JACC: HEART FAILURE VOL. ■, NO. ■, 2020
© 2020 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER

A Machine Learning Approach to Management of Heart Failure Populations

Linyuan Jing, PhD,[a] Alvaro E. Ulloa Cerna, PhD,[a] Christopher W. Good, DO,[b] Nathan M. Sauers, PharmD,[c] Gargi Schneider, MD,[d] Dustin N. Hartzel, BS,[e] Joseph B. Leader, BA,[e] H. Lester Kirchner, PhD,[f] Yirui Hu, PhD,[f] David M. Riviello, BS,[g] Joshua V. Stough, PhD,[a,h] Seth Gazes, MS,[c] Allyson Haggerty, MBA,[a] Sushravya Raghunath, PhD,[a] Brendan J. Carry, MD,[b] Christopher M. Haggerty, PhD,[a,b,*] Brandon K. Fornwalt, MD, PhD[a,b,i,*]

ABSTRACT

BACKGROUND Heart failure is a prevalent, costly disease for which new value-based payment models demand optimized population management strategies.

OBJECTIVES This study sought to generate a strategy for managing populations of patients with heart failure by leveraging large clinical datasets and machine learning.

METHODS Geisinger electronic health record data were used to train machine learning models to predict 1-year all-cause mortality in 26,971 patients with heart failure who underwent 276,819 clinical episodes. There were 26 clinical variables (demographics, laboratory test results, medications), 90 diagnostic codes, 41 electrocardiogram measurements and patterns, 44 echocardiographic measurements, and 8 evidence-based "care gaps": flu vaccine, blood pressure of <130/80 mm Hg, A_{1c} of <8%, cardiac resynchronization therapy, and active medications (active angiotensin-converting enzyme inhibitor/angiotensin II receptor blocker/angiotensin receptor-neprilysin inhibitor, aldosterone receptor antagonist, hydralazine, and evidence-based beta-blocker) were collected. Care gaps represented actionable variables for which associations with all-cause mortality were modeled from retrospective data and then used to predict the benefit of prospective interventions in 13,238 currently living patients.

RESULTS Machine learning models achieved areas under the receiver-operating characteristic curve (AUCs) of 0.74 to 0.77 in a split-by-year training/test scheme, with the nonlinear XGBoost model (AUC: 0.77) outperforming linear logistic regression (AUC: 0.74). Out of 13,238 currently living patients, 2,844 were predicted to die within a year, and closing all care gaps was predicted to save 231 of these lives. Prioritizing patients for intervention by using the predicted reduction in 1-year mortality risk outperformed all other priority rankings (e.g., random selection or Seattle Heart Failure risk score).

CONCLUSIONS Machine learning can be used to priority-rank patients most likely to benefit from interventions to optimize evidence-based therapies. This approach may prove useful for optimizing heart failure population health management teams within value-based payment models. (J Am Coll Cardiol HF 2020;■:■–■)
© 2020 by the American College of Cardiology Foundation.

From the [a]Department of Translational Data Science and Informatics, Geisinger, Danville, Pennsylvania; [b]Heart Institute, Geisinger, Danville, Pennsylvania; [c]Center for Pharmacy Innovation and Outcomes, Geisinger, Danville, Pennsylvania; [d]Department of Medicine, Geisinger, Danville, Pennsylvania; [e]Phenomic Analytics and Clinical Data Core, Geisinger, Danville, Pennsylvania; [f]Department of Population Health Sciences, Geisinger, Danville, Pennsylvania; [g]Steele Institute for Health Innovation, Geisinger, Danville, Pennsylvania; [h]Department of Computer Science, Bucknell University, Lewisburg, Pennsylvania; and the [i]Department of Radiology, Geisinger, Danville, Pennsylvania. *Drs. Haggerty and Fornwalt contributed equally to this work and are joint senior authors. This work was supported by a Quality Fund Award from Geisinger Health Plan. Geisinger receives funding from Tempus for ongoing development of predictive modeling technology and commercialization. Tempus and Geisinger have jointly applied for a patent related to the work. None of the Geisinger authors has ownership interest in any of the intellectual property resulting from the partnership. The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

Manuscript received July 10, 2019; revised manuscript received January 2, 2020, accepted January 2, 2020.

ISSN 2213-1779/$36.00 https://doi.org/10.1016/j.jchf.2020.01.012

ABBREVIATIONS AND ACRONYMS

ARB = angiotensin II receptor blocker

ACE = active angiotensin-converting enzyme

AUC = area under the receiver-operating characteristic curve

BP = blood pressure

CRT = cardiac resynchronization therapy

EBBB = evidence-based beta-blocker

ECG = electrocardiogram

EHR = electronic health record

HF = heart failure

HFpEF = heart failure with preserved ejection fraction

HFrEF = heart failure with reduced ejection fraction

In response to the rising cost of chronic conditions, such as heart failure (HF), new models of health care and reimbursement are being developed (1). In these value-based care models, management is extending beyond single patient-physician encounters to instead treat disease at a population scale. The general goal of such models is to improve patient outcomes while reducing/containing costs by delivering care that keeps patients optimally managed and reduces the frequency of high cost/high acuity encounters. Optimizing this kind of management at a population level requires an effective means to identify and stratify patients in need of intervention and, ideally, identify the appropriate intervention to deploy. At present, there is a critical lack of validated, data-driven models to support these population health goals.

