± 0.04	0.02 ± 0.12	0.19 ± 0.05	0.17 ± 0.04
End-systolic volume	0.17 ± 0.06	-0.04 ± 0.08	0.02 ± 0.11	0.06 ± 0.09	-0.01 ± 0.18	0.04 ± 0.21
LVESVi	0.21 ± 0.06	-0.13 ± 0.10	0.12 ± 0.11	-0.07 ± 0.12	0.02 ± 0.14	0.17 ± 0.19
Minimum weight	-0.66	-0.60	-0.57	-0.54	-0.19	-0.44
Maximum weight	0.42	0.75	0.77	0.43	0.31	0.50

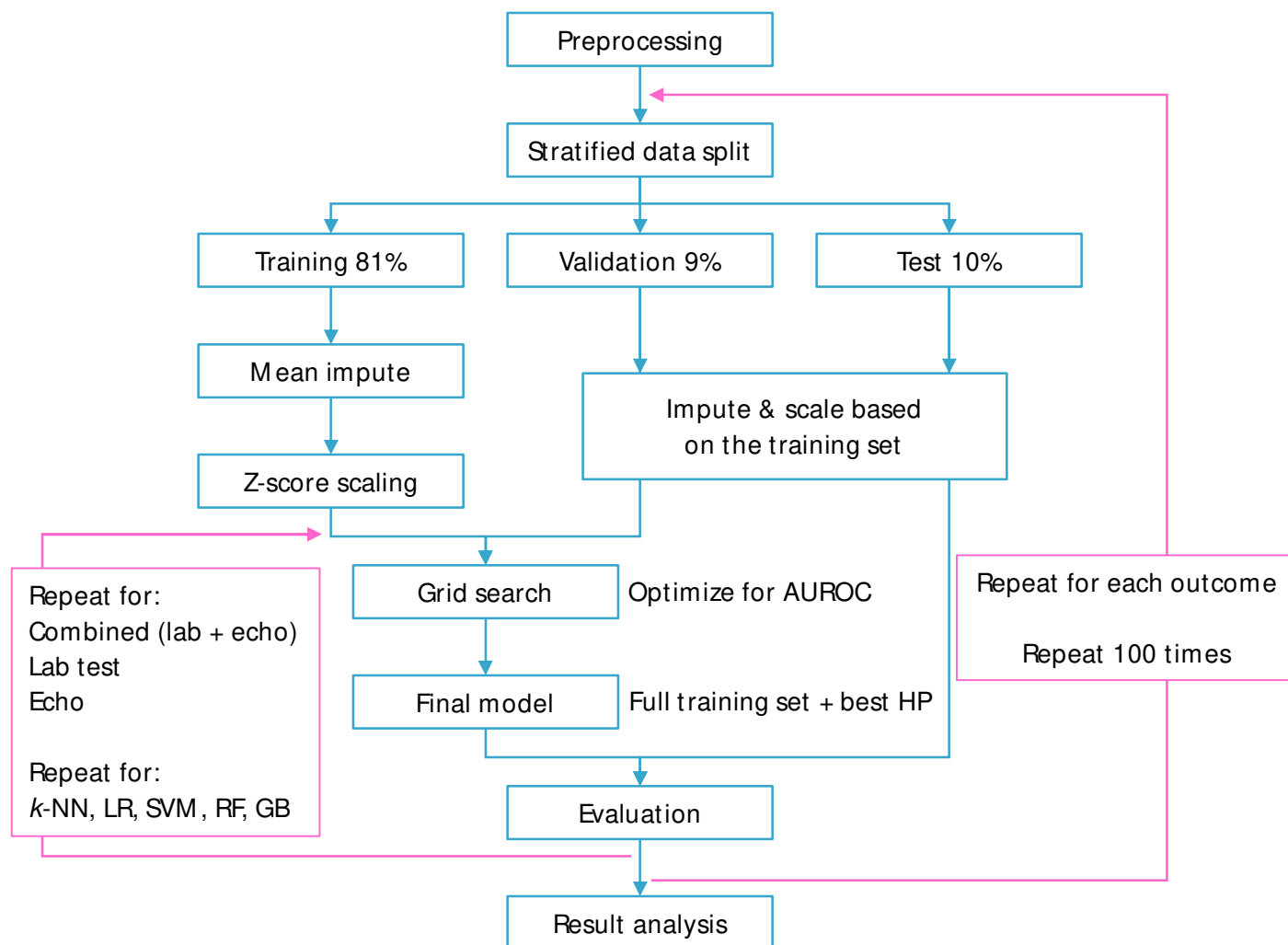
HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction; CTRCD - cancer therapy-related cardiac dysfunction; AST - aspartate aminotransferase; LVEF - left ventricular ejection fraction; LVESVi - left ventricular end-systolic volume index.

Table S8. Predictive variables for the individual outcomes.

Outcome	Variables
HF	Hypertension Peripheral edema Shortness of breath Age Creatinine LVEF Systolic blood pressure End-systolic volume LVESVi
AF	Age
CAD	Sex Family history Tobacco use Hypertension Hyperlipidemia Chest pain Age Sodium LVEF
MI	Hyperlipidemia Age Blood glucose White blood cell AST Diastolic blood pressure
Stroke	Family history Hypertension Hyperlipidemia Fatigue Age Sodium Carbon dioxide Albumin LVEF Heart rate Systolic blood pressure
<i>de novo</i> CTRCD	Sex Family history Hypertension Shortness of breath Fatigue Age LVEF Heart rate Systolic blood pressure

