Journal of the American Heart Association

ORIGINAL RESEARCH

PrevCardioOncAI: Machine Learning Algorithms for Predicting Cardiovascular Disease in Cancer Survivors

Sherry-Ann Brown, MD, PhD; Michelle Z. Fang, BS; Rodney Sparapani, PhD; Yadi Zhou, PhD; Kristen Osinski, MS; Bradley Taylor, MBA; Duo Yu, PhD; Jeffrey Blessing, PhD; Rushabh Shah, BS; Patrick Collier, MD, PhD; Mehri BagheriMohamadiPour, MSc; Jun Zhang, PhD; Anai Kothari, MD, MS; Gift Echefu, MD; John Rickards, MD; Cameron Otto, BS; Zanele Sanchez; Jessica Olson, PhD; Adelaide Arruda-Olson, MD, PhD; Yee Chung Cheng, MD; Feixiong Cheng, PhD; for the Cardio-Oncology Artificial Intelligence Informatics and Precision Equity, and Patient Similarity Algorithms in the Prevention of Cardiovascular Toxicity Research Team Investigators

BACKGROUND: Cardiovascular disease (CVD) is a leading cause of death in cancer survivors. Predicting CVD risk in this population remains challenging. Risk prediction models that use machine learning algorithms could provide an objective method for accurate risk prediction to facilitate the prevention and management of CVD in cancer survivors. We evaluated previously tested machine learning algorithms and logistic regression (regularized by machine learning methods), in addition to testing and validating newer, more complex machine learning algorithms, for CVD prediction in cancer survivors.

METHODS: This multicenter study used a database of 3835 multiracial cancer survivors with 89 clinical, laboratory, and echocardiographic features over 20 years. Models were trained using repeated random and time-split samples and tested on a separate cohort of 329 patients. Model performance was assessed using the area under the receiver operating characteristic curve.

RESULTS: Regularized logistic regression achieved an area under the receiver operating characteristic curve of 0.845 (heart failure), 0.783 (atrial fibrillation), 0.792 (coronary artery disease), and 0.806 (composite CVD). These are comparable to 0.837 and 0.848 for heart failure, using Bayesian additive regression tree and random forest as more advanced machine learning models, respectively. De novo composite CVD (post-cancer diagnosis) was also predicted with an area under the receiver operating characteristic curve of 0.826 using regularized logistic regression, compared with 0.735 and 0.802 using decision tree and random forest, respectively.

CONCLUSIONS: Regularized logistic regression and advanced machine learning models demonstrated similar predictive performance, with institutional transferability. These tools may support risk stratification and prevention strategies in cardio-oncology using longitudinal data.

REGISTRATION: URL: https://www.clinicaltrials.gov; Unique identifier: NCT05377320.

Key Words: artificial intelligence ■ cancer survivors ■ cardio-oncology ■ cardiotoxicity ■ echocardiography ■ machine learning

More than 18 million individuals in the United States are cancer survivors,[1,2] and this is expected to increase to >22 million by 2030.[3] Cardiologists are tasked with determining which cancer survivors may be at greatest risk for cardiovascular disease (CVD), and initiating cardioprotective

Correspondence to: Sherry-Ann Brown, MD, PhD, Heart Innovation and Equity Research (HIER) Group, 1200 Brickell Avenue, Suite 1950, #1005, Miami, FL 33131, USA. Email: connect@myheartopulence.com

This article was sent to Tochukwu M. Okwuosa, DO, Associate Editor, for review by expert referees, editorial decision, and final disposition.

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.123.030363

For Sources of Funding and Disclosures, see page 16.

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

CLINICAL PERSPECTIVE

What Is New?

- In a large database with 20 years of longitudinal data, regularized logistic regression and advanced machine learning algorithms developed at a tertiary institution accurately predicted cardiovascular disease in cancer survivors.
- These models maintained predictive performance when externally validated at an independent tertiary institution, confirming transferability across health systems.

What Are the Clinical Implications?

- These predictive models could be used to identify cancer survivors at elevated cardiovascular risk and facilitate timely prevention strategies.
- With demonstrated cross-institutional applicability, such tools may support broader implementation of precision prevention in cardio-oncology.

Nonstandard Abbreviations and Acronyms

BART	Bayesian additive regression tree
CatBoost	categorical boosting
F&MCW	Froedtert and the Medical College of Wisconsin
LightGBM	light gradient boosting machine
SHAP	Shapley additive explanations
XGBoost	extreme gradient boosting

medications and imaging surveillance.[4–9] Using current assessments, our ability to predict which patients are at greatest risk is limited. In recent years, several cardio-oncology consensus recommendations have become available for managing cardiovascular risk for cancer survivors.[10–13] Yet, we remain challenged with how best to predict those patients at highest risk for developing CVD and in greatest need of aggressive prevention and monitoring according to guidelines.

The development and use of machine learning artificial intelligence algorithms could help improve the prediction ability of data modeling and associated clinical management. The field of artificial intelligence may be considered as using computer models to perform tasks that would typically require human intelligence. Machine learning algorithms are a type of artificial intelligence algorithms and are used to construct methods that learn from training data to make predictions/decisions without being explicitly programmed to do so. These computer models can take in vast amounts of labeled training data, process the data to identify trends and correlations, and then predict future events based on these insights especially over time. Over the past few years, machine learning has become increasingly prevalent in cardiovascular research, including use in preliminary studies for the prediction of drug-induced cardiovascular complications.[14,15] At the same time, the predictive ability of machine learning algorithms can be compared with more traditional methods such as logistic regression, even when regularization (considered a machine learning method) is applied to logistic regression typically considered to be a statistical modeling technique. Logistic regression, even when not regularized, is a powerful tool that remains popular, due to simplicity, interpretability, and efficacy in predicting outcomes.[16] Comparing machine learning algorithms to more traditional logistic regression has usefulness, given that a recent study demonstrated machine learning algorithms outperformed logistic regression in predicting CVD in that population.[17] The literature remains inconclusive on the efficacy of machine learning algorithms compared with logistic regression, because other studies vary in their results.[18,19] Investigating and comparing the performance of these methods is therefore pertinent.

The accumulation of longitudinal clinical data for oncology patients offers an exciting opportunity for developing predictive models for clinical practice using machine learning compared with logistic regression. Due to interindividual differences in response to cardioprotective or preventive measures, machine learning algorithms may enhance CVD risk prediction that is truly personalized to facilitate the prevention of cancer-related cardiovascular toxicity for cancer survivors. Such artificial intelligence algorithms present a novel opportunity to revolutionize identification of cancer survivors at greatest CVD risk, with subsequent implementation of aggressive CVD preventive efforts. Artificial intelligence, therefore, has the potential to attenuate the CVD deaths that occur in 11% of cancer survivors.[20]

Recently developed/validated machine learning algorithms were used to predict CVD in cancer survivors in a robust database from a single institution.[21] The machine learning algorithms were used to predict heart failure, atrial fibrillation, coronary artery disease, myocardial infarction, and stroke in cancer survivors. However, the algorithms have not been validated on patient data externally to the institution where the algorithms were trained. A recent study indicated that models trained on national data may be more transferable than models trained on local data alone.[22] It is therefore unknown whether the existing machine learning algorithms trained locally to predict CVD in cancer survivors are applicable and transferable to external sites. Additionally, the current conflicting literature on the predictive performance of machine learning

models compared with logistic regression remains a point of interest.

In this study, we further investigated whether supervised machine learning models could accurately predict the risk of developing CVD in cancer survivors, with external applicability and transferability. We used machine learning models to predict 3 different types of cardiovascular outcomes, including heart failure, atrial fibrillation, and coronary artery disease. We used patient data to externally train and validate machine learning algorithms (at a new and different institution) that had been created and previously trained and validated internally at another single institution for predicting and facilitating the prevention of CVD in individuals with a history of cancer. The new data (from the new institution) were tested on the models that were initially developed, trained, and validated at the previous single-center institution. Furthermore, to develop the robustness of our study, we also implemented newer, advanced machine learning algorithms to determine added predictive benefit, and compared the performance of all of the machine learning models with regularized logistic regression.

METHODS

Study Design and Study Population

Informed consent was waived for this retrospective study approved by the Froedtert and the Medical College of Wisconsin (F&MCW) institutional review board. Upon reasonable request from the corresponding author and approval by the institutional review board, all deidentified data used in this study may be made available. Data gathering, management, and analysis in this retrospective review were conducted at F&MCW and Cleveland Clinic, in collaboration with the Milwaukee School of Engineering and the University of Wisconsin–Milwaukee. Our multicenter, multidisciplinary research team of collaborators has prior group publications[21,23,24] and includes specialists in artificial intelligence, cardio-oncology, preventive cardiology, clinical decision aids, shared decision-making, integration of personalized care tools in the electronic health record, and innovation in racial and ethnic minorities. The F&MCW clinical research data warehouse, which houses clinical and demographic information for >1.2 million patients from the Froedtert Health System spanning >20 years (from the early 2000s to the present) and several sites and clinical models across the health system, was queried to identify all cancer survivors aged ≥18 years who have had measurement of their left ventricular ejection fraction within 3 years before cancer diagnosis (Figure 1). The clinical research data warehouse can be queried for specific criteria including, but not limited to, patient demographics, laboratory test results, medication orders, tumor registry, and clinical encounter characteristics. Thus, the clinical research data warehouse was used as a resource to capture patient data from ≈4000 individuals meeting these requirements. Additional data were also obtained directly from the electronic health records to obtain echocardiogram reports dating back to the mid-1990s. Baseline epidemiological findings from this cohort of ≈4000 patients were previously described and provided results from a variety of parameters, some of which are summarized in Table 1.[23]

Deidentified clinical, laboratory, and echocardiographic data (Table 1; 104 features when expanded in detail and with various forms of interrogation) for 4221 patients who underwent quality control and were curated to match the format required for the machine learning algorithms previously described.[21] The machine learning algorithms were developed to aid in assessing the risk of CVD, and the source code is available at https://github.com/ChengF-Lab/CO-ML. The algorithms were built on the evolution of 20-year data and validated on recent 5- to 10-year data for cancer survivors in a robust multiracial tertiary institution database at Cleveland Clinic.[21] Minor modifications were made to the machine learning algorithms with the purpose of optimizing its alignment with the F&MCW patient population and facilitate data input based on the distribution of structured versus unstructured data from F&MCW, a population external to Cleveland Clinic where the models were initially validated. The adaptation of the artificial intelligence algorithms to our F&MCW data was unrelated to modeling cardiovascular risk and focused on the best way to incorporate the format of the new external population data. Adjustments to the model were therefore minimal and only what was necessary to incorporate the F&MCW data. The clinical, laboratory, and echocardiographic data were then incorporated into the machine learning algorithms as input. Information about the development of any CVD, including heart failure, atrial fibrillation, and coronary artery disease, was included as outcomes, regardless of the timing of cancer diagnosis.