Data science approaches, including machine learning, are well-suited to assist with these tasks. For example, one of the first papers on this subject in 1995 showed that a neural network could use echocardiography data to predict 1-year mortality in 95 patients with HF with accuracy that was superior to a linear model or clinical judgment (2). Since then, numerous additional studies with thousands of patients have shown significant promise for machine learning to predict hospitalization (3), readmission (4), or death (3,5,6) in patients with HF.

Previously published models using machine learning for risk predictions in patients with HF have 2 primary limitations with regard to their utility in optimizing clinical population health management. First, most models have used small, systematically collected and annotated datasets, such as from a clinical trial (7), or focused on an important, but narrow, clinical setting, such as in-hospital mortality during HF hospitalization for acute decompensation (8). Such approaches, although valid and appropriate within their respective constraints, are not necessarily generalizable to a broad and heterogeneous HF population, as characterized in real-world clinical data. The second limitation is that none of the published findings using machine learning models have led to clinically relevant, actionable results. Ahmad et al. (6) reported associations between therapies and outcomes in 44,886 patients with HF, which showed that retrospective associations can be used to drive prospective predictions. Levy et al. (5) took this 1 step further to model predicted hazard ratios at a population level for the addition of therapies such as angiotensin-converting enzyme (ACE) inhibitors and beta-blockers. This type of approach may be able to generate individual patient-level predictions about those most likely to benefit from certain therapies and, thus, can be used to direct resources at a population level.

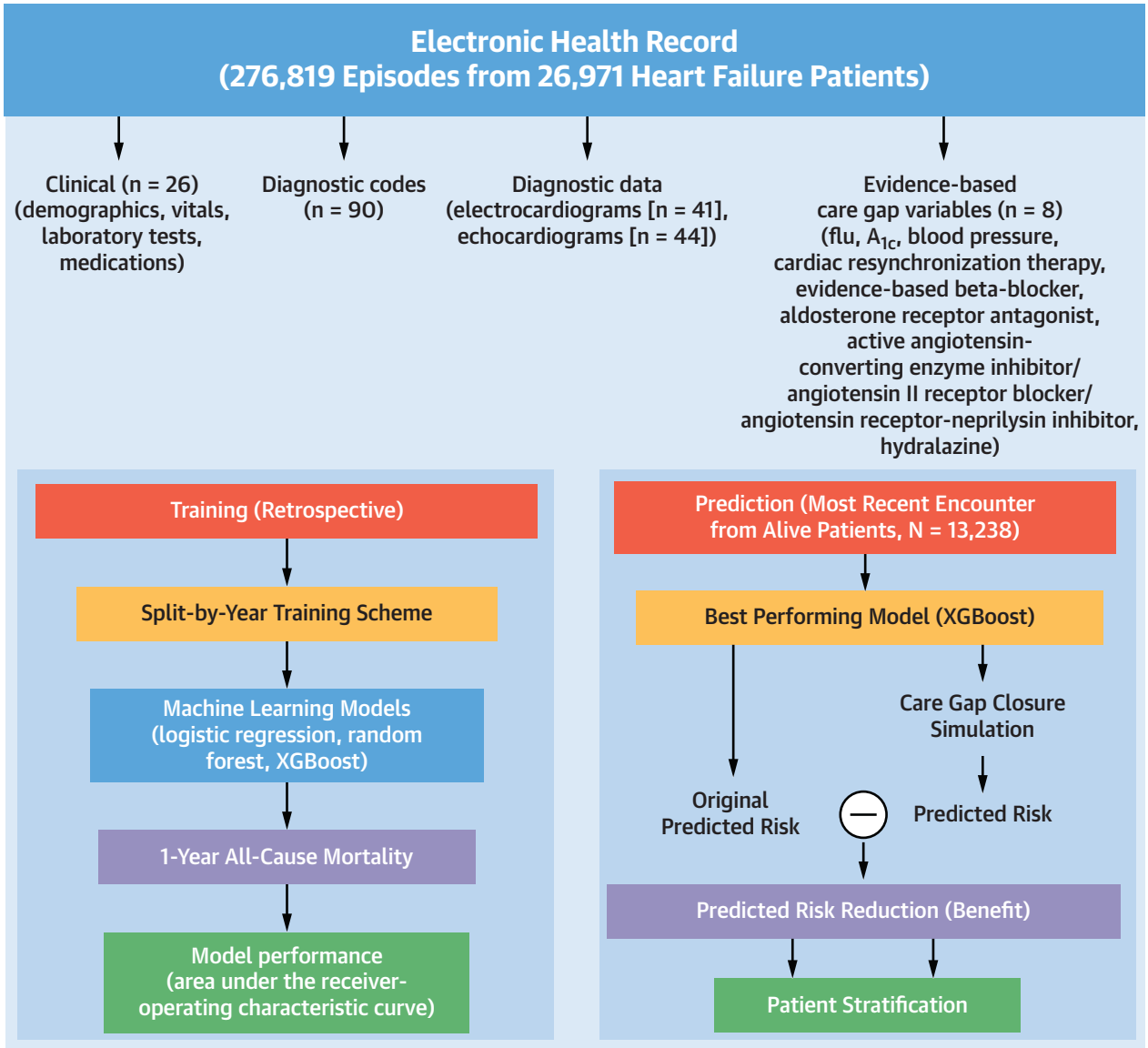
In the present study, we leverage a large 20-year retrospective dataset derived from a health system (Geisinger) that was an early adopter of electronic health record (EHR) technology to develop a predictive model for all patients with HF using machine learning. This model included a comprehensive set of input variables, including 8 care gap indicators. Importantly, this novel incorporation of evidence-based care gaps into a predictive model represents a methodology for driving clinical action from a machine learning model (not just predicting risk but predicting modifiable reduction in risk, or benefit, as a result of action). Moreover, we demonstrate how such insights might be used through population health management efforts to simultaneously stratify risk and therapeutic benefit at an individual patient level to optimally deploy health care resources.