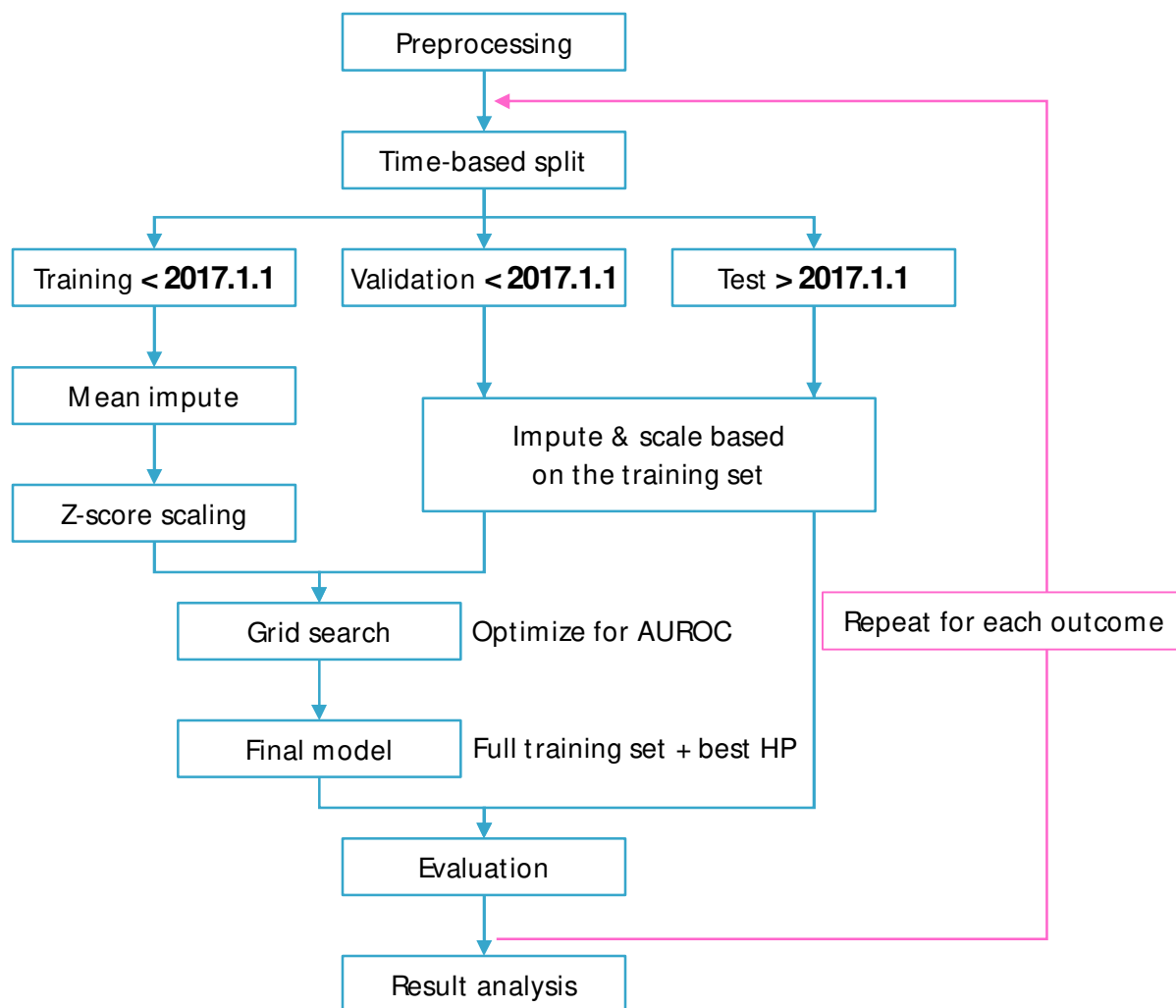
HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction; CTRCD - cancer therapy-related cardiac dysfunction; AST - aspartate aminotransferase; LVEF - left ventricular ejection fraction; LVESVi - left ventricular end-systolic volume index.

Figure S1. Workflow for the classification method and feature set selection and evaluation.



A training-validation-test procedure was performed for 100 times for each outcome. In each iteration, all data were split to training, validation, and test sets. Five methods, *k*-NN, LR, SVM, RF, and GB, and three feature sets, echo only, lab test only, echo and lab test combined were tested. The training set and validation set were used for hyperparameter tuning. Then, they were combined and trained using the optimal hyperparameter set that achieved the highest area under the receiver operating characteristic curve for the validation set. The final model was then used to predict the test set, and the performance of the test set was reported for each outcome.

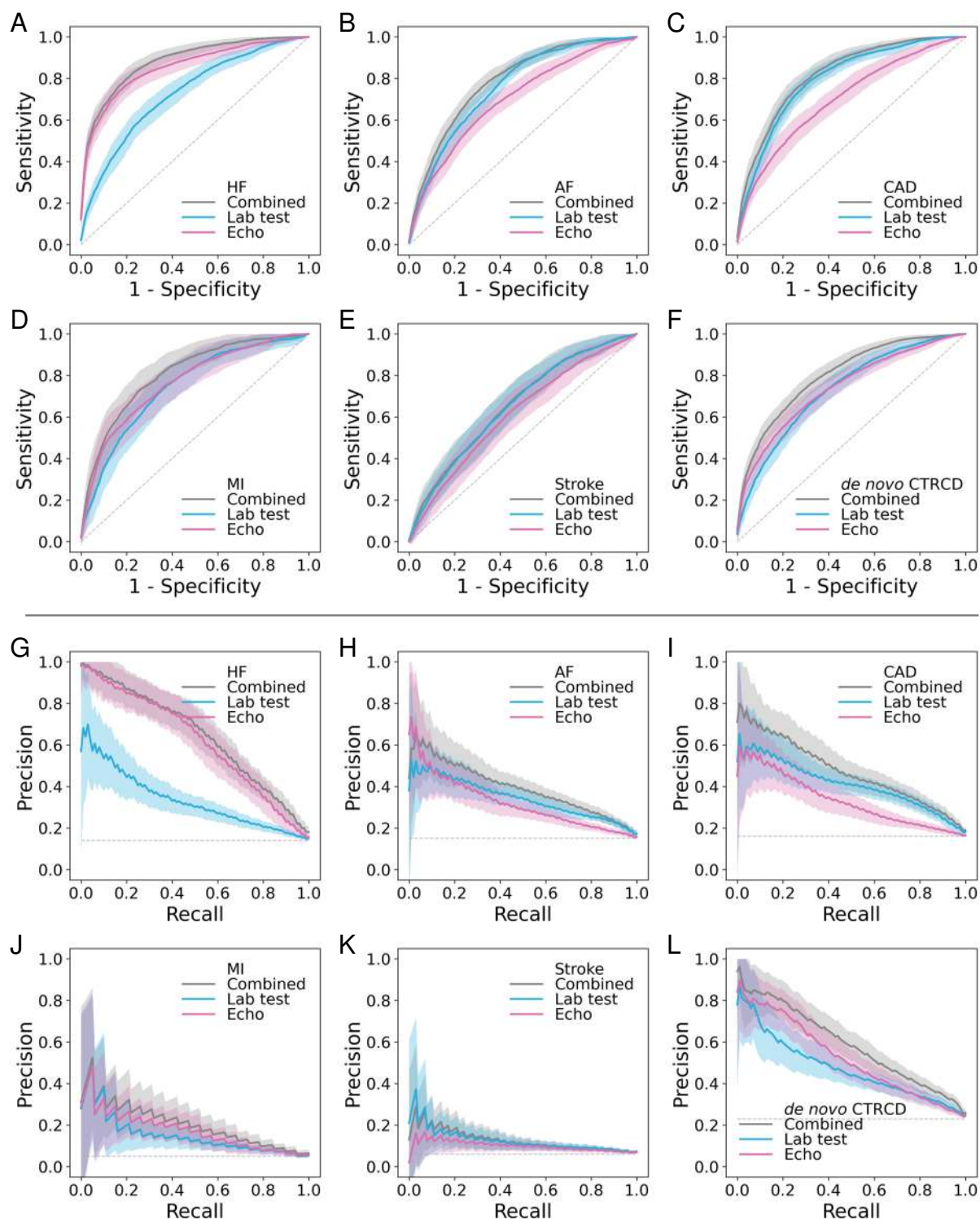
Figure S2. Workflow for the model generalizability test by splitting data chronologically.