We subsequently repeated validation of the algorithms, on a new set of 329 patients from F&MCW, who were not included in the training or test set. Outcomes examined were the same: CVD, heart failure, atrial fibrillation, and coronary artery disease. In addition, for the entire data set we examined de novo CVD, which we define as having developed at least 1 form of CVD (in particular heart failure, atrial fibrillation, or coronary artery disease) after the first cancer diagnosis.

Inclusion Criteria

Inclusion criteria were the following: individuals with a history of cancer, ages 18 years or older, measurement of their left ventricular ejection fraction within 3 years

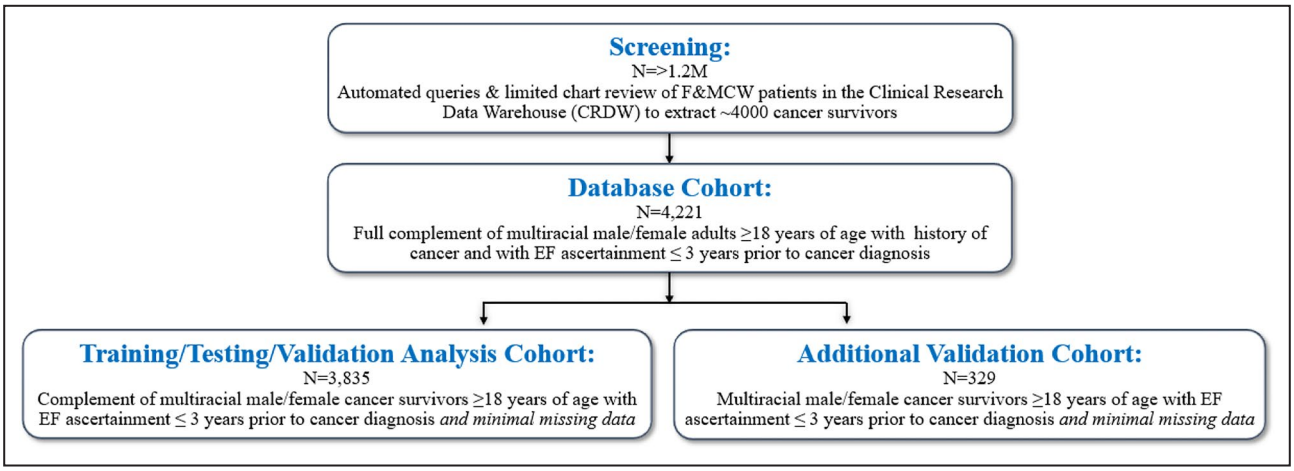

Figure 1. Study population flowchart.
The study population was extracted from a data warehouse of >1.2 million patients in the F&MCW. A database was ascertained (N=4221), based on abstracting all male and female adult patients with a history of cancer and with EF results available within 3 years before cancer diagnosis. After excluding records with substantial missing data, a training, testing, and validation analysis cohort of 3835 patients was determined. An additional 329 patients were also derived as an additional validation set. Of note, 78% of the cohort was non-Hispanic White, 18% non-Hispanic Black, 2% Hispanic, and 2% other/unknown. CRDW indicates clinical research data warehouse; EF, ejection fraction; and F&MCW, Froedtert and Medical College of Wisconsin.

before cancer diagnosis, and from the mid-1990s to 2020 (at the time of initial data querying for this study).

Preparation of Cardiovascular Outcomes Data

CVD outcomes, including heart failure, atrial fibrillation, and coronary artery disease, were extracted using *International Classification of Diseases, Ninth and Tenth Revision* (*ICD-9*, *ICD-10*) diagnosis codes and spot-checked with manual verification. This study included both inpatient and outpatient codes. According to the date of diagnosis of these 5 cardiac events, we classified those diagnosed before cancer therapy as preexisting cardiac events and those diagnosed after as de novo CVD. All variables were collected for each patient using all available data. Each patient had 2 distinct sets of clinical variables: laboratory and echocardiographic. The results of laboratory tests included estimates of glomerular filtration rate, glycated hemoglobin, glucose, calcium, and total protein, among others. Left ventricular ejection fraction, left ventricular end systolic volume index, and left ventricular end diastolic volume index were all measured using echocardiography. Due to the longitudinal nature of the available echocardiographic data, we extracted several features for each echocardiographic variable: the maximum value obtained for a variable from among all measurements obtained over the entire longitudinal follow-up period (all follow-up, ≈20 years); the minimum of all follow-up; the slope of all follow-up, defined as the coefficient from a simple linear regression line fitted to that patient's repeated measurements over time; the maximum increase within 3 months; and the maximum decrease within 3 months. Finally, clinical variables were used as features in the development of machine learning models for cardiovascular outcomes (see complete clinical variable names in Table S1).

Table 1. Characteristics of the Final Study Population Cumulative Over 20 Years (N=3835)

Characteristic	Value
No.	3835
Age, y, median [interquartile range]	70 [61–79]
Sex, n (%)	
Women	1734 (45.2)
Men	2101 (54.8)
LDL (mg/dL), mean±SD	92.4±31.4
HDL (mg/dL), mean±SD	49.7±15.8
Minimum left ventricle ejection fraction, mean±SD	51.7±12.4
Diabetes, n (%)	997 (26.0)
Hypertension, n (%)	3091 (80.6)
Cardiovascular disease, n (%)	2521 (65.7)
Heart failure, n (%)	1831 (47.7)
Atrial fibrillation, n (%)	939 (24.5)
Coronary artery disease, n (%)	1530 (39.9)

HDL indicates high-density lipoprotein; and LDL, low-density lipoprotein.

Regularized Logistic Regression for Risk Prediction

In our work, the regularized logistic regression model was used. Logistic regression as a statistical method estimates the probability of a given observation belonging to a particular category by modeling the log-odds of the outcome as a linear combination of predictor

Table 2. Machine Learning (Artificial Intelligence) Algorithm Variables

Variable
Laboratory test (including demographics)
Sex
Race
Tobacco use
Diabetes
Hyperlipidemia
Orthopnea
Shortness of breath
Age
Estimated glomerular filtration rate
Hematocrit
Mean corpuscular volume
Blood glucose
Total protein
Potassium
Carbon dioxide
Platelet
Alanine aminotransferase
Albumin
Bilirubin
Alcohol use
Peripheral edema
Fatigue
Red blood cell
Mean corpuscular hemoglobin
Sodium
White blood cell
Aspartate aminotransferase
Family history
Hypertension
Chest pain
Body mass index
Mean corpuscular hemoglobin concentration
Calcium
Chloride
Creatinine
Alkaline phosphatase
Echocardiographic
Heart rate
Systolic blood pressure
End-diastolic volume
Left ventricular end-diastolic volume index
Diastolic blood pressure
Left ventricular end-systolic volume index
Left ventricular ejection fraction

Used with permission.[20]

variables.[25] When working with a large number of potentially correlated predictors, regularization can help reduce overfitting and improve model applicability and transferability. Although standard logistic regression does not include a mechanism for penalizing model complexity, regularized variants introduce a penalty during training that shrinks coefficient magnitudes.[26]

In our model, we used ridge (L2 penalty) regularization, which penalizes the sum of the squared regression coefficients. This technique is particularly useful when many predictors may contribute to an outcome and where retaining all variables is desired. Unlike lasso (L1) regularization, which can shrink some coefficients to 0 and thereby perform variable selection, ridge regression retains all predictors in the model by shrinking their coefficients toward 0.[27] In logistic regression, these coefficients are interpreted as log-odds ratios, reflecting the change in the log-odds of the outcome per unit increase in the predictor, holding other variables constant.[25] We preferred ridge over lasso to allow all clinical features to contribute to risk prediction while mitigating overfitting due to multicollinearity.

Advanced Machine Learning Algorithms for Risk Prediction

Various machine learning algorithms were studied in this work, including the decision tree classifier method. This method was chosen due to intrinsic interpretability. A decision tree can explicitly highlight variables and their cutoff values that are used in splitting. The decision tree algorithm makes predictions by splitting the data based on feature values that build a tree-like model of decisions from the training data. Each tree node represents a question on the features to determine which path the sample should follow to ultimately reach the end of the tree (an outcome). Ultimately, the model predicts the target variable by traversing from root to leaf, making decisions at each tree node based on feature values. Consequently, the decision tree model was used, recognizing that the advantage of interpretability of the decision tree model is often traded with the disadvantage of a lack of transferability.

Regularized logistic regression models the log-odds of the outcome as a linear combination of predictor variables, assuming additivity on the logistic scale.[28] Although logistic regression can accommodate nonlinear relationships and interactions when appropriate terms (such as splines, polynomial expansions, or interaction terms) are explicitly modeled,[25] our implementation used only the original variables without such transformations. This modeling choice was made to preserve interpretability and reduce the risk of overfitting in high-dimensional feature space but may have

limited the model's capacity to capture complex associations in the data. In contrast, nonlinear machine learning models, such as random forest and Bayesian additive regression tree (BART) are capable of automatically modeling such relationships without the need for manual specification of nonlinear terms.[29]

To improve transferability, a random forest machine learning algorithm was also used. The random forest model is an ensemble method that combines many decorrelated decision trees into 1 model.[30] The integration of multiple decision trees improves the transferability of the predictive results. Of note, although regularized logistic regression is a linear model (because the algorithm assumes a linear relationship between the features and the outcome), conversely, the random forest model is nonlinear. Nonlinear models possess the potential capacity to capture complex relationships between features and the target variable that may not be linear. However, use of nonlinear models, such as random forest, tends to decrease interpretability. This necessitates additional methods of interpretation. We have therefore used Shapley additive explanations (SHAP) bar plots and beeswarm plots to aid in visualization of the most influential variables for predictions using the random forest algorithm (Figure S1).

To further develop the robustness of our investigation of the predictive capacity of machine learning algorithms, we used 4 additional advanced machine learning models for our analysis. These algorithms include BART and 3 gradient boosting models, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and categorical boosting (CatBoost). The BART model is an ensemble machine learning method that combines multiple decision trees and analyzes these sequentially (boosting) and independently (bagging) in the tree growing process. This model is distinct from the random forest algorithm, which only uses bagging techniques. In addition, the BART model is specifically framed within a Bayesian framework, which provides not only a single averaged prediction (as done in random forest), but rather a distribution of predictions that reflects the uncertainty of the model parameters. This enables the model to better handle variability and uncertainty. The gradient boosting machine learning models focus on enhancing the predictive performance of models by sequentially combining decision trees.[31] Gradient boosting begins by comparing the predictions of a single decision tree to the actual values, computing the residual errors. A new model is then trained to predict these residual errors. This model is added to the ensemble to adjust the previous predictions. This cycle repeats until no further improvement results. The final boosted value results from sequential error correction in the decision trees. XGBoost, LightGBM, and CatBoost are 3 efficient implementations of these gradient boosting algorithms.