METHODS

EHR DATA COLLECTION. Patients with HF over the last 19 years (January 2001 through November 2019) were identified from the Geisinger EHR, comprising data from 13 regional hospitals and a network of primary and specialty clinic sites. HF was defined by using the validated Electronic Medical Records and Genomics phenotype (9) and the "definite" category (i.e., probable or possible HF were not included). All clinical encounters since 6 months before the HF diagnosis date, including outpatient office visits, hospital admissions, emergency department visits, laboratory tests, and cardiac diagnostic studies (e.g., echocardiograms or electrocardiograms [ECGs]), were identified, grouped into episodes, and used as independent samples (see details in the Supplemental Appendix). Briefly, episodes were defined by consolidating data across encounters within 2 weeks of each other, up to a maximum 2-month span. To provide an external validation set, all patient data from 1 member hospital were excluded from model training and used exclusively for validation of the final model.

MODEL INPUTS. A total of 209 variables were collected from the EHR (see **Central Illustration**): 26 clinical variables: age, sex, height, weight, smoking status, heart rate, systolic and diastolic blood pressures (BPs), use of loop diuretics, antihypertensive and antidiabetic medications, and laboratory test values (hemoglobin, estimated glomerular filtration rate, creatine kinase-muscle/brain, lymphocytes,

CENTRAL ILLUSTRATION Overall Schematic



Jing, L. et al. J Am Coll Cardiol HF. 2020;■(■):■-■.

We studied 1-year all-cause mortality in a large cohort of patients with heart failure using machine learning models to integrate clinical variables, measures from diagnostic studies (e.g., echocardiography and electrocardiography) and evidence-based care gap variables from electronic health records. Mean area under the receiver-operating characteristic curve from a split-by-year training scheme was reported to evaluate model performance. The best-performing model was then used to estimate risk reduction (potential benefit) by artificially closing care gaps in a prospective prediction set and to evaluate the efficiency of benefit-driven patient prioritization.

high-density lipoprotein, low-density lipoprotein, uric acid, sodium, potassium, N-terminal pro-B-type natriuretic peptide, troponin T, hemoglobin A_{1c}, troponin I, creatinine, and total cholesterol); 90 cardiovascular-related International Classification of Diseases-Tenth Revision diagnostic codes (Supplemental Table 1); 44 nonredundant echocardiographic variables; 41 ECG measurements (such as QRS duration) and patterns (such as atrial fibrillation); and 8 care gap variables (described in the next

TABLE 1 Care Gap Definitions

Care Gap	Inclusion	Exclusion	Gap Closure
Flu vaccine	All patients eligible	Allergy	Received flu vaccine in the current flu season
BP in goal	All patients eligible	N/A	Open (not in goal) if the most recent reading and ≥1 of the 4 prior readings in the past 12 months are >130 mm Hg for systolic or >80 mm Hg for diastolic
A_{1c} in goal	Diagnosis of diabetes*	N/A	Most recent A_{1c} within the last 6 months: <8%
EBBB	Diagnosis of heart failure with most recent LVEF of <40%	• Bradycardia (heart rate of <60 beats/min by averaging up to 5 most recent readings in last 6 months) • On inotropic therapy • History of second or third degree heart block without implantable cardioverter-defibrillator or pacemaker • Hypotension (systolic pressure of <100 mm Hg by averaging last 5 readings in past 6 months) • Severe COPD or asthma • Allergy or contraindications	Currently taking EBBB
ACE inhibitor/ ARB/ARNI	Diagnosis of heart failure with most recent LVEF of <40%	• Pregnancy • History of angioedema • Hypotension • Serum creatinine of >2 in any of preceding 3 measurements • Potassium of >5 in any of previous 3 measurements • Allergy or contraindications • Currently taking hydralazine and isosorbide dinitrate/mononitrate	Currently taking ACE inhibitor or ARB or ARNI
Hydralazine	• Diagnosis of heart failure with most recent LVEF of <40% • Not currently taking ACE inhibitor/ARB/ARNI • Having any of the following conditions: pregnancy, history of angioedema, allergy or contraindications for ACE inhibitor/ARB/ARNI, serum creatinine of >1.6 in any of preceding 3 measurements, and potassium of >5 in any of preceding 3 measurements	• Hypotension • Allergy or contraindications	Currently taking combination of hydralazine and isosorbide dinitrate/mononitrate
ARA	Diagnosis of heart failure with most recent LVEF of <35%	• Hypotension • Serum creatinine of >2 in any of preceding 3 measurements • Potassium of >5 in any of preceding 3 measurements • On dialysis • Allergy or contraindications	Currently taking ARA
CRT	• Diagnosis of heart failure • LVEF of ≤35% • Left bundle branch block from ECG in last 12 months • QRS duration ≥150 ms from ECG in last 12 months	N/A	Currently have a CRT-D or CRT-P device

*Diabetes definition is described in the Supplemental Appendix.