All patients were split by the date 2017.1.1. Patients that received cancer therapy before this date were used for model training, and patients that received cancer therapy after this date comprised the test sets for evaluation of the model performances. Logistic regression and combined feature set were used. A hyperparameter tuning was performed to optimize for the area under the receiver operating characteristic curve (AUROC) of the validation set for each outcome. Then, the training

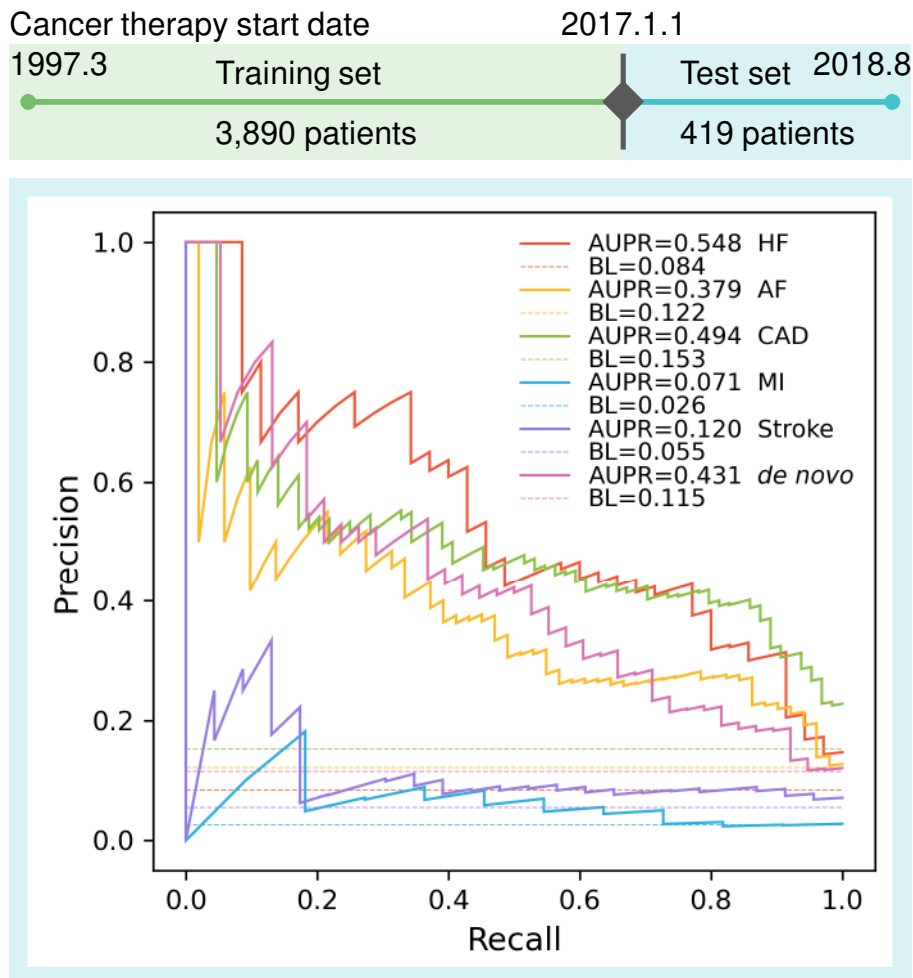
set and validation set were combined and trained using the optimal hyperparameter set that achieved the highest AUROC. The final model was then used to predict the test set, and the performance of the test set was reported for each outcome.

Figure S3. Comparison of the performances of the three feature sets.