Model Workflow

Figure S2 shows the workflow for the analysis of all outcomes. To test for future applicability and transferability, the data were first time-split using a cancer diagnosis date cutoff of January 1, 2020, producing a 9:1 ratio of training data to test data. To find the optimal hyperparameters, the subset of training data was again randomly split using a 9:1 ratio for 20 iterations. The models were trained and validated using each combination of the hyperparameters in Tables S2 and S3 through a grid search. The hyperparameter grid search for logistic regression included a coefficient for regularization to control for overfitting. The area under the receiver operating curve (AUROC) was used to evaluate the prediction performance of a classification model. The set of hyperparameters producing the largest AUROC, averaged across the 20 iterations on each validation set, was chosen for the final models. We then combined all data from before our date cutoff (January 1, 2020) into 1 large training set to train using the determined hyperparameters. We tested those models on the time-split test set (data on or after January 1, 2020). Additionally, we used the additional 329 patient validation set as a second test to evaluate transferability of our models.

For the de novo CVD analysis, the data were not time-split, due to lower numbers of outcomes in recent years. Instead, the data were split into training (81%), validation (9%), and test (10%) sets for 20 iterations. In each iteration, a grid search was performed using each combination of hyperparameters to train the models on the training set and test on the validation set (Table S2). The set of hyperparameters producing the greatest AUROC on the validation set was then used when training the models on the combined training and validation sets. Finally, the models were tested on the test set, with performance measured using AUROC (Figure S3).

Clinically Relevant Variables

To determine statistically robust clinical predictors for de novo CVD, we used previously defined criteria to assess the coefficients of the logistic regression model.[21] To replicate the methodology of this prior study, the logistic regression models were fitted for an additional 80 iterations for 100 iterations total to assess coefficient stability.[32] We identified statistically consistent clinical variables using 2 criteria: (1) the absolute coefficient of variation (the ratio of SD and mean) for the logistic regression coefficient was low, ensuring that the coefficient fluctuated little over the 100 repetitions; (2) the absolute associated coefficient was

high in comparison with the extremum coefficient for that outcome (relative coefficient). According to these criteria, the magnitude of the coefficient of variation for a feature must be <0.5 and the relative coefficient must be >0.3 to identify the statistically consistent predictors.[21] Although these variables were statistically stable, we note that this does not imply clinical modifiability, causal relevance, or direct clinical usefulness.

The random forest algorithm is not intrinsically interpretable. To understand the variables influential in this algorithm's predictions, we used SHAP plots to highlight the top 5 most impactful features on the model outcome. SHAP has been used in a variety of studies to add interpretability to black box machine learning algorithms.[33–36] The SHAP value is derived from cooperative game theory to calculate each feature's average marginal contribution toward the model's prediction. For every patient, combinations (coalition) of features are generated. The difference between a combination of features that includes a particular feature and combination that excludes particular features produces a marginal contribution of that feature toward the model's prediction. This marginal contribution is averaged over all coalition of features, per feature, to understand which features were most impactful on the model's prediction.

Statistical Analysis

We excluded patients who were missing data in >60% of features and excluded features with >80% missing values. Missing values were imputed using most frequent imputation, because this method produced the strongest model performance during the validation process, and continuous variables were standardized using *Z* scores. To test for future applicability and transferability, the data were time-split using a cancer diagnosis date cutoff of January 1, 2020, producing a 9:1 ratio of training data to test data. To evaluate the performance of machine learning models, we used the AUROC, computed using the metrics.roc_auc_score and metrics.average_precision_score functions from the scikit-learn Python package.

To assess whether differences in model discrimination were statistically significant, we conducted pairwise comparisons of AUROC using the DeLong test.[37,38] This nonparametric method uses a 2-sample hypothesis test to compare receiver operating characteristic curves generated from predictions on the same data set. Comparisons were performed among regularized logistic regression and each advanced machine learning model (BART, random forest, CatBoost, XGBoost, and LightGBM) for each of the 4 clinical outcomes: composite CVD, heart failure, atrial fibrillation, and coronary artery disease. These statistical analyses were conducted in R.

Table 3. Machine Learning Predictive Performance (Area Under the Receiver Operating Characteristic Curve) of Previously Tested Algorithms and Regularized Logistic Regression on Testing Set Using Python

Model	CVD	HF	AFib	CAD
Logistic regression	0.80	0.85	0.78	0.79
Decision tree	0.74	0.74	0.76	0.73
Random forest	0.79	0.85	0.75	0.79

AFib indicates atrial fibrillation; CAD, coronary artery disease; CVD, cardiovascular disease; and HF, heart failure.

Software Packages and Codes

The initial model performance evaluations (Tables 3 and 4) were performed using Python 3.9.12. Classification models were done using the package scikit-learn 1.0.2. SHAP plots were created using the SHAP package 0.41.0. The nomograms were created in R version 4.3.2 using the RMS package. Subsequent modeling was also pursued using R to incorporate the more advanced machine learning algorithms (Tables 5 and 6), as well as DeLong test analyses (Tables S4 through S8). The source codes for this machine learning study are available at https://github.com/ChengF-Lab/CO-ML and https://github.com/rsparapa/MLcompare. Minor modifications were made to the machine learning algorithms to accommodate the new F&MCW data before the training and validation process.

RESULTS

Epidemiology

In this study, we built a large, longitudinal cardio-oncology cohort with ≈4000 oncology patients collected from our F&MCW institutional electronic medical record database (Figure 1). The epidemiological findings of this cohort were previously published.[23] Briefly, in the longitudinal cohort, 78% of participants were non-Hispanic White, 18% were non-Hispanic Black, 2% were Hispanic, 1% were other races, and 1% were unknown.

The majority of cancers represented were breast cancer (11%), lung cancer (16%), hematologic cancers (16%), and prostate cancer (9%). In the total cohort

Table 4. Machine Learning Predictive Performance (Area Under the Receiver Operating Characteristic Curve) of Previously Tested Algorithms and Regularized Logistic Regression on Patient Validation Set Using Python

Model	CVD	HF	AFib	CAD
Logistic regression	0.71	0.68	0.83	0.64
Decision tree	0.73	0.71	0.79	0.71
Random forest	0.76	0.75	0.76	0.68

AFib indicates atrial fibrillation; CAD, coronary artery disease; CVD, cardiovascular disease; and HF, heart failure.

Table 5. Out-of-Sample Predictive Performance of Logistic Regression and Newer More Advanced Machine Learning Algorithms on the Testing Set (N=360) Extracted From the Total Sample Size Used After Removing Samples With Substantial Missing Values (N=3835)

Model	CVD	HF	AFib	CAD
Logistic regression	0.795 (0.740–0.837)	0.844 (0.778–0.886)	0.757 (0.673–0.836)	0.790 (0.734–0.840)
Decision tree	0.610 (0.553–0.664)	0.634 (0.556–0.699)	0.598 (0.503–0.694)	0.655 (0.593–0.721)
Random forest	0.800 (0.751–0.839)	0.848 (0.794–0.890)	0.759 (0.680–0.842)	0.778 (0.719–0.836)
BART	0.789 (0.731–0.830)	0.837 (0.773–0.887)	0.756 (0.666–0.828)	0.787 (0.732–0.838)
XGBoost	0.751 (0.703–0.794)	0.769 (0.706–0.820)	0.667 (0.594–0.740)	0.711 (0.653–0.779)
LightGBM	0.753 (0.704–0.800)	0.798 (0.744–0.849)	0.720 (0.638–0.816)	0.748 (0.688–0.807)
CatBoost	0.782 (0.735–0.821)	0.841 (0.777–0.885)	0.759 (0.684–0.835)	0.754 (0.692–0.808)

Data are presented as area under the receiver operating characteristic curve (95% CI). AFib indicates atrial fibrillation; BART, Bayesian additive regression tree; CAD, coronary artery disease; CatBoost, categorical boosting; CVD, cardiovascular disease; HF, heart failure; LightGBM, light gradient boosting machine; and XGBoost, extreme gradient boosting.

(n=4626), 1845 (40%) patients received any pharmacologic cancer therapy, 980 (21%) patients received alkylating agents, 412 (9%) patients received various antibodies, 57 (1%) patients received human epidermal growth factor receptor 2 (HER2) agents, 655 (14%) patients received endocrine therapies, 275 (6%) patients received enzyme inhibitors, 667 (14%) patients received mitotic inhibitors, and 1032 (22%) patients received treatments from other cancer drug classes. Radiation therapy was administered to 564 (25%) people in the available cohort (n=2300; 2015–2020) from the MOSAIQ radiation oncology database. As expected, cardiomyopathy and cardiomegaly tended to be noted more frequently in those who received pharmacologic cancer treatment, including antineoplastic antibiotics (anthracyclines) and antibodies (including trastuzumab).

The overall cohort at baseline (n=4626) had (primarily diastolic) heart failure in 2253 (49%) people, cardiomyopathy in 459 (10%) people, atrial fibrillation in 1027 (22%), coronary artery disease in 2037 (44%) people, ischemic or hemorrhagic stroke in 924 (20%) people, peripheral artery disease in 1265 (27%) people, and lung cancer in 1092 (24%) people. In the years following cancer diagnosis in the cohort, new systolic or diastolic heart failure developed in 573 (13%) people, new cardiomyopathy in 352 (8%) people, new atrial fibrillation in 487 (11%) people, new coronary artery disease in 632 (14%) people, new ischemic or hemorrhagic strokes in 387 (8%) people, new peripheral artery disease in 602 (13%) people, new diabetes in 351 (8%) people, new hypertension in 420 (9%) people, and new hyperlipidemia in 485 (11%) people.

Predictive Analytics

Subsequently, excluding for missing values produced a final analytic data set of 3835 patients and 89 features (eg, left ventricular ejection fraction, creatinine). Median age was 70 years (interquartile range, 61–79 years), with roughly an even split between men (54.8%) and women (45.2%). CVD was noted in 2521 individuals, with heart failure in 1831 individuals, coronary artery disease in 1530 individuals, and atrial fibrillation in 939 individuals (Table 1).