ACE = active angiotensin-converting enzyme; ARB = angiotensin II receptor blocker; ARNI = angiotensin receptor-neprilysin inhibitor; ARA = aldosterone receptor antagonist; BP = blood pressure; COPD = chronic obstructive pulmonary disease; CRT = cardiac resynchronization therapy; CRT-D = cardiac resynchronization therapy defibrillator; CRTP = cardiac resynchronization therapy pacemaker; EBBB = evidence-based beta-blocker; LVEF = left ventricular ejection fraction; N/A = not applicable.

section). Two of the care gap variables were rules defined to modify other variables (BPs and A_{1c}), leaving 207 true input variables used in the machine learning models. Laboratory values, vital signs, ECG and echocardiographic measurements (recorded in the Xcelera database) within 12 months before the encounter date were extracted. If no measurements were available within the specified time window, the variable was set to missing. ECG measures were extracted from clinical ECG reports, which are reviewed and verified by physicians after clinical standards. EHR data preprocessing and cleaning are further detailed in the Supplemental Appendix.

CARE GAP VARIABLES. We introduced 8 evidence-based, actionable interventions (care gap variables) to study their association with patient outcomes: 1) flu vaccine administration; 2) hemoglobin A_{1c} in goal range (<8%); 3) BP in goal (<130/80 mm Hg); 4) active evidence-based beta-blocker (EBBB); 5) active ACE

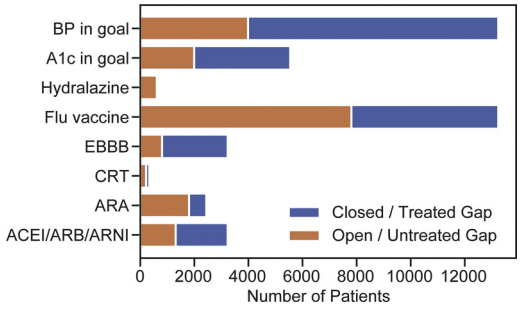
TABLE 2 Basic Demographics and Patient Characteristics

	All (276,819 Episodes From 26,971 Patients)		Validation Set (548 Episodes/ Patients)		Prediction Set (13,238 Episodes From Living Patients)	
	Value	% Missing	Value	% Missing	Value	% Missing
Age, yrs	76 (67-84)	0	73 (64-83)	0	75 (65-84)	0
Male	53	0	57	0	53	0
Smoking history	63	0	59	0	62	0
Height, cm	168 (157-175)	10	168 (160-176)	2	168 (160-175)	5
Weight, kg	84 (70-102)	2	89 (72-111)	1	87 (72-105)	2
Diastolic pressure, mm Hg	68 (60-75)	1	70 (64-78)	1	70 (62-78)	2
Systolic pressure, mm Hg	124 (112-137)	1	124 (112-138)	1	124 (113-138)	2
Heart rate, beats/min	72 (64-80)	2	72 (64-81)	1	72 (64-82)	2
Left ventricular ejection fraction, %	52 (35-57)	48	47 (32-57)	44	52 (37-57)	48
HDL, mg/dl	42 (35-52)	41	41 (35-51)	47	43 (35-53)	45
LDL, mg/dl	74 (56-97)	35	79 (57-102)	47	73 (55-97)	42
NT-proBNP, pg/ml	2,137 (766-5,567)	63	2,044 (804-4,026)	53	1,755 (591-4,695)	56
Troponin T, ng/ml	0.02 (0.01-0.06)	58	0.02 (0.01-0.05)	51	0.03 (0.02-0.06)	58

Values are median, (interquartile range) or %. *Smoking history: current or ever smoking.
HDL = high-density lipoprotein; LDL = low-density lipoprotein; NT-proBNP = N-terminal-pro hormone B-type natriuretic peptide.

inhibitor, angiotensin II receptor blocker (ARB), or angiotensin receptor-neprilysin inhibitor; 6) active aldosterone receptor antagonist; 7) active hydralazine and isosorbide dinitrate; and 8) cardiac resynchronization therapy (CRT). These care gap variables were defined with assistance from a cardiologist, a physician trained in medical informatics, and a pharmacist with HF expertise following national guidelines and recommendations for evidence-based HF therapies (10). Detailed inclusion/exclusion criteria are listed (Table 1). By definition, gaps 4 through 8 apply only to HF with reduced ejection fraction (HFrEF), whereas gaps 1 through 3 apply to all patients with HF. A blinded chart review validation of each care gap variable is detailed in Supplemental Table 2.

FIGURE 1 Care Gap Prevalence



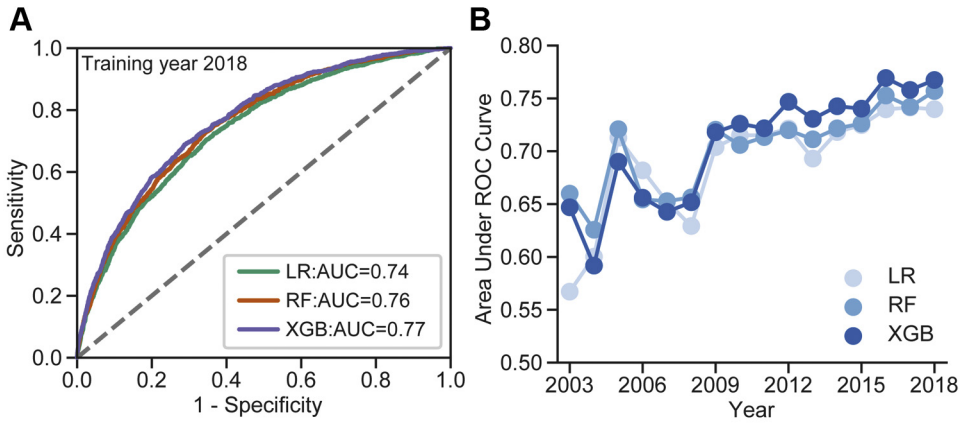
Number of patients with an open/untreated gap **(orange)** and with a closed/treated gap **(blue)** as of the most recent encounter date in living patients (N = 13,238). The sum of the orange and blue bars represents the total number of patients eligible (i.e., that fit the inclusion criteria) for that gap. These numbers are summarized in Supplemental Table 4.
ACEI = active angiotensin-converting enzyme inhibitor; ARA = aldosterone receptor antagonist; ARB = angiotensin II receptor blocker; ARNI = angiotensin receptor-neprilysin inhibitor; BP = blood pressure; CRT = cardiac resynchronization therapy; EBBB = evidence-based beta-blocker.