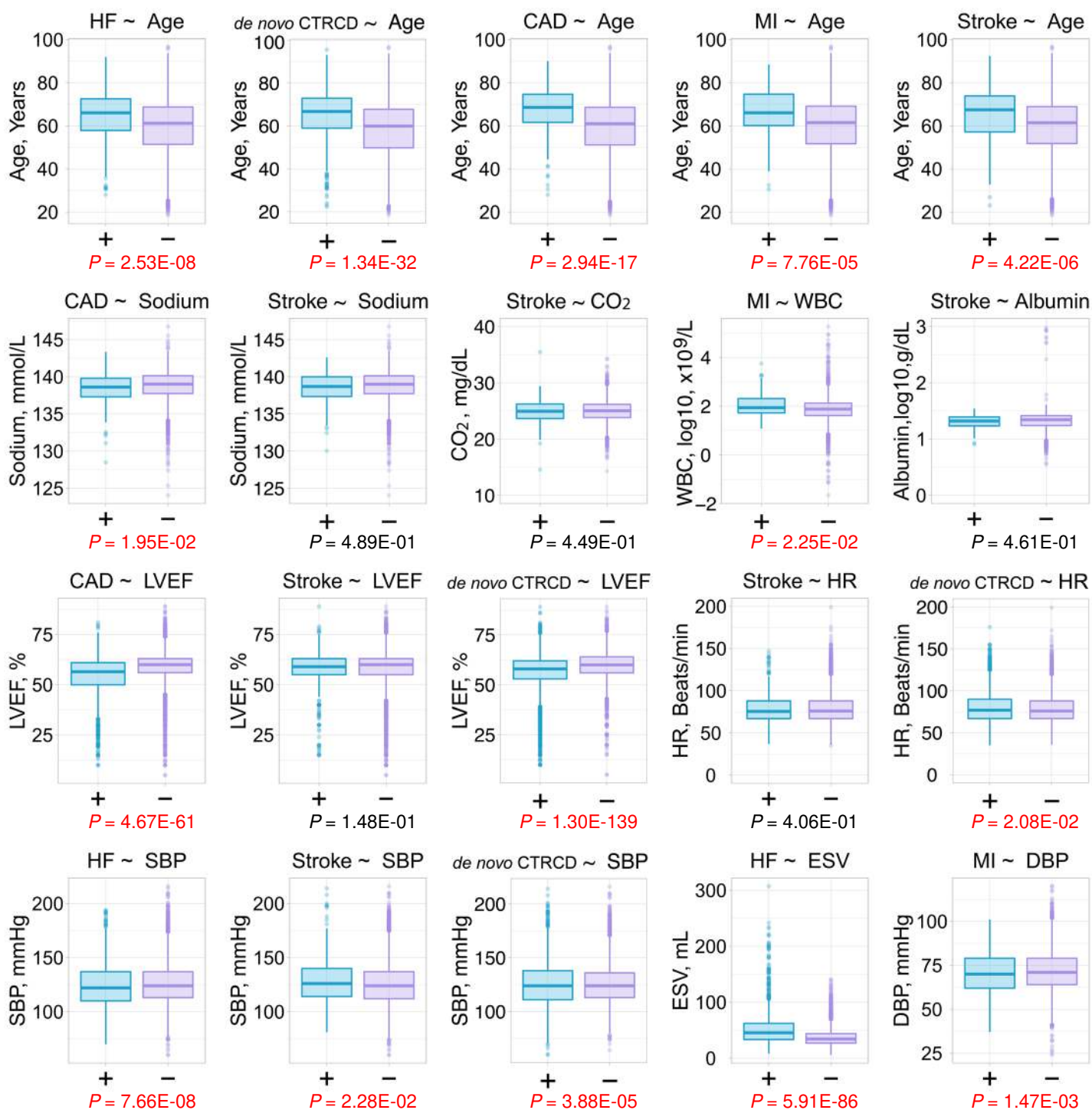
Three feature sets, echo only, lab test only, both sets combined were evaluated based on the test sets performances from 100 iterations using logistic regression. (A-F) receiver operating characteristic and (G-L) precision-recall curves show that combined feature sets achieved better performances than either feature set alone. In addition, for heart failure (HF), myocardial infarction (MI), and *de novo* cancer therapy-related cardiac dysfunction (CTRCD), echo feature set achieved better results than lab test (AUROC, $P = 8.6 \times 10^{-46}$, 2.7×10^{-3} , 3.7×10^{-2} , respectively, two-sided paired sample *t*-test). For atrial fibrillation (AF), coronary artery disease (CAD), and stroke, lab test outperformed echo (AUROC, $P = 4.0 \times 10^{-28}$, 8.3×10^{-44} , 1.1×10^{-9} , respectively, two-sided paired sample *t*-test).

Figure S4. Precision-recall curves of the time-split test sets.



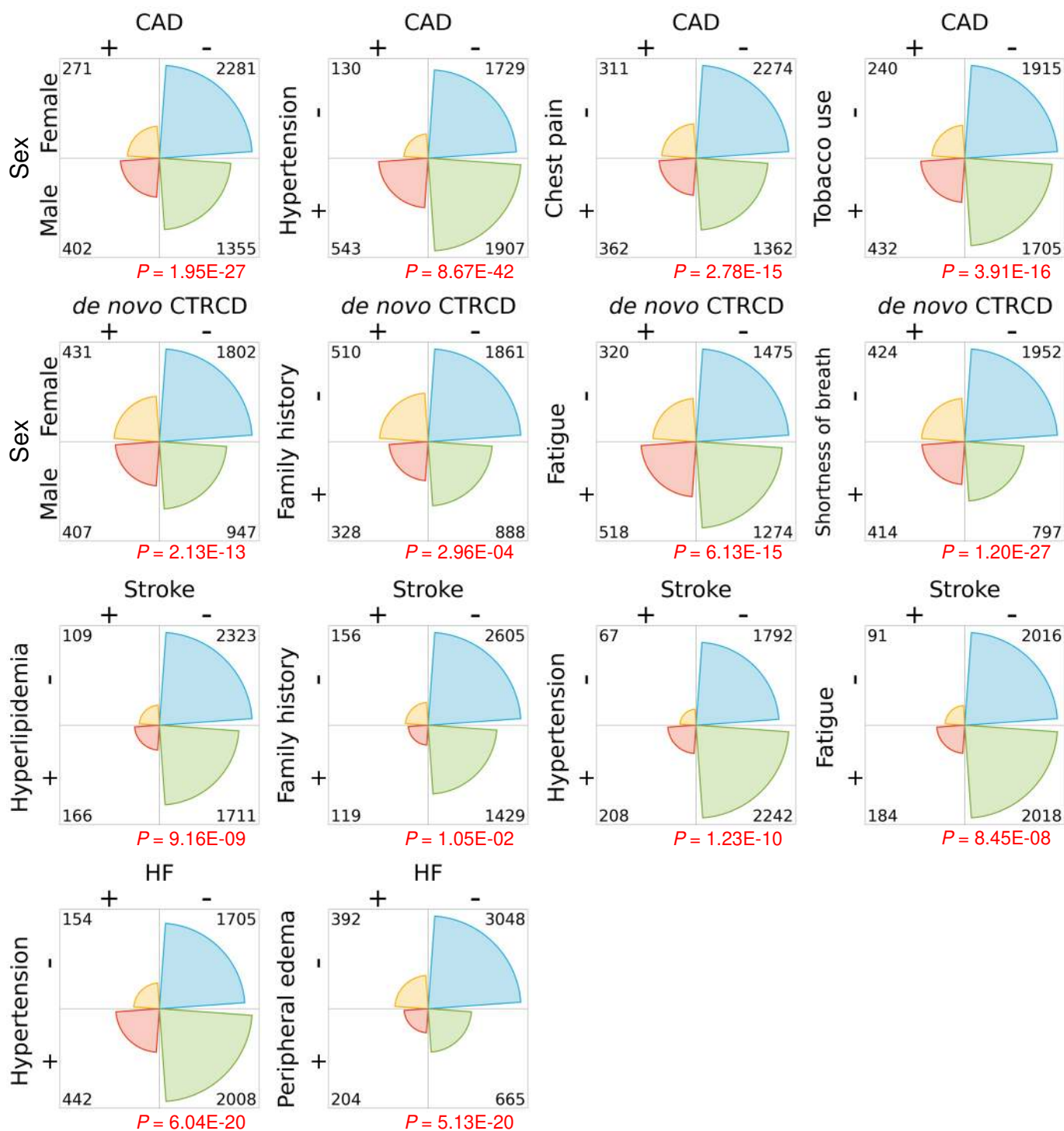
Logistic regression and combined feature sets were used to train the models. Dotted lines indicate the baseline (BL) performances of random classifiers. Our models achieved positive performances in terms of area under the precision-recall curve (AUPR) for all outcomes (all AUPRs are higher than the baselines).