For regularized logistic regression, the mean AUROCs were 0.795 for CVD, 0.774 for heart failure, 0.709 for atrial fibrillation, and 0.762 for coronary artery disease. We then examined the performance of more advanced machine learning models in predicting these

Table 6. Out-of-Sample Predictive Performance of Logistic Regression and Newer More Advanced Machine Learning Algorithms on the Validation Set (N=329)

Model	CVD	HF	AFib	CAD
Logistic regression	0.735 (0.687–0.786)	0.704 (0.632–0.778)	0.781 (0.687–0.861)	0.710 (0.628–0.783)
Decision tree	0.685 (0.617–0.752)	0.640 (0.551–0.729)	0.751 (0.600–0.901)	0.591 (0.490–0.692)
Random forest	0.736 (0.681–0.778)	0.723 (0.651–0.783)	0.781 (0.688–0.880)	0.690 (0.608–0.763)
BART	0.741 (0.684–0.790)	0.751 (0.680–0.822)	0.799 (0.714–0.877)	0.679 (0.603–0.752)
XGBoost	0.731 (0.666–0.771)	0.663 (0.544–0.702)	0.644 (0.509–0.739)	0.637 (0.522–0.707)
LightGBM	0.700 (0.629–0.748)	0.666 (0.535–0.690)	0.833 (0.755–0.912)	0.660 (0.570–0.735)
CatBoost	0.721 (0.663–0.778)	0.712 (0.636–0.787)	0.759 (0.597–0.920)	0.668 (0.586–0.749)

Data are presented as area under the receiver operating characteristic curve (95% CI). AFib indicates atrial fibrillation; BART, Bayesian additive regression tree; CAD, coronary artery disease; CatBoost, categorical boosting; CVD, cardiovascular disease; HF, heart failure; LightGBM, light gradient boosting machine; and XGBoost, extreme gradient boosting.

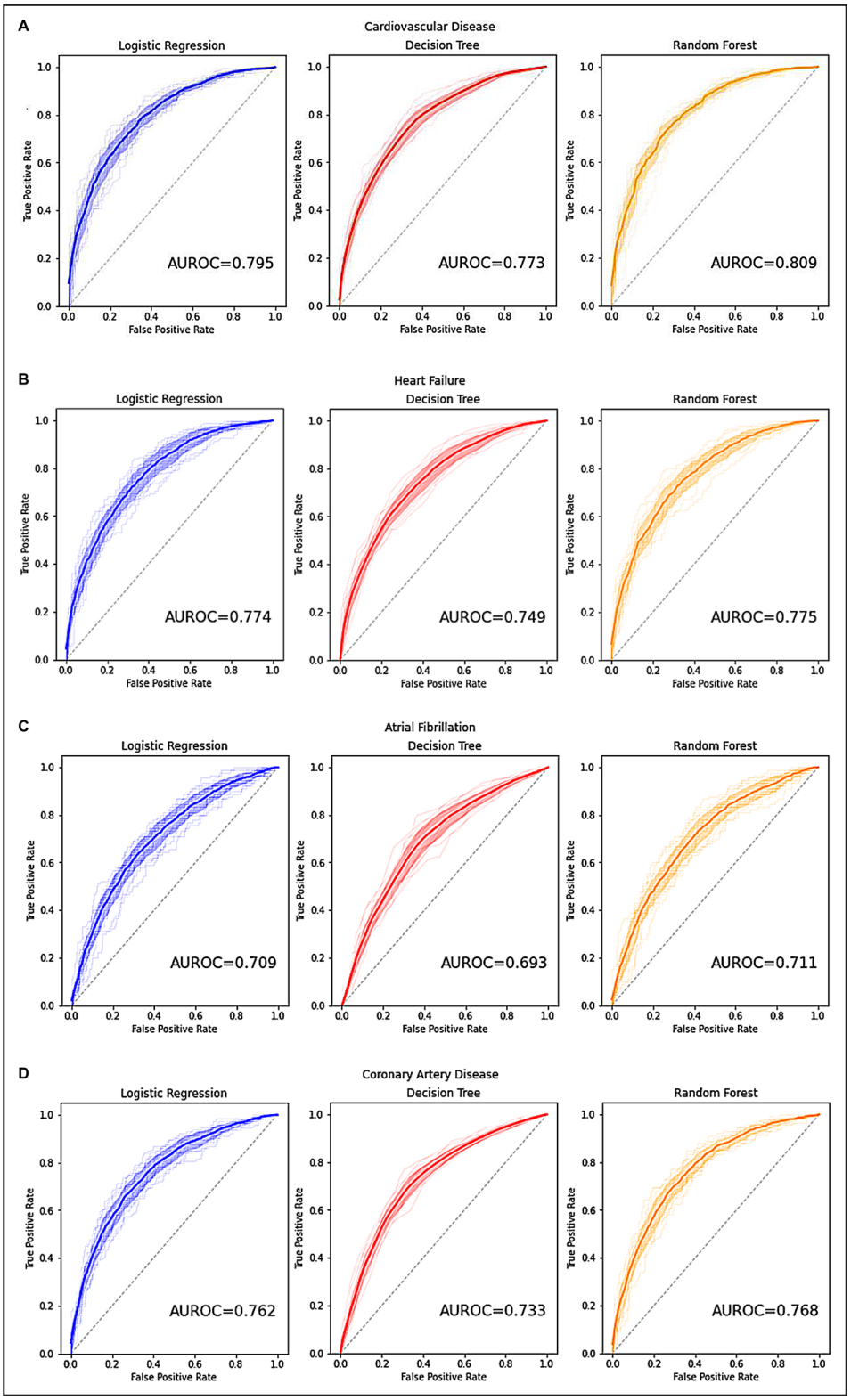

Figure 2. ROC curves for the regularized logistic regression, decision tree, and random forest algorithms during the 20-iteration validation process (N=3835).
Outcomes examined are cardiovascular disease (composite of heart failure and atrial fibrillation and coronary artery disease) (**A**), heart failure (**B**), atrial fibrillation (**C**), and coronary artery disease (**D**). For each iteration, the ROC of the model that produced the highest AUROC is shown. AUROC indicates area under the receiver operating characteristic curve; and ROC, receiver operating characteristic.

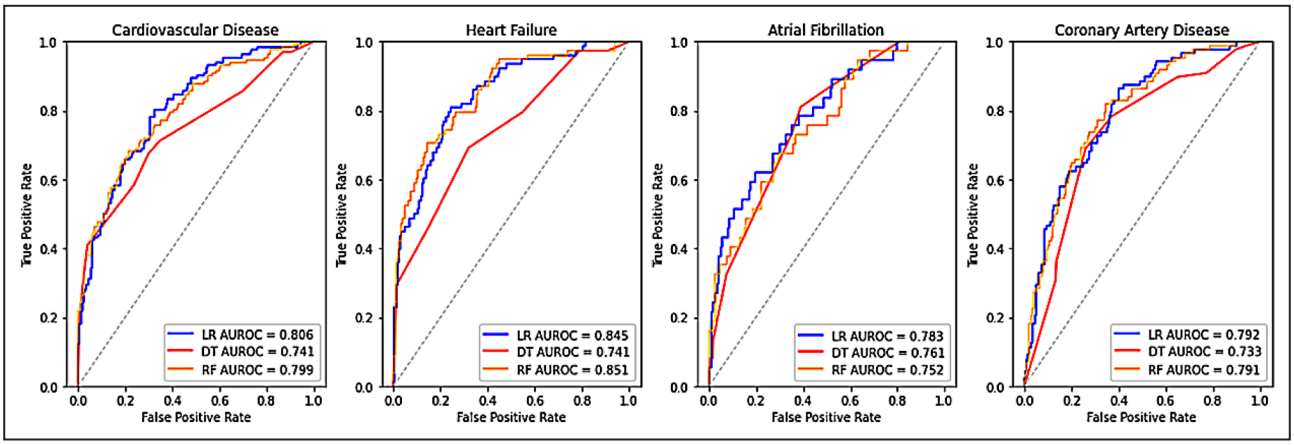

Figure 3. Receiver operating characteristic curves for the regularized LR, DT, and RF algorithms when tested on the time-split test set (N=3835).

Outcomes examined are cardiovascular disease, heart failure, atrial fibrillation, and coronary artery disease. AUROC values are shown at the bottom of the figure. Optimal hyperparameters determined for each model are shown in Tables S1 and S2 in the Supplemental Material. AUROC indicates area under the receiver operating characteristic curve; DT, decision tree; LR, logistic regression; and RF, random forest.

CVD outcomes during the training/validation process. Figure 2 shows the average AUROCs produced during the 20 iterations of the validation process for CVD outcomes regardless of the timing of cancer diagnosis. The decision tree mean AUROCs were 0.773 for CVD, 0.749 for heart failure, 0.693 for atrial fibrillation, and 0.733 for coronary artery disease.

Figure S4 shows the decision trees produced for CVD, heart failure, atrial fibrillation, and coronary artery disease for the training set using the optimal hyperparameters. Table S1 contains the full names for the data variables displayed in Figure S4. In the decision trees, the samples number indicates the number of patients predicted in each class. The class represents whether the algorithm was predicting for the development of disease or for the absence of disease (no outcome). The true patient distribution is depicted in the bracket values by (number of patients with correctly predicted outcome, number of misclassified patients).

To account for the lack of transferability of decision trees, we also tested the random forest algorithm. The random forest model produced mean AUROCs for the validation set of 0.809 for CVD, 0.775 for heart failure, 0.711 for atrial fibrillation, and 0.768 for coronary artery disease.

Time-Split Testing

We then examined the performance of the machine learning algorithms on the time-split testing data set (Table 3, Figure 3). AUROCs were highest for the regularized logistic regression algorithms, with AUROCs being 0.806, 0.845, 0.783, and 0.792 for CVD, heart failure, atrial fibrillation, and coronary artery disease, respectively. We provide nomograms for the time-split analysis regularized logistic regression models in Figure S5. The decision tree model produced AUROCs of 0.741, 0.741, 0.761, and 0.733 for these outcomes, respectively. The AUROCs produced by testing the random forest model on the time-split data were 0.799, 0.851, 0.752, and 0.791 for these outcomes, respectively.

Random forest overall improved the performance for the time-split test set compared with decision tree. SHAP bar plots were used to highlight the top 5 most impactful features on the random forest models (Figure S1). In these plots, the larger the mean absolute SHAP value, the greater the impact of that feature on the model's prediction. For example, for CVD, pr_carmeg (cardiomegaly noted before cancer diagnosis) was the most influential variable in predicting overall CVD in composite, indicated by the highest mean absolute value SHAP score (Figure S1). In addition to bar plots, beeswarm plots were also created for each cardiac event to interpret the direction of impact of the most influential features. Each point is an observation on the beeswarm plot, with the *y* value being the feature, the color of the point representing the relative contribution of the feature for 1 patient, and the *x* value representing the magnitude of influence of that feature on predicting that cardiovascular outcome. Positive SHAP values indicate the feature was more likely to predict the patient would develop that cardiovascular outcome.