PRIMARY OUTCOME. We used machine learning models to predict all-cause mortality 1 year after the end of an episode. Survival duration was calculated from the date of death (cross-referenced with national death index databases monthly) or last living encounter from the EHR.

MACHINE LEARNING MODEL TRAINING AND EVALUATION. We compared performances between a linear logistic regression classifier and nonlinear models, including random forest and XGBoost (11) (a scalable gradient tree boosting system). These nonlinear models were hypothesized to improve predictive accuracy by capturing more complex, nonlinear relationships among input variables. The best-performing model was selected for subsequent analysis of care gap closure effect estimation. We evaluated models using a split-by-year cross-validation to simulate clinical deployment, as described in the Supplemental Appendix and illustrated by Supplemental Figure 1.

BENEFIT PREDICTION IN LIVING PATIENTS BY SIMULATION OF CARE GAP CLOSURE. To study the predicted effect of closing care gaps on reducing 1-year all-cause mortality, we artificially closed care gaps while keeping all other variables unchanged as follows: for binary gap variables, changing the value

FIGURE 2 Model Evaluation Using Split-by-Year Cross-Validation



(A) Receiver operating characteristic curves for year 2018 and **(B)** areas under the receiver-operating characteristic (ROC) curves (AUCs) for all years using linear logistic regression (LR) and nonlinear random forest (RF) and XGBoost (XGB).

from 1 (open/untreated) to 0 (closed/treated); for continuous variables, changing the value to goal (A_{1c}: 8% or BP: 130/80 mm Hg) if the original value was above the goal and the corresponding medication (antidiabetic or antihypertensive) to 1 (taking medication) if original value was 0. A care gap was not closed for patients who met the exclusion criteria for that care gap (i.e., a patient with bradycardia who could not be treated with EBBB) or who had a missing value (no A_{1c} test result in last 6 months or BP in last 12 months).

After the simulation, we calculated the change in risk score for each patient, that is, the difference between the baseline risk score with care gaps unchanged and the updated risk score with care gaps closed, which was further translated into an estimated benefit of reduction in estimated mortality rate. The cumulative sum of the benefit from all patients was then used to provide an estimated number of lives that could be saved by closing care gaps.

Model evaluation and care gap simulation were performed for all patients with HF and separately on patients with HF with HFrEF and preserved ejection fraction (HFpEF). Detailed methods and results of this subgroup analysis are provided in the supplement.

RESULTS

STUDY POPULATION. Within our EHR, 26,971 patients with HF who collectively underwent 276,819 episodes (median age: 76 years; 53% male) satisfied the inclusion criteria. On average, each patient had 10 episodes (interquartile range: 4 to 14). The median follow-up duration was 3.4 years (interquartile range: 1.3 to 6.4 years) using reverse Kaplan-Meier estimation (12), and 13,733 (51%) patients had a recorded death. The external validation set contained 548 episodes/patients from a separate Geisinger hospital as of January 1, 2018 (the cutoff date for the 2018 model), of which 42 (8%) died within 1 year. **Table 2** and Supplemental Table 3 show summary statistics.

Of the 13,238 patients who were living as of November 16, 2019, there were 3,772 (28%) with HFrEF and 6,784 (51%) with HFpEF. The remaining patients had either midrange EF (40% to 50%; 1,424; 11%) or no available EF measurement (1,258; 10%). A total of 10,516 (79%) had at least 1 open care gap, and 788 (6%) had 4 or more care gaps open as of their most recent clinical episode. **Figure 1** shows the number of patients for each gap for which the gap was open/untreated (orange) or closed/treated (blue). The sum of orange and blue represents the number of patients who were eligible for the gap (i.e., who fit the inclusion criteria). Depending on the gap, 25% to 91% of eligible patients had an open gap (additional details available in Supplemental Table 4).