Figure S5. Distributions of the identified continuous clinically relevant variables in the patients.



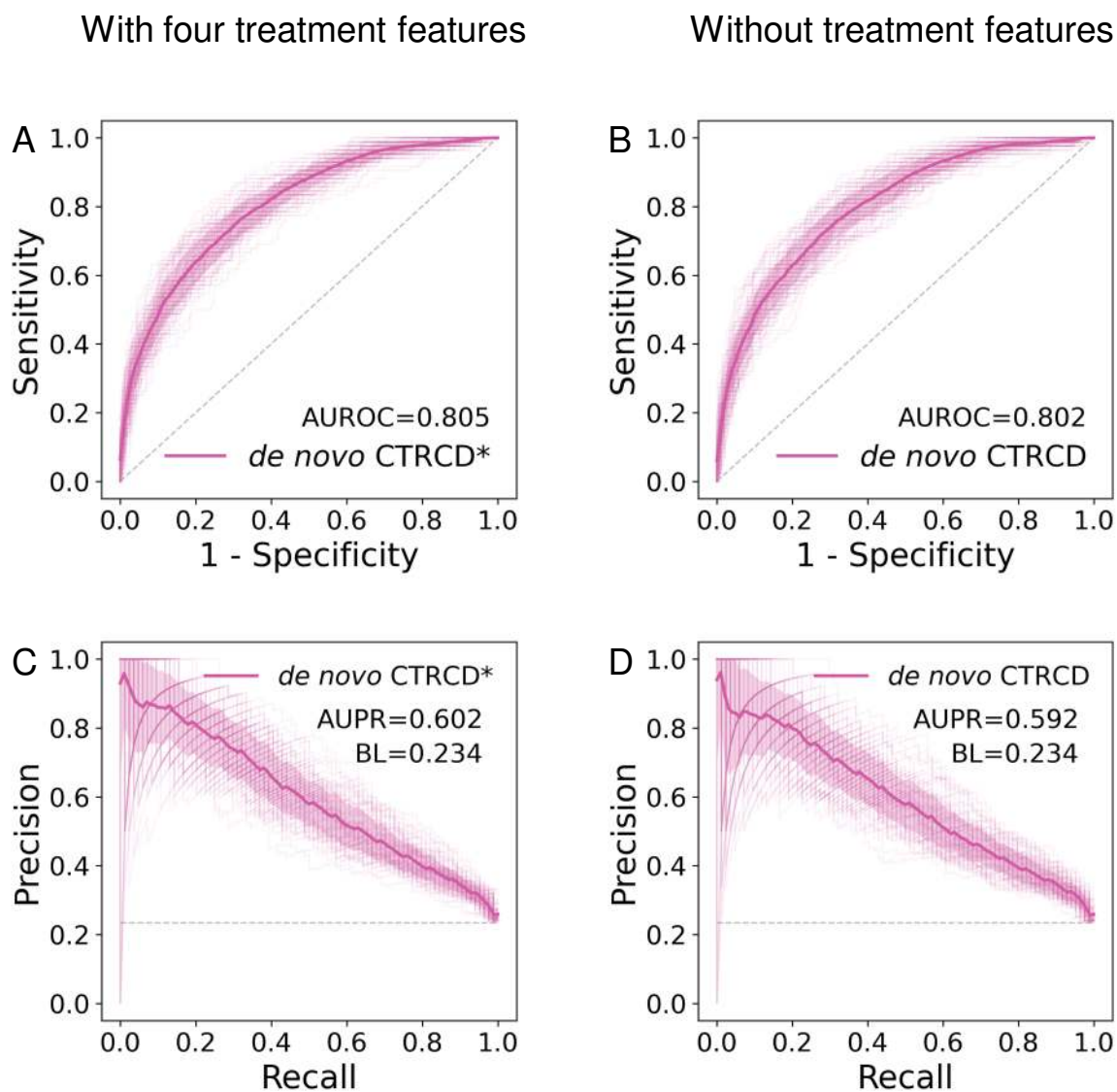
+, patients with the outcome; -, patients without the outcome. $P < 0.05$ are highlighted in red (Kolmogorov-Smirnov test). HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction; CTRCD - cancer therapy-related cardiac dysfunction. WBC - white blood cell; LVEF - left ventricular ejection fraction; HR - heart rate; SBP - systolic blood pressure; DBP - diastolic blood pressure; ESV - end-systolic volume.