Additional Validation

The AUROCs for the regularized logistic regression, decision tree, and random forest algorithms when tested on the additional 329-patient validation set were similar to the time-split test set performances (Table 4,

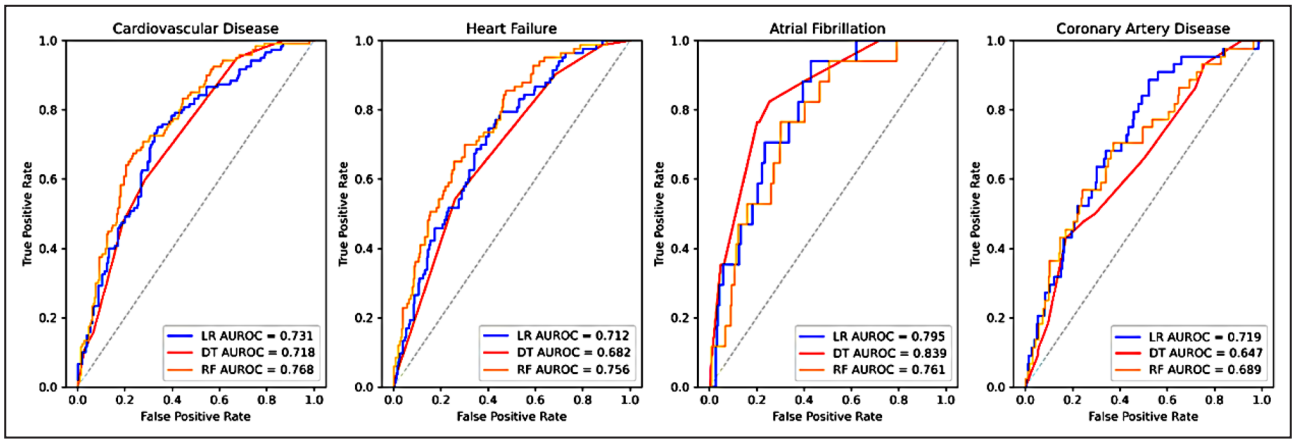

Figure 4. Receiver operating characteristic curves for the regularized LR, DT, and RF algorithms when tested on the additional 329-patient validation set.
Outcomes examined are cardiovascular disease, heart failure, atrial fibrillation, and coronary artery disease. AUROC values are shown at the bottom of the figure. AUROC indicates area under the receiver operating characteristic curve; DT, decision tree; LR, logistic regression; and RF, random forest.

Figure 4). Outcomes examined were the same: CVD, heart failure, atrial fibrillation, and coronary artery disease.

De Novo Predictions

We also tested whether our method could predict de novo CVD. Figure 5 shows the average AUROCs produced for de novo CVD in the data set of 3835 patients, with 0.826 using regularized logistic regression, 0.735 using decision tree classification, and 0.802 using random forest.

Newer Advanced Machine Learning Algorithms

We expanded our investigation to implement 4 additional machine learning methods, namely the BART and 3 gradient boosting models, XGBoost, LightGBM, and CatBoost, yielding new data. The AUROC performance of the BART model was 0.79, 0.84, 0.76, and 0.79, for CVD, heart failure, atrial fibrillation, and coronary artery disease, respectively. For XGBoost, the AUROCs produced were 0.75, 0.77, 0.67, and 0.711 for these same outcomes, respectively. The LightGBM model produced AUROCs of 0.75, 0.80, 0.72, and 0.75 for these outcomes, respectively. The AUROCs for CatBoost were 0.78, 0.84, 0.76, and 0.75 for these outcomes, respectively (Table 5). When tested on the additional 329-patient validation set, the AUROCs for the 4 new algorithms were similar to the time-split test set performance (Table 6). Outcomes examined were the same: CVD, heart failure, atrial fibrillation, and coronary artery disease. In our group's previous work, multiple clinically significant variables were identified as

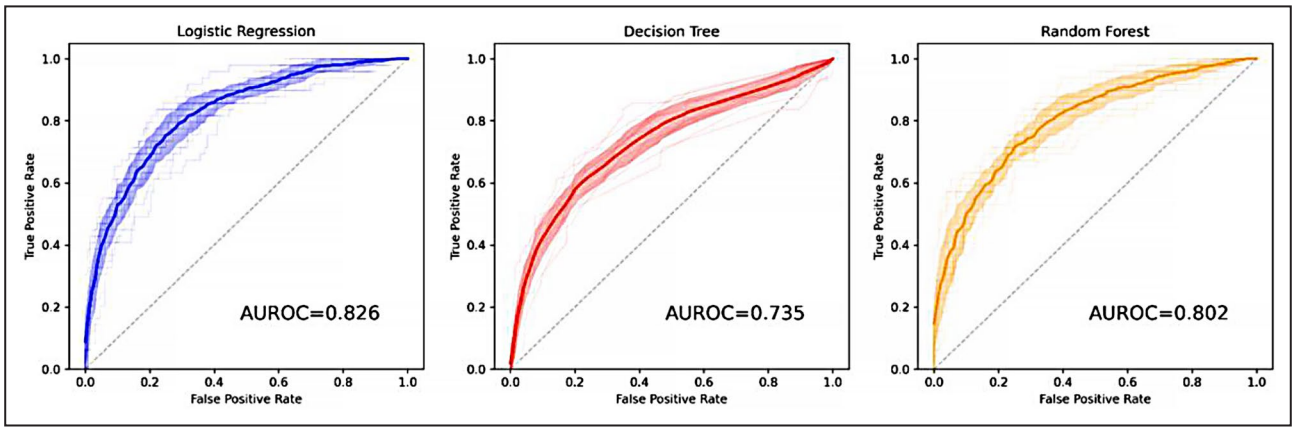

Figure 5. Receiver operating characteristic curves for the regularized logistic regression, decision tree, and random forest algorithms for the prediction of de novo cardiovascular disease as a composite of any heart failure, atrial fibrillation, or coronary artery disease after cancer diagnosis in the data set of 3835 cancer survivors.
AUROCs are shown. AUROC indicates area under the receiver operating characteristic curve.

being significantly associated with CVD, including age, hypertension, glucose levels, left ventricular ejection fraction, creatinine, and aspartate aminotransferase levels.[21] Clinically significant variables were confirmed in the current work, including prior hyperlipidemia, prior peripheral edema, sodium, chloride, and natriuretic peptide B prohormone.

Model Performance Comparisons

To formally compare model performance, we conducted pairwise comparisons of AUROC using the DeLong test among regularized logistic regression and each of the advanced machine learning models (BART, CatBoost, XGBoost, LightGBM) and random forest across all 4 cardiovascular outcomes. These results are presented in Tables S4 through S8. Logistic regression was significantly outperformed by XGBoost for heart failure (P=0.0131) and coronary artery disease (P=0.0065). No significant differences were observed between logistic regression and other models for atrial fibrillation or composite CVD (all P>0.05), indicating similar discriminatory performance between logistic regression and each of the advanced machine learning algorithms in those cases.

DISCUSSION

In this study, we evaluated predictive machine learning models for cardiac risk assessment among 3 types of cardiovascular outcomes, including heart failure, atrial fibrillation, and coronary artery disease, as well as a composite of the 3, regardless of the timing of cancer diagnosis, and also specifically after cancer diagnosis. We sought to ensure internal validity by using repeated cross-validation and regularization techniques to minimize overfitting and preserve consistency in model performance within the training data set, following established practices in traditional statistical modeling and machine learning prognostic tools.[39,40] All outcomes received relatively high or high AUROCs, ranging from 0.7 to 0.8 using decision tree and random forest, in our retrospective cohort of ≈4000 individuals spanning at least 20 years. Both regularized logistic regression and the BART advanced machine learning algorithm received AUROCs of above or near 0.8 for all outcomes. In addition, models built using time-split data suggested high applicability and transferability of our models for potential future clinical implementation. Of note, almost 20% of participants were non-Hispanic Black. Demographics have not traditionally been reported in cardio-oncology studies, which have not typically focused on ensuring multiracial cohorts.[41] Consequently, our group's database may be of substantial benefit to the cardio-oncology community, because we detailed the demographics and epidemiology of our cohort.[23]

A recent study indicated that models trained on national data may be more transferable than models trained on local data alone.[22] We evaluated the models by conducting testing on independent patient populations from distinct institutions and time periods, mirroring the standards used in similar studies to assess external applicability, recognizing that real-world application requires consistency across populations, time periods, and clinical settings.[42] The models' ability to retain predictive accuracy across diverse clinical contexts reinforces transferability.[42] This reflects the machine learning approach of empirical transferability and resonates with previous machine learning studies.[39,40] Before the current study, it was unknown whether the preexisting machine learning algorithms trained locally to predict CVD in cancer survivors would be applicable and transferable. Our study suggests that these algorithms can be applicable and transferable with accurate prediction of CVD in cancer survivors. The caveat would be that we retrained and retested and revalidated the models and algorithms on our external population. Further studies would be needed to determine to what extent the models would be equally externally transferable with limited additional retesting, retraining, and revalidation on each external data set. Although machine learning offers scalable potential for clinical prediction and screening, real-world implementation may require population-specific calibration, high-quality input data, and validation in diverse clinical settings to ensure model sensitivity, optimize thresholds, and assess cost-effectiveness.[39]

Although national heterogenous databases are being built in cardio-oncology, until we have a robust universal database for national use in cardio-oncology, perhaps we should expect to continue to retrain, retest, and revalidate models appropriately. Combining resources in multi-institutional research groups such as ours may facilitate this.[23] Notably, the algorithms were used to predict CVD regardless of the timing of cancer diagnosis, as well as specifically after cancer diagnosis. This suggests that the algorithms may be able to predict CVD in the general population without cancer diagnoses, as well as CVD in cancer survivors. These algorithms may have usefulness in preventive cardiology and also in cardio-oncology.