ACCURACY FOR PREDICTING ALL-CAUSE MORTALITY USING MACHINE LEARNING. The split-by-year cross-validation showed that performance was highly variable in early years (before 2009) for all 3 machine learning methods (**Figure 2**), likely due to small sample sizes (Supplemental Table 5) and more missingness (Supplemental Table 3). The performance improved in subsequent years, and the nonlinear

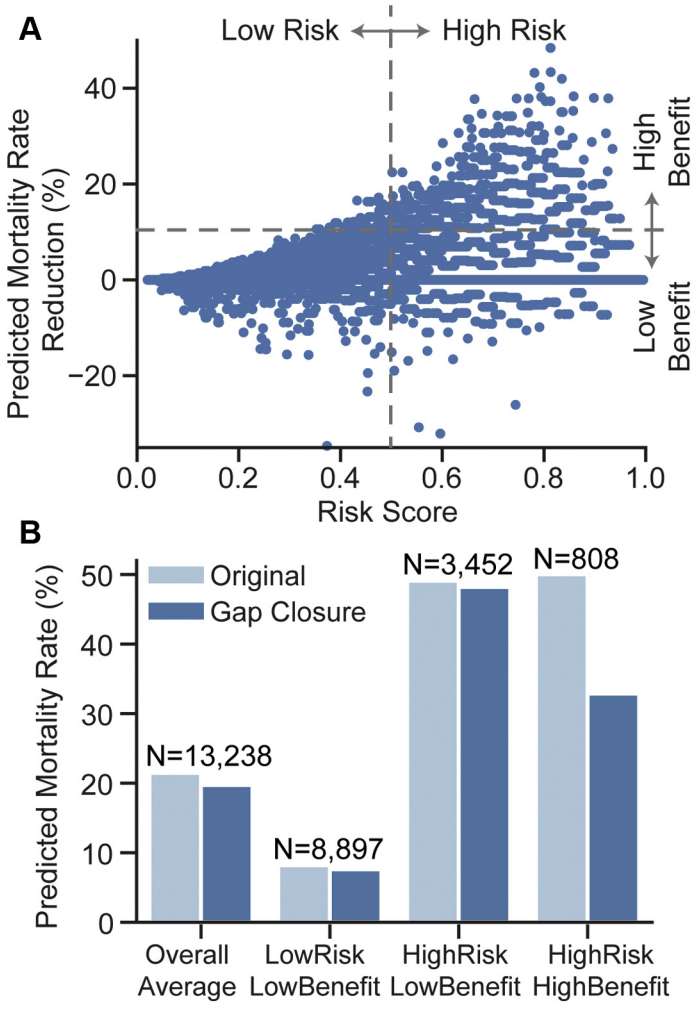
XGBoost model consistently achieved the best area under the receiver-operating characteristic curve (AUC).

In the final and largest test set (year 2018), all 3 models predicted 1-year all-cause mortality with AUCs above 0.70, superior to the performance of the Seattle HF Model (AUC: 0.57). The nonlinear models achieved higher AUCs (random forest: 0.76; XGBoost: 0.77) compared with linear logistic regression (0.74). Results were similar in patients with HFrEF and HFpEF (Supplemental Figure 2). Additionally, in the holdout validation set of 548 episodes/patients from a separate hospital, the XGBoost model had an AUC of 0.78. Calibration on the 2018 test set showed that the XGBoost model had a tendency to slightly overestimate risk (Supplemental Figure 3).

PREDICTING BENEFIT OF CLOSING CARE GAPS. A final model was fit on all training samples (all but the most recent episodes for currently living patients) by using XGBoost to predict the benefit of closing care gaps in the living patients based on data from their most recent episode. The distribution of risk scores is available in the supplement (Supplemental Figure 4). Of the 13,238 living patients, based on the estimated mortality rate, 2,844 (21.5%) patients were predicted to die within 1 year. Simulating closure of the 8 care gaps resulted in 2,613 (19.7%) patients being predicted to die within 1 year. Hence, the aggregate predicted absolute mortality rate reduction was 1.7% (individual patient range -35% to 48%, absolute), and 231 (8.1% of 2,844) additional patients were predicted to survive beyond 1 year assuming all 8 care gaps could be closed. Of these 231 patients, 102 of them had HFrEF and 87 had HFpEF.

We further investigated the relationship between risk and benefit by comparing the predicted benefits among several subgroups. **Figure 3** shows that the overall average benefit ("Overall Average") was predicted to be relatively small and was primarily driven by the large group of patients with low mortality risk at baseline (risk score: ≤0.5) and low benefit after closing the care gaps (≤10% reduction in mortality rate) ("Low Risk, Low Benefit"). There was, however, a subgroup of patients predicted to have high mortality risk (risk score: >0.5) who were also predicted to have high benefit after closing gaps (>10% reduction in mortality rate; "High Risk, High Benefit"). However, not all high-risk patients were predicted to have high benefit, as evidenced by another subgroup of patients with similarly high baseline risk but minimal risk reduction after closing the care gaps ("High Risk, Low Benefit"). The distribution of patients with HFrEF and HFpEF according to this categorization is shown in Supplemental Figure 5. Compared with the low-benefit group, high-benefit patients had, in general, more open care gaps, poorer blood pressure and A_{1c} control, lower EF, and higher left ventricular dimensions (Supplemental Table 6).

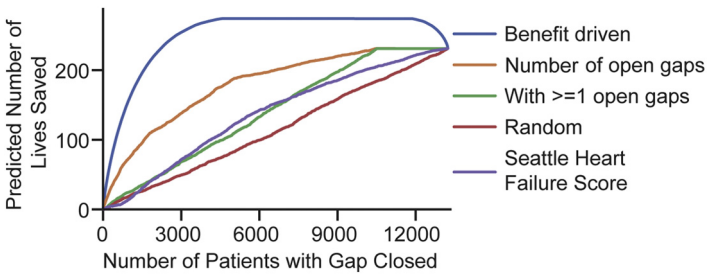
FIGURE 3 Relationship Between Predicted Mortality Risk and Benefit (Risk Reduction)



(A) Scatterplot of risk score and corresponding benefit for individual patients in the prediction set (N = 13,238). Negative reductions in mortality rate reflect a detrimental effect of closing care gaps on mortality risk, as predicted by the XGBoost model, in a small proportion of patients. **(B)** Average mortality rate before and after care gap closure simulation in selected groups. Risk is not equivalent to benefit because patients at similarly high mortality risk levels do not have the same predicted benefit of closing care gaps. Additionally, 81 patients had relatively low risk and high benefit (data not shown).