Figure S6. Distributions of the identified categorical clinically relevant variables in the patients.



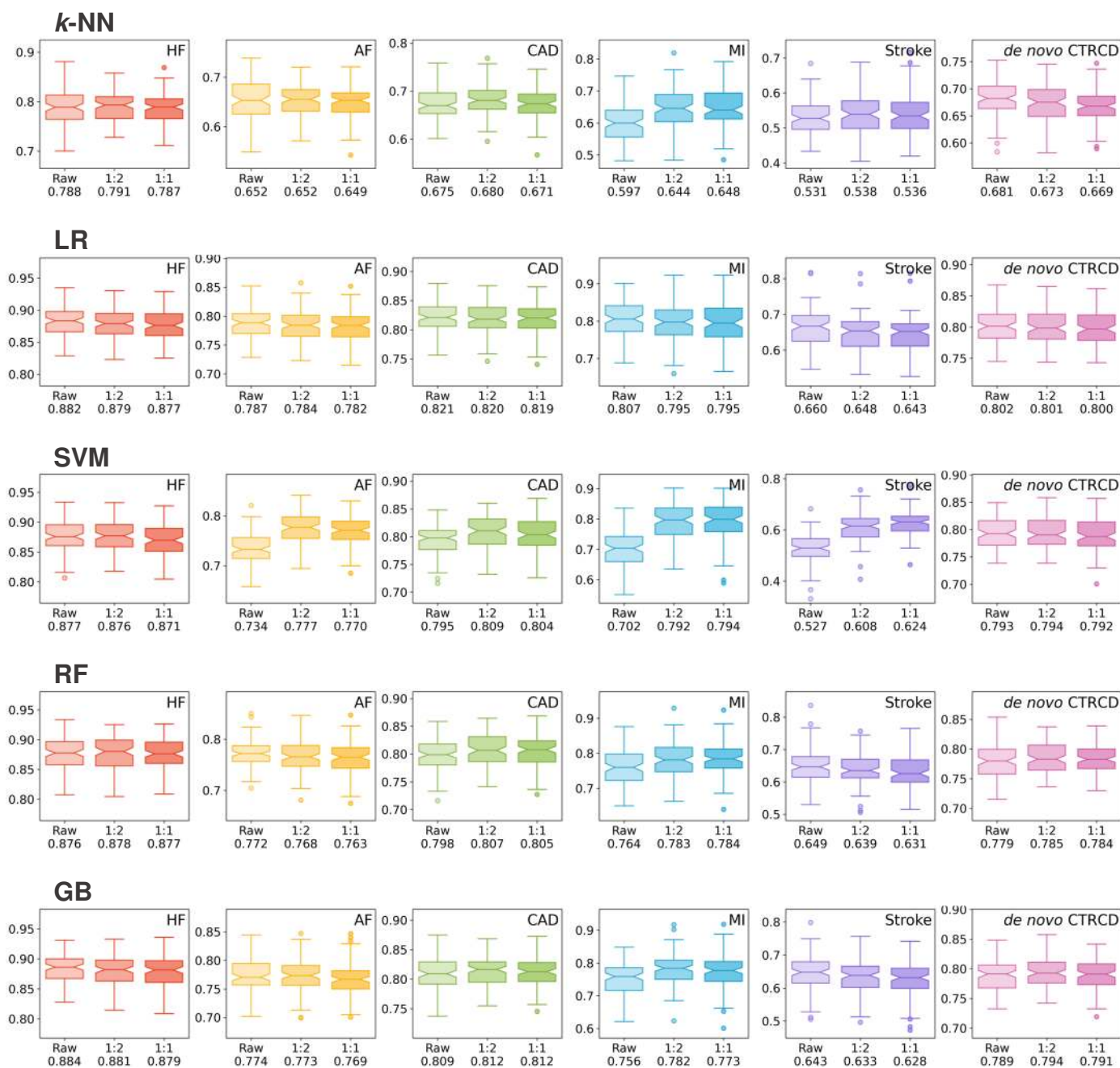
For the outcomes, + indicates patients with the outcome, - indicates patients without the outcome. For the variables, + and - indicates whether patients have this symptom. $P < 0.05$ are highlighted in red (χ^2 test). HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction; CTRCD - cancer therapy-related cardiac dysfunction.

Figure S7. Performance comparisons of the models with (A and C) and without (B and D) the cancer therapy-related features.