Several cardio-oncology consensus recommendations and risk stratification tools for the management of cancer therapy related cardiotoxicities have become available.[10–13,43] Additionally, a position statement recommends proformas that risk stratify patients based on baseline features into low-, moderate-, and high-risk categories for developing CVD from cancer treatment.[44] These proformas/tools are limited by imprecision and challenges with individualizing the recommendations. Even the scoping guidelines available for use are not based on precision medicine data and

lack information from innovative studies that can truly personalize care. Furthermore, the guidelines and proformas are for a subset of patients receiving specific cancer treatments and are not universally applicable or generalizable for cardio-oncology patients undergoing therapy that were not included. Our models overcome these challenges by using machine learning for more precise prediction, with AUROCs that match other machine learning models used for prediction of CVD in individuals with or without cancer.[21,39,45–50]

Our group's validated models complement previous work studying the development of various forms of computational frameworks for analyzing and predicting patient outcomes. For example, prior work has investigated machine learning applied to phenotypic, therapeutic, chemical, and genomic properties for predicting drug interactions as a common cause of adverse drug reactions.[51] Another example from a different group is the DeepProg machine learning framework, which uses multiple omics data matrices and existing survival information as input, to model future patient survival. DeepProg uses RNA sequencing, methylation, and microRNA data, paired with information in the Cancer Genome Atlas.[52] From another group, a multimodel data set included histopathological, radiologic, and clinico-genomic machine learning models to predict the prognosis of patients with ovarian cancer.[53] Our models in the current study rely on data from the electronic health record, which is broadly available across patient populations. Future models will incorporate genomic and other forms of emerging data becoming available in the cardio-oncology population. Comparing the effectiveness of each kind of predictive model with varying forms of precision medicine data will likely provide valuable insight on the necessity of multiomics data combined with the use of artificial intelligence algorithms to predict patient outcomes.

In our models, probabilities are generated for each outcome, although risk levels are not explicitly forecast. A system of risk-based tertiles or quartiles based on our data may also be incorporated in the next iteration of the models.[54,55] Our future work may additionally take a variety of forms. We can continue to improve the models as more data are collected, because in previous work we observed a marginal increase in model performance as training sizes increased,[21] highlighting the importance of large-scale cohort studies for machine learning research. We also previously noted that restricting variables to a specific time period (ie, within 1, 5, and 10 years of the first diagnosis for the outcome) led to similar model performance, although some outcomes may be slightly improved.[21] Our current models give excellent prediction without the inclusion of information on cancer types and treatment; however, we can incorporate this information into future models. Additionally, our teams are actively incorporating additional forms of imaging data[39,56–59] directly into models via convolutional neural networks to further improve their performance. Furthermore, we intend to integrate machine learning-based risk assessment with clinically relevant online tools, as well as various forms of omics and other emerging precision medicine data, for use in cardio-oncology practice.

Although spot checks were conducted for verification, it should be noted that potential errors in *ICD-9* and *ICD-10* codes may affect machine learning performance. There has been a long tradition of using *ICD-9* diagnoses to define outcomes or comorbidities, particularly with large administrative databanks such as Medicare.[60,61] More recent work has moved on to validated *ICD-9* from/to *ICD-10* crosswalks, in concert with the advent of more data availability via electronic health records.[62,63] In addition to diagnosis codes, condition identification is also supplemented by procedural codes. Procedural and pharmaceutical codes may be more precise than diagnosis coding that may vary among clinicians and health systems.[64,65] The usefulness of additional codes beyond diagnosis codes for the prediction of CVD in cancer survivors can potentially be explored in the future, particularly using procedural and pharmaceutical codes that are associated with CVD outcomes.

There exists some overlap between the fields of machine learning and traditional statistics, such as in the case of the regularized logistic regression model, which can be classified as being part of both fields.[66] Machine learning and traditional statistics often complement each other. For example, machine learning algorithms are often considered as black box models, although methods such as SHAP plots can help with the interpretation of the contributions of the features (Figure S1). Traditional statistical analyses, such as nomograms, can offer easy-to-understand tools that can be used to visualize risk (Figure S5). Although SHAP provides a ranking of the most influential variables, nomograms provide a comprehensive list of the influence of all variables. An important point of consideration is that SHAP, not nomograms, can be used for interpreting nonlinear models such as random forest. We have therefore provided both SHAP and nomograms as modes of visualization, because they each possess distinct value. To further elaborate on the philosophical and methodological distinctions between machine learning and traditional statistical approaches in our study, we provide Table S9. This table contrasts perspectives on generalizability, internal and external validity, and calibration, all of which are critical concepts for both predictive modeling and inference.

Bridging machine learning and traditional statistics can be especially helpful as we seek explainable artificial intelligence for reliable translation into clinical practice. Albeit with challenges and limitations, applying

these algorithms for personalized risk prediction and tailoring care of adult cancer survivors at risk of developing various forms of CVD remains an emerging concept. The results of this study suggest that machine learning approaches, alongside traditional statistics methodology, can be used to stratify cardiac risk among oncology patients by using large-scale, longitudinal data from health care systems. Findings from this study may provide valuable guidance and tools for oncologists, cardiologists, informaticians, and administrators tasked with improving prediction and care of cancer survivors in cardio-oncology.

The main advantage of the decision tree algorithm is interpretability. The variables at the top of the decision tree, therefore most influential for the development of new CVD and the development of new heart failure, were preexisting restrictive cardiomyopathy and troponin T, and preexisting shortness of breath, respectively (Figure S4). For the development of new atrial fibrillation, preexisting restrictive cardiomyopathy, creatine kinase, age, and NT-proBNP (N-terminal pro-B-type natriuretic peptide) were highly relevant in predicting this outcome. For the development of new coronary artery disease developed after cancer diagnosis, preexisting hyperlipidemia, preexisting restrictive cardiomyopathy, creatine kinase myocardial bound, and unspecified peripheral vascular disease were influential. It is clinically consistent that preexisting restrictive cardiomyopathy, shortness of breath, hyperlipidemia, and unspecified peripheral vascular disease present before cancer diagnosis, as well as age, troponin T, creatine kinase, creatine kinase myocardial bound, and NT-proBNP assessed after cancer diagnosis, can predict the future development of various forms of new CVD. It is important to highlight that the newly developed CVD diagnoses predicted by the algorithm were not present at the time of measurement of troponin T, creatine kinase, creatine kinase myocardial bound, or NT-proBNP. Additionally, no single predictor was sufficient on its own to predict these forms of CVD. Together, the variety of components were modeled for accurate prediction. Finally, although the decision tree is interpretable, the algorithm is vulnerable to overfitting. To minimize overfitting, the model was tested against both a time-split data set and an additional 329-patient validation set.

To further address the potential weakness of the decision tree algorithm, we also applied the random forest algorithm, which integrates the results of multiple decision trees to avoid the issue of a single tree overfitting. This algorithm showed stronger performance for certain outcomes, though at the expense of being less interpretable. We used SHAP values to aid in the interpretability of this algorithm. In Figure S1, patients who did not develop a particular cardiovascular outcome are represented in the blue color, whereas patients who developed the cardiovascular outcome are colored in red. Variable importance is ranked vertically, with most influential variables at the top of the plots. Points with a positive SHAP value indicate data points that contributed toward predicting the outcome of interest, whereas points with a negative SHAP value did not contribute toward predicting cardiovascular outcome. For example, for the first beeswarm plot (for developing CVD in composite overall), pr_carmeg (the presence of preexisting cardiomegaly before cancer diagnosis) is the most influential variable. For this variable, the blue points are on the left (negative SHAP), and the red points are on the right (positive SHAP), indicating that patients who were positive for this pr_carmeg variable (ie, they had a preexisting history of cardiomegaly before cancer diagnosis) were more likely to develop the CVD outcome, whereas patients who did not have previous cardiomegaly were less likely to develop the CVD outcome. It is important to note the relevance of transition of SHAP values from blue (left) to right (red). For example, for the variable age, the points change from blue to red from left to right. This means that the greater the patient's age, the more likely a patient was to develop CVD.

SHAP values were also determined for calculating heart failure, atrial fibrillation, and coronary artery disease, separately from the composite cardiovascular outcome. For the development of new heart failure, pr_carmeg (the presence of preexisting cardiomegaly before cancer diagnosis) was the most predictive, followed by pr_carmyor (preexisting restrictive cardiomyopathy), troponin_T (troponin T), pr_pad (preexisting peripheral artery disease), and pr_sob (preexisting shortness of breath). For the development of new atrial fibrillation, pr_carmeg (the presence of preexisting cardiomegaly before cancer diagnosis) is the most influential variable, followed by age, pr_carmyor (preexisting restrictive cardiomyopathy), creatine kinase, and nt-probnp (NT-proBNP). For the development of new coronary artery disease (developed after cancer diagnosis), pr_hlp (preexisting hyperlipidemia) was the most influential, followed by pr_pad (preexisting peripheral artery disease), ckmb (creatine kinase myocardial bound), pr_upvd (unspecified peripheral vascular disease), and pr_carmyor (preexisting restrictive cardiomyopathy). These findings were overall consistent with those from the decision tree algorithms and may potentially have clinical usefulness.

We further assessed the predictive capacities of machine learning by using 4 additional algorithms. We used varying complex algorithms including BART, and 3 gradient boosting models, XGBoost, LightGBM, and CatBoost. The performance of these additional machine learning models varied. Overall, BART yielded the highest AUROCs across these 4 newer models, with AUROCs comparable to regularized logistic

regression and random forest for predictive performance (Table 5). This trend was also noted when tested in the 329-sample validation group (Table 6). Another study evaluating machine learning algorithms showed that BART outperformed other nonlinear models and performed similarly to linear regression.[67] Additionally, they highlighted the accuracy and ease of use of BART.

In comparison with the substantial history of use of traditional logistic regression in statistical inference and predictive modeling, the newer advancements of machine learning remain relatively recent. We acknowledge that logistic regression fundamentally possesses a statistical background. We have therefore treated regularized logistic regression as both a statistical (logistic regression) and a more traditional (or older) machine learning method (regularized) in comparison with the newer, advanced machine learning models we have also used. In the testing set for all 4 outcomes (CVD, heart failure, atrial fibrillation, and coronary artery disease), regularized logistic regression and the machine learning algorithms overall performed comparably, yielding similar AUROCs. For composite CVD, for example, BART demonstrated an AUROC of 0.789, compared with 0.795 for regularized logistic regression. Notably, a possible explanation for this similar performance may be due to the ability of logistic regression to handle both linear and nonlinear data, similar to the advanced machine learning algorithms. To capture nonlinear relationships, logistic regression models the log-odds of the outcome as a linear combination of predictor variables, corresponding to an additive model on the logistic scale. In this way, logistic regression can accommodate nonlinear effects and interaction terms, through appropriate modeling techniques, such as spline functions and polynomial expansions. Although our implementation did not incorporate such transformations, regularized logistic regression appeared to capture complex patterns in our data, compared with more advanced machine learning. Logistic regression relatively consistently outperformed decision tree, with AUROCs overall >0.7 compared with >0.6 (Tables 5 and 6), with significant difference confirmed with the DeLong test *P* values <0.05 (Tables S4, through S8). There was 1 significant DeLong test *P* value, when comparing logistic regression to the more advanced machine learning algorithms. This was noted for comparing regularized logistic regression (AUROC, 0.790) to XGBoost (AUROC, 0.711), in favor of logistic regression. This was specifically for coronary artery disease (*P*=0.0065) and was not observed for composite CVD (*P*=0.185). Importantly, the AUROCs of the 2 models may appear numerically similar (ie, 0.790 and 0.711). A DeLong test *P*<0.05 indicates a statistically significant difference in discriminatory performance, likely driven by consistent directional differences across predictions in our relatively large sample; yet, this statistical separation should be interpreted cautiously, because even a modest but significant difference (eg, 0.01 AUROC units) may lack clinical relevance unless supported by meaningful improvements in decision-making, calibration, or patient outcomes. Taking all of these findings and deliberations into consideration, our overall assessment would be that although some prior literature has shown advanced machine learning models outperforming regularized logistic regression, our results do not show this substantially in our data set.[17] Our results suggest that logistic regression continues to serve a prominent role in the prediction of CVD in cancer survivors. This can be beneficial for advancing trust and adoption of artificial intelligence prediction in clinical practice, because clinicians are generally familiar with logistic regression. With that said, further exploration and comparison of more advanced machine learning methods and regularized logistic regression may continually be warranted.