PATIENT PRIORITIZATION TO EFFICIENTLY CLOSE CARE GAPS THROUGH POPULATION HEALTH MANAGEMENT. Assuming that a population health management team could be assembled and deployed to

FIGURE 4 Simulation of Care Gap Closure Using XGBoost



Prioritization of patients according to predicted benefit is the most optimal resource allocation method based on having the highest predicted patient survival (*y*-axis) relative to the number of patients needed to treat (*x*-axis). The slopes of the plotted lines are inversely proportional to the number needed to treat, and thus, steeper lines represent more efficient patient prioritization. The small drop in lives saved at the far right-hand side of the line corresponding to the "Benefit Driven" model reflects the patients for whom closing the care gaps had a predicted negative impact on mortality risk, as shown in Figure 3A.

close care gaps, the efficiency of its efforts would depend on effective guidance as to which patients to target first in a rank-ordered fashion. To demonstrate the potential value of machine learning to optimize care team resource deployment in this setting, we plotted the number of lives predicted to be saved versus the number of patients receiving an intervention (in which all eligible gaps were subsequently assumed closed) for several different prioritization strategies:

1. Random prioritization.
2. Randomly prioritizing any patient with at least 1 open care gap.
3. Rank-ordering patients by the number of open care gaps.
4. Stratifying patients using the Seattle Heart Failure risk score (5).
5. Stratifying patients according to the XGBoost model's predicted benefit (i.e., mortality risk reduction).

Figure 4 shows that the proposed machine learning benefit stratification model (strategy 5) was the most efficient. That is, benefit stratification had the steepest slope of any prioritization strategy and, thus, in a resource-constrained environment, maximized the predicted total number of lives saved for a given number of patient interventions.

DISCUSSION

Optimized population health management demands novel, data-driven approaches for allocating health care resources, particularly within new value-based care models. This study has made considerable advances toward the development of such an approach for HF that combines extensively and carefully curated clinical data and machine learning. The model incorporates important clinical variables, quantitative measures from common diagnostic studies such as echocardiography and ECG, and evidence-based interventions in the form of care gaps. Our results show that a machine learning model with these inputs can achieve good accuracy to predict 1-year all-cause mortality in patients with HF.

Several studies have been published in recent years using machine learning to predict outcomes (mostly survival) in patients with HF, as summarized by Tripoliti et al. (13). These studies used various methods, from traditional classification (e.g., logistic regression, random forest) to custom-developed algorithms (contrast pattern-aided logistic regression with probabilistic loss function). The reported performances (AUC) vary from 0.61 to 0.94, while mostly centered around 0.75 to 0.8. Our model performance is, therefore, comparable to these prior studies, but with critical differences that enhance the clinical utility and generalizability of our model. Namely, we used large-scale EHR data with comprehensive features from a general HF population and leveraged machine learning with a novel split-by-year design for optimal evaluation and deployment in a real-world clinical setting. Detailed comparison with previous studies is included in the Supplemental Appendix.

Furthermore, our explicit representation of clinical care gaps in the model represents a new paradigm for guiding clinical action with machine learning. Specifically, we showed how these care gap inputs can be used to predict risk reduction associated with specific interventions on an individual patient level. These model predictions can provide guidance to integrated health systems working to optimally distribute scarce clinical resources (e.g., care teams) to patients who need them the most. Importantly, most published models and clinical scoring systems rely heavily on risk prediction, which could be used to prioritize distribution of health care resources. However, risk is not equivalent to modifiable risk (i.e., benefit), and thus, patients with identical risk of 1-year mortality can have very different predicted benefit from interventions. Therefore, deployment of resources based simply on risk is unlikely to be optimal. We demonstrated support for this claim by showing the

superiority of our model's predicted performance over the Seattle Heart Failure score for prioritizing patient interventions.

Despite the fact that these interventions (care gaps) are recommended in national guidelines based on demonstrated benefit—for example, even flu vaccination has been associated with decreased all-cause mortality in HF (14)—the prevalence of open care gaps remains a significant problem in medicine. For example, in patients with HF, therapies proven to prolong life are used at staggeringly low rates: only 57% are receiving ACE inhibitors, 34% are receiving evidence-based beta blockers, and 32% are receiving mineralocorticoid antagonists (15). Additionally, although some gaps can be easily addressed in a single setting (e.g., flu vaccine), others are more difficult to manage and require close monitoring and frequent follow-up (e.g., BP and A_{1c} control). This problem is highly complex and unlikely to be solved by relying on individual providers to change practice. However, new value-based care models can likely address this problem more effectively by creating organized care teams. These teams will require accurate, reliable data science, such as that presented in this article, to successfully allocate resources.

A small portion of patients had a negative predicted reduction in mortality rate (1.8% of the living patients with benefit of <-5%), indicating that closing gaps for these patients was predicted to have a detrimental effect on mortality risk. This is likely due to the nonlinear relationship between care gaps such as A_{1c} or BP and mortality. Despite the evidence-based guidelines showing that lower BPs are associated with reduced risk of adverse events in HF (16), the so-called BP paradox has been noted in multiple studies where lower BP or pronounced changes in BP (increases or decreases) was associated with poor outcomes (17,18). Additional investigation is warranted to fully understand the role of BP and to determine optimal BP targets in HF.