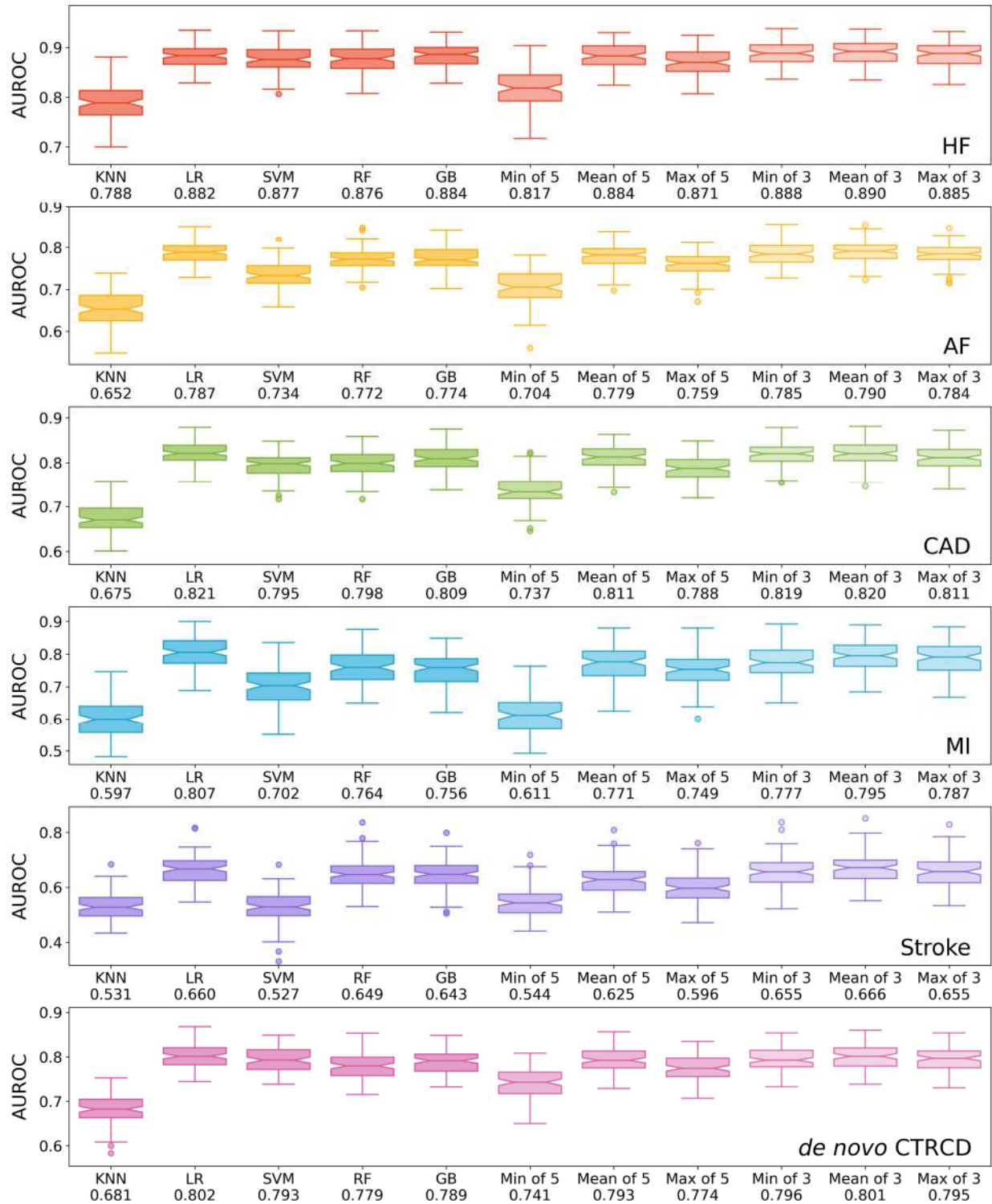
* indicates these models were trained and evaluated with four additional categorical features: chemotherapy, radiation, anthracycline, and trastuzumab usage. Integrating these features marginally improved the model performances (AUROC = 0.805 vs. 0.802, $P > 0.1$, t -test).

Figure S8. Model performances using synthetic minority oversampling technique for the training sets.



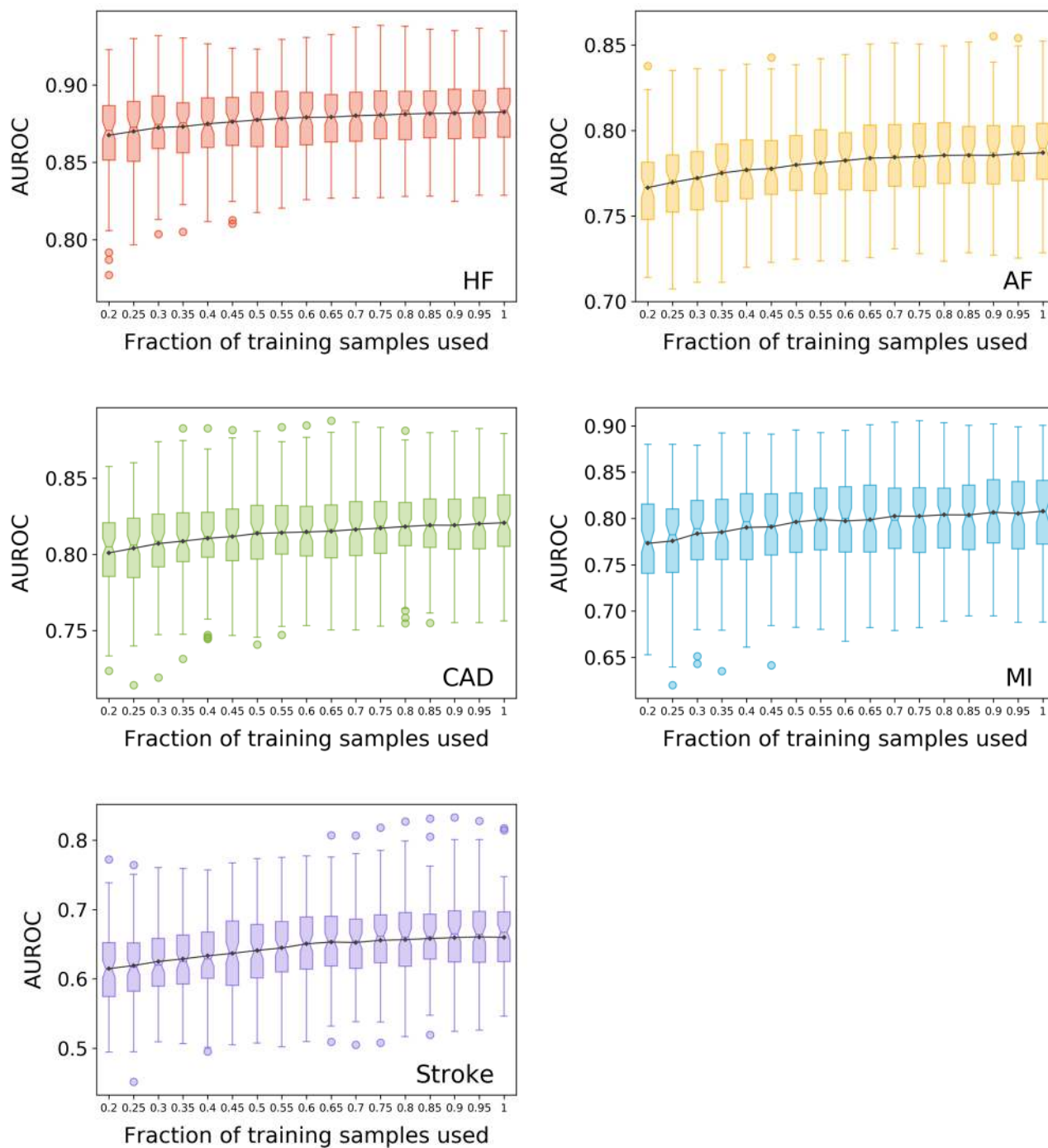
Synthetic minority oversampling technique (SMOTE) was used to generate new training data from existing training data. No improvement was observed for logistic regression. Y axis shows the area under the receiver operating characteristic (AUROC) values of 100 repeats. Raw, original model performances without using SMOTE. 1:2, the minority class was resampled to a ratio of 1:2 to the majority class. 1:1, both classes were balanced. The number below each group shows the average AUROC value. HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction; CTRCD - cancer therapy-related cardiac dysfunction; k-NN - k-nearest neighbors; LR - logistic regression; SVM - support vector machine; RF - random forest; GB - gradient tree boosting.