Notwithstanding, we noted various limitations in our study. We recognize that restricting our previous analyses to patients who already have ascertainment of ejection fraction at baseline may select for patients with at least suspicion of or potential for future cardiomyopathy, such as those with a preexisting cardiopulmonary complaint or those planned for cardiotoxic drugs such as anthracyclines. Nevertheless, we report on preexisting and subsequently developed CVDs in this population and also on the array of cancer medications used, in our epidemiologic study report.[23] We would expect that these algorithms may be transferable to cancer survivors with similar characteristics. Information on cancer type and treatment were not explicitly included in the models, which gave excellent prediction without this information. Cancer type and treatment information can be considered for inclusion in future models. Future studies may use alternative means of baseline assessment and subsequent prediction, such as our new studies currently underway using baseline ECGs instead of baseline echocardiograms for prediction; of course, the caveat will also be that those patients will have had baseline ECGs, which are more common than baseline echocardiograms and are often obtained in the absence of preexisting cardiovascular complaints.

Our study suggests applicability and transferability in our F&MCW population of these machine learning algorithms, and (regularized) logistic regression models, that were previously validated in a cardio-oncology population at Cleveland Clinic. Our results therefore indicate applicability, transferability, and preservation of model discrimination when applied to a temporally and institutionally independent data set. Consequently, our results suggest transferability in terms of discrimination performance across our F&MCW clinic and hospital sites in a robust health system, consistent with results

from the previous Cleveland Clinic population study. Although the model demonstrated stable discrimination metrics in this separate population (F&MCW after Cleveland Clinic), we did not assess calibration. Given this limitation, we refrain from describing our findings as evidence of external validation or generalizability in the formal statistical sense. Instead, we describe our results as supportive of potential model transportability with respect to discrimination performance. Rigorous calibration of machine learning or statistical models to a specific population of interest would be an important future step for both clinical relevance and public health usefulness.[40] Calibration refers to the assessment of whether predicted probabilities match observed outcomes. Population-specific thresholds may be necessary, and initial calibration should be grounded in representative validation data. Calibration should be coupled with stringent quality control of data inputs, to maintain predictive integrity. Moreover, further evaluation of cost-effectiveness and clinical impact, especially when integrated with additional clinical and laboratory variables, would facilitate responsible deployment of such tools.

CONCLUSIONS

Machine learning models for CVD outcomes in the oncology population were evaluated and suggested multi-institutional transferability, using a systematic analysis of classification methods and feature sets. Models developed using time-split data showed moderately high performance and potential real-world transferability over time. This may indicate potential for external applicability, though we recognize the importance of calibration to formally establish external validity in statistical studies. The future versions of our models may also incorporate tertile or quartile risk stratification to aid clinical decision-making and ultimately improve patient outcomes. Furthermore, our group is currently developing accessible online tools that incorporate outcomes from our models. In summary, our findings suggest that machine learning tools may prove useful for assessing cardiac risk in cancer patients by integrating large-scale, longitudinal data from health care systems.

ARTICLE INFORMATION

Received April 21, 2023; accepted June 9, 2025.

Affiliations

Department of Medicine, Medical College of Wisconsin, Milwaukee, WI (S.-A.B.); Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN (S.-A.B., A.A.); Heart Innovation and Equity Research (HIER) Institute, Miami, FL (S.-A.B.); Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH (M.Z.F., F.C.); Institute for Health and Equity, Medical College of Wisconsin, Milwaukee, WI (R.S., J.O.); Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH (Y.Z., F.C.); Clinical and Translational Science Institute, Medical College of Wisconsin, Milwaukee, WI (K.O., B.T.); Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI (D.Y., C.O.); Department of Computer Science, Milwaukee School of Engineering, Milwaukee, WI (J.B.); Medical College of Wisconsin, Milwaukee, WI (R.S., M.B.); Department of Cardiovascular Diseases, Cleveland Clinic, Cleveland, OH (P.C.); Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI (M.B., J.Z.); Division of Surgical Oncology, Department of Surgery, Medical College of Wisconsin, Milwaukee, WI (A.K.); Department of Internal Medicine, Baton Rouge General Medical Center, Baton Rouge, LA (G.E.); Department of Medicine, Mercer University School of Medicine, Mercer University, Macon, GA (J.R.); Miami Dade College, Miami, FL (Z.S.); and Division of Hematology and Oncology, Department of Medicine, Medical College of Wisconsin, Milwaukee, WI (Y.C.C.).

Acknowledgments

The authors are grateful to Dr Widlansky at Medical College of Wisconsin for his mentorship in this work, and to K. Hansen for her collaboration as a patient advocate representative in our research group, as well as Dr Lian at Mercer University School of Medicine for his invaluable assistance with providing helpful feedback on the statistical components. Artificial intelligence tools were used in the development of this article.

Sources of Funding

This publication was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant numbers UL1TR001436 and KL2TR001438. This work was also 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. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Disclosures

None.

Supplemental Material

Tables S1–S9
Figures S1–S5

REFERENCES

1. Miller KD, Nogueira L, Devasia T, Mariotto AB, Yabroff KR, Jemal A, Kramer J, Siegel RL. Cancer treatment and survivorship statistics, 2022. *CA Cancer J Clin*. 2022;72:409–436. doi: 10.3322/caac.21731
2. Siegel RL, Kratzer TB, Giaquinto AN, Sung H, Jemal A. Cancer statistics, 2025. *CA Cancer J Clin*. 2025;75:10–45. doi: 10.3322/caac.21871
3. 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. doi: 10.3322/caac.21565
4. Dent SF, Kikuchi R, Kondapalli L, Ismail-Khan R, Brezden-Masley C, Barac A, Fradley M. Optimizing cardiovascular health in patients with cancer: a practical review of risk assessment, monitoring, and prevention of cancer treatment-related cardiovascular toxicity. *Am Soc Clin Oncol Educ Book*. 2020;40:1–15. doi: 10.1200/EDBK_286019
5. Sadler D, Chaulagain C, Alvarado B, Cubeddu R, Stone E, Samuel T, Bastos B, Grossman D, Fu CL, Alley E, et al. Practical and cost-effective model to build and sustain a cardio-oncology program. *Cardio-Oncology*. 2020;6:9. doi: 10.1186/s40959-020-00063-x
6. Fradley MG, Brown AC, Shields B, Viganego F, Damrongwatanasuk R, Patel AA, Hartlage G, Roper N, Jaunese J, Roy L, et al. Developing a comprehensive cardio-oncology program at a cancer institute: the Moffitt Cancer Center experience. *Oncol Rev*. 2017;11:340. doi: 10.4081/oncol.2017.340
7. Okwuosa TM, Akhter N, Williams KA, DeCara JM. Building a cardio-oncology program in a small- to medium-sized, nonprimary cancer center, academic hospital in the USA: challenges and pitfalls. *Futur Cardiol*. 2015;11:413–420. doi: 10.2217/FCA.15.43
8. Parent S, Pituskin E, Paterson DI. The cardio-oncology program: a multidisciplinary approach to the Care of Cancer Patients with Cardiovascular Disease. *Can J Cardiol*. 2016;32:847–851. doi: 10.1016/j.cjca.2016.04.014

9. Snipelisky D, Park JY, Lerman A, Mulvagh S, Lin G, Pereira N, Rodriguez-Porcel M, Villarraga HR, Herrmann J. How to develop a cardio-oncology clinic. *Heart Fail Clin*. 2017;13:347–359. doi: 10.1016/j.hfc.2016.12.011
10. Armenian SH, Lacchetti C, Barac A, Carver J, Constine LS, Denduluri N, Dent S, Douglas PS, Durand JB, Ewer M, et al. Prevention and monitoring of cardiac dysfunction in survivors of adult cancers: American Society of Clinical Oncology clinical practice guideline. *J Clin Oncol*. 2017;35:893–911. doi: 10.1200/JCO.2016.70.5400
11. Herrmann J, Lerman A, Sandhu NP, Villarraga HR, Mulvagh SL, Kohli M. Evaluation and management of patients with heart disease and cancer: cardio-oncology. *Mayo Clin Proc*. 2014;89:1287–1306. doi: 10.1016/j.mayocp.2014.05.013
12. Lyon AR, López-Fernández T, Couch LS, Asteggiano R, Aznar MC, Bergler-Klein J, Boriani G, Cardinale D, Cordoba R, Cosyns B, et al. 2022 ESC guidelines on cardio-oncology developed in collaboration with the European Hematology Association (EHA), the European Society for Therapeutic Radiology and Oncology (ESTRO) and the international cardio-oncology society (IC-OS). *Eur Heart J*. 2022;43:4229–4361. doi: 10.1093/eurheartj/ehac244
13. Herrmann J, Lenihan D, Armenian S, Barac A, Blaes A, Cardinale D, Carver J, Dent S, Ky B, Lyon AR, et al. Defining cardiovascular toxicities of cancer therapies: an international cardio-oncology society (IC-OS) consensus statement. *Eur Heart J*. 2022;43:280–299. doi: 10.1093/eurheartj/ehab674
14. 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. doi: 10.1021/acs.jcim.7b00641
15. 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. doi: 10.1021/acs.jcim.8b00769
16. Schober P, Vetter TR. Logistic regression in medical research. *Anesth Analg*. 2021;132:365–366. doi: 10.1213/ANE.0000000000005247
17. Xi Y, Wang H, Sun N. Machine learning outperforms traditional logistic regression and offers new possibilities for cardiovascular risk prediction: a study involving 143,043 Chinese patients with hypertension. *Front Cardiovasc Med*. 2022;9:1025705. doi: 10.3389/fcvm.2022.1025705
18. Sievering AW, Wohlmuth P, Geßler N, Gunawardene MA, Herrlinger K, Bein B, Arnold D, Bergmann M, Nowak L, Gloeckner C, et al. Comparison of machine learning methods with logistic regression analysis in creating predictive models for risk of critical in-hospital events in COVID-19 patients on hospital admission. *BMC Med Inform Decis Mak*. 2022;22:309. doi: 10.1186/s12911-022-02057-4
19. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open*. 2020;3:e1918962. doi: 10.1001/jamanetworkopen.2019.18962
20. Sturgeon KM, Deng L, Bluethmann SM, Zhou S, Trifiletti DM, Jiang C, Kelly SP, Zaorsky NG. A population-based study of cardiovascular disease mortality risk in US cancer patients. *Eur Heart J*. 2019;40:3889–3897. doi: 10.1093/eurheartj/ehz766
21. Zhou Y, Hou Y, Hussain M, Brown SA, Budd T, Tang WHW, Abraham J, Xu B, Shah C, Moudgil R, et al. Machine learning-based risk assessment for cancer therapy-related cardiac dysfunction in 4300 longitudinal oncology patients. *J Am Heart Assoc*. 2020;9:e019628. doi: 10.1161/JAHA.120.019628
22. Kim E, Pieczkiewicz DS, Castro MR, Caraballo PJ, Simon GJ. Multi-task learning to identify outcome-specific risk factors that distinguish individual micro and macrovascular complications of type 2 diabetes. AMIA Joint Summits on Translational Science Proceedings.2018:122–31.
23. Brown SA, Sparapani R, Osinski K, Zhang J, Blessing J, Cheng F, Hamid A, Berman G, Lee K, BagheriMohamadiPour M, et al. Establishing an interdisciplinary research team for cardio-oncology artificial intelligence informatics precision and health equity. *Am Heart J Plus*. 2022;13:10094. doi: 10.1016/j.ahjo.2022.100094
24. Brown S-A. Patient similarity for decision-making in prevention of cardiovascular toxicity: a feasibility study. 2022 https://clinicaltrials.gov/ct2/show/NCT05377320.
25. Sperandei S. Understanding logistic regression analysis. *Biochem Med*. 2014;24:12–18. doi: 10.11613/BM.2014.003
26. Friedrich S, Groll A, Ickstadt K, Kneib T, Pauly M, Rahnenführer J, Friede T. Regularization approaches in clinical biostatistics: a review of methods and their applications. *Stat Methods Med Res*. 2023;32:425–440. doi: 10.1177/09622802221133557
27. Šinkovec H, Heinze G, Blagus R, Geroldinger A. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Med Res Methodol*. 2021;21:1–15. doi: 10.1186/s12874-021-01374-y
28. Norton EC, Dowd BE. Log odds and the interpretation of logit models. *Health Serv Res*. 2018;53:859–878. doi: 10.1111/1475-6773.12712
29. Luo C, Daniels MJ. Variable selection using Bayesian additive regression trees. *Stat Sci*. 2024;39:286–304. doi: 10.1214/23-sts900
30. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32. doi: 10.1023/A:1010933404324
31. Chen T, He T, Benesty M, Khotilovich V, Tang Y. Xgboost: extreme gradient boosting. R package version 04-2. 2015:1–4.
32. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat Med*. 2016;35:1159–1177. doi: 10.1002/sim.6782
33. Oh AR, Park J, Lee JH, Kim H, Yang K, Choi JH, Ahn J, Sung JD, Lee SH. Association between perioperative adverse cardiac events and mortality during one-year follow-up after noncardiac surgery. *J Am Heart Assoc*. 2022;11:e024325. doi: 10.1161/JAHA.121.024325
34. Imran M, Bhatti A, King DM, Lerch M, Dietrich J, Doron G, Manlik K. Supervised machine learning-based decision support for signal validation classification. *Drug Saf*. 2022;45:583–596. doi: 10.1007/s40264-022-01159-2
35. Naegelin M, Weibel RP, Kerr JI, Schinazi VR, La Marca R, von Wangenheim F, Hoelscher C, Ferrario A. An interpretable machine learning approach to multimodal stress detection in a simulated office environment. *J Biomed Inform*. 2023;139:104299. doi: 10.1016/j.jbi.2023.104299
36. Ladbury C, Zarinshenas R, Semwal H, Tam A, Vaidehi N, Rodin AS, Liu A, Glaser S, Salgia R, Amini A. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. *Transl Cancer Res*. 2022;11:3853–3868. doi: 10.21037/tcr-22-1626
37. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845. doi: 10.2307/2531595
38. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. doi: 10.1186/1471-2105-12-77
39. Akbilgic O, Butler L, Karabayir I, Chang PP, Kitzman DW, Alonso A, Chen LY, Soliman EZ. ECG-AI: electrocardiographic artificial intelligence model for prediction of heart failure. *Eur Heart J Digit Health*. 2021;2:626–634. doi: 10.1093/ehjdh/ztab080
40. Attia IZ, Tseng AS, Benavente ED, Medina-Inojosa JR, Clark TG, Malyutina S, Kapa S, Schirmer H, Kudryavtsev AV, Noseworthy PA, et al. External validation of a deep learning electrocardiogram algorithm to detect ventricular dysfunction. *Int J Cardiol*. 2021;329:130–135. doi: 10.1016/j.ijcard.2020.12.065
41. Prasad P, Branch M, Asemota D, Elsayed R, Addison D, Brown S-A. Cardio-oncology preventive care: racial and ethnic disparities. *Curr Cardiovasc Risk Rep*. 2020;14:18. doi: 10.1007/s12170-020-00650-8
42. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130:515–524. doi: 10.7326/0003-4819-130-6-199903160-00016
43. Brown SA, Okwuosa TM, Barac A, Volgman AS. The role of angiotensin-converting enzyme inhibitors and β-blockers in primary prevention of cardiac dysfunction in breast cancer patients. *J Am Heart Assoc*. 2020;9:e015327. doi: 10.1161/JAHA.119.015327
44. Lyon AR, Dent S, Stanway S, Earl H, Brezden-Masley C, Cohen-Solal A, Tocchetti CG, Moslehi JJ, Groarke JD, Bergler-Klein J, et al. Baseline cardiovascular risk assessment in cancer patients scheduled to receive cardiotoxic cancer therapies: a position statement and new risk assessment tools from the cardio-oncology study Group of the Heart Failure Association of the European Society of Cardiology in collaboration with the international cardio-oncology society. *Eur J Heart Fail*. 2020;22:1945–1960. doi: 10.1002/ejhf.1920
45. Hou Y, Zhou Y, Hussain M, Budd GT, Tang WHW, Abraham J, Xu B, Shah C, Moudgil R, Popovic Z, et al. Cardiac risk stratification in cancer patients: a longitudinal patient-patient network analysis. *PLoS Med*. 2021;18:e1003736. doi: 10.1371/journal.pmed.1003736

46. Güntürkün F, Akbilgic O, Davis RL, Armstrong GT, Howell RM, Jefferies JL, Ness KK, Karabayir I, Lucas JT Jr, Srivastava DK, et al. Artificial intelligence-assisted prediction of late-onset cardiomyopathy among childhood cancer survivors. *JCO Clin Cancer Inform*. 2021;5:459–468.
47. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*. 2019;394:861–867. doi: 10.1016/S0140-6736(19)31721-0
48. Attia ZI, Kapa S, Yao X, Lopez-Jimenez F, Mohan TL, Pellikka PA, Carter RE, Shah ND, Friedman PA, Noseworthy PA. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *J Cardiovasc Electrophysiol*. 2019;30:668–674. doi: 10.1111/jce.13889
49. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25:70–74. doi: 10.1038/s41591-018-0240-2
50. Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, Kapa S, Friedman PA, Lopez-Jimenez F. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. *Circ Arrhythm Electrophysiol*. 2020;13:e007988. doi: 10.1161/CIRCEP.119.007988
51. Cheng F, Zhao Z. Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *J Am Med Inform Assoc*. 2014;21(e2):e278–e286. doi: 10.1136/amiajnl-2013-002512
52. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med*. 2021;13:112. doi: 10.1186/s13073-021-00930-x
53. Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García I, Zamarin D, Long Roche K, Liu Y, Patel D, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Can*. 2022;3:723–733. doi: 10.1038/s43018-022-00388-9
54. 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. doi: 10.1016/j.jaccao.2019.10.008
55. 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. doi: 10.1093/eurheartj/ehz460
56. Gunturkun F, Davis RL, Armstrong GT, Jefferies JL, Ness KK, Green DM, Lucas JT, Srivastava D, Hudson MM, Robison LL, et al. Deep learning for improved prediction of late-onset cardiomyopathy among childhood cancer survivors: a report from the St. Jude lifetime cohort (SJLIFE). *J Clin Oncol*. 2020;38:10545–10545. doi: 10.1200/JCO.2020.38.15_suppl.10545
57. Akbilgic O, Butler L, Karabayir I, Chang P, Kritzman D, Alonso A, Chen L, Soliman E. Artificial intelligence applied to ECG improves heart failure prediction accuracy. *JACC*. 2021;77:3045.
58. Henglin M, Stein G, Hushcha PV, Snoek J, Wiltschko AB, Cheng S. Machine learning approaches in cardiovascular imaging. *Circ Cardiovasc Imaging*. 2017;10. doi: 10.1161/CIRCIMAGING.117.005614
59. 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. *Circulation*. 2018;138:1623–1635. doi: 10.1161/CIRCULATIONAHA.118.034338
60. Roper WL, Winkenwerder W, Hackbarth GM, Krakauer H. Effectiveness in health care. An initiative to evaluate and improve medical practice. *N Engl J Med*. 1988;319:1197–1202.
61. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol*. 1992;45:613–619. doi: 10.1016/0895-4356(92)90133-8
62. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43:1130–1139. doi: 10.1097/01.mlr.0000182534.19832.83
63. Sun JW, Rogers JR, Her Q, Welch EC, Panozzo CA, Toh S, Gagne JJ. Adaptation and validation of the combined comorbidity score for ICD-10-CM. *Med Care*. 2017;55:1046–1051. doi: 10.1097/MLR.0000000000000824
64. Fishman PA, Goodman MJ, Hornbrook MC, Meenan RT, Bachman DJ, O'Keeffe Rosetti MC. Risk adjustment using automated ambulatory pharmacy data: the RxRisk model. *Med Care*. 2003;41:84–99. doi: 10.1097/00005650-200301000-00011
65. Johnson ML, El-Serag HB, Tran TT, Hartman C, Richardson P, Abraham NS. Adapting the Rx-risk-V for mortality prediction in outpatient populations. *Med Care*. 2006;44:793–797. doi: 10.1097/01.mlr.0000218804.41758.ef
66. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2:56–67. doi: 10.1038/s42256-019-0138-9
67. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat*. 2011;20:217–240. doi: 10.1198/jcgs.2010.08162