Figure S9. Performance comparisons of the stacked models.



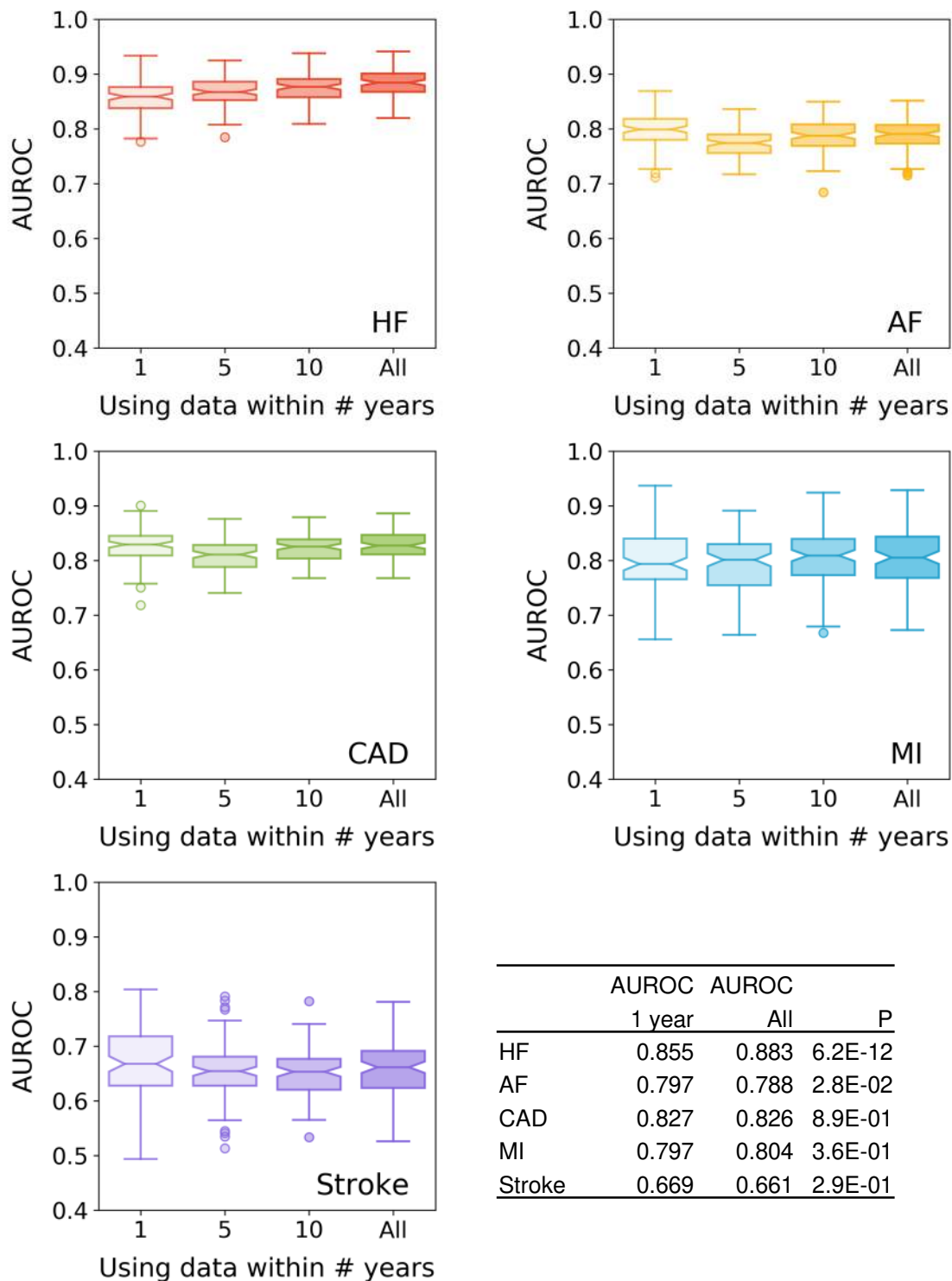
The outputs of the original models of the five algorithms were stacked using three methods: minimum, mean, and maximum. We evaluate the results by stacking all five algorithms (Min of 5, Mean of 5, and Max of 5). In addition, we also excluded *k*-NN and SVM due to their overall lower performances than the other three (Min of 3, Mean of 3, and Max of 3). The number below each group shows the average area under the receiver operating characteristic (AUROC) value of 100 repeats. HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction; CTRCD - cancer therapy-related cardiac dysfunction; *k*-NN - *k*-nearest neighbors; LR - logistic regression; SVM - support vector machine; RF - random forest; GB - gradient tree boosting.

Figure S10. Performance comparisons of the models using different sizes of training set.



Results are based on 100 repeats. In each run, a portion of the training samples was used. Black lines indicate the means. All outcomes benefited from an increasing number of training samples. HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction.

Figure S11. Performance comparisons of the models using variables collected within different time points.



Models performed similarly overall. Slightly improvement may be achieved by experimenting of a model specific variable selection procedure. AUROCs in the tables are the averages. HF - heart failure; AF - atrial fibrillation; CAD - coronary artery disease; MI - myocardial infarction.