It is critically important to note that this concept needs prospective evaluation because, currently, our models are based on association, not causation. It is logical to infer some causative relationship between evidence-based therapies and survival, which has been demonstrated in many prior studies (10,19–21). However, this needs to be evaluated prospectively, and we have thus launched a randomized prospective study (NCT03804606).

STUDY LIMITATIONS. First, by treating each episode as an independent training sample, longitudinal information for individual patients was not captured and, thus, could compromise model performance. However, by using the split-by-year cross-validation approach, a patient's historic information was still used in training to make predictions from a current episode. Future work will explore the use of approaches to more optimally leverage longitudinal information.

Second, the care gaps selected in the current study primarily focus on patients with HFrEF with a few exceptions, because of lack of evidence-based treatment for HFpEF. Some important treatments for HF were not included for various reasons, such as newer therapies (e.g., ivabradine) for which we do not yet have enough retrospective data or difficulty in capturing the therapy from structured EHR data (e.g., implantable cardioverter-defibrillator devices). Additionally, we did not account for optimized dosing of medications in the current study. Future studies will leverage natural-language processing to more accurately capture these features from unstructured data and explore their potential impact on mortality risk.

Third, the predicted benefit is for short-term mortality risk reduction, which is highly relevant because of the high 1-year mortality rate in HF. However, this may not account for longer-term benefits of treatment, and future investigations are required. Finally, we used EHR data from a single health system, which may limit generalizability. An external, independent dataset from another large health care system may help further validate and improve our model. However, our dataset represents a broad, heterogeneous population covering a large geographic area populated by approximately 3 million people served by 13 hospitals and >100 clinics. Moreover, we validated our model using a holdout from a geographically separate independent hospital within our system.

CONCLUSIONS

We presented a machine learning model to predict 1-year all-cause mortality with good accuracy in a large cohort of patients with HF. Our results leveraging 276,819 episodes from nearly 27,000 patients show that these models can be used to not only risk-stratify

patients but also efficiently prioritize patients based on predicted benefits of clinically relevant evidence-based interventions. This approach will likely prove useful for assisting HF-population health management teams within new value-based payment models.

ACKNOWLEDGMENTS The authors acknowledge Paul M. Berry, Susan A. Kilbride, and Christopher D. Nevius for their assistance with the manual chart reviews for care gap validation.

ADDRESS FOR CORRESPONDENCE: Dr. Brandon K. Fornwalt, Department of Imaging Science and Innovation, Geisinger, 100 North Academy Avenue, Danville, Pennsylvania 17822-4400. E-mail: bkf@gatech.edu.

PERSPECTIVES

COMPETENCY IN PATIENT CARE AND PROCEDURAL SKILLS: A machine learning model can integrate vast amounts of clinically acquired electronic health record data to optimally direct heart failure population health management in new value-based care models.

TRANSLATIONAL OUTLOOK: The clinical adoption of data science and machine learning models to optimize population health management may ultimately improve survival and care delivery in populations of patients with HF.

REFERENCES

1. O'Connor CM. Bundle up for value-based heart failure care. J Am Coll Cardiol HF 2015;3:931–2.

2. Ortiz J, Ghefter CG, Silva CE, Sabbatini RM. One-year mortality prognosis in heart failure: a neural network approach based on echocardiographic data. J Am Coll Cardiol 1995;26:1586–93.

3. Shah SJ, Katz DH, Selvaraj S, et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation 2015;131:269–79.

4. Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of machine learning techniques for heart failure readmissions. Circ Cardiovasc Qual Outcomes 2016;9:629–40.

5. Levy WC, Mozaffarian D, Linker DT, et al. The Seattle Heart Failure Model: prediction of survival in heart failure. Circulation 2006;113:1424–33.

6. Ahmad T, Lund LH, Rao P, et al. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. J Am Heart Assoc 2018;7:e008081.

7. Subramanian D, Subramanian V, Deswal A, Mann DL. New predictive models of heart failure mortality using time-series measurements and ensemble models. Circ Hear Fail 2011;4:456–62.

8. Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin WJ. Risk stratification for in-hospital mortality in acutely decompensated heart failure:classification and regression tree analysis. JAMA 2005;293:572–80.

9. Bielinski SJ. Heart Failure (HF) With Differentiation Between Preserved and Reduced Ejection Fraction. PheKB. Available at: https://phekb.org/phenotype/heart-failure-hf-differentiation-between-preserved-and-reduced-ejection-fraction. Accessed December 19, 2019.

10. Yancy CW, Jessup M, Bozkurt B, et al. 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure. J Am Coll Cardiol 2017;70:776–803.

11. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min - KDD '16. March 2016: 785-94.

12. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials 1996;17:343–6.

13. Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Comput Struct Biotechnol J 2017;15:26–47.

14. Modin D, Jørgensen ME, Gislason G, et al. Influenza vaccine in heart failure. Circulation 2019;139:575–86.

15. Callender T, Woodward M, Roth G, et al. Heart failure care in low- and middle-income countries: a systematic review and meta-analysis. PLoS Med 2014;11:e1001699.

16. SPRINT Research Group, Wright JT, Williamson JD, et al. A randomized trial of intensive versus standard blood-pressure control. N Engl J Med 2015;373:2103–16.

17. Ventura HO, Messerli FH, Lavie CJ. Observations on the blood pressure paradox in heart failure. Eur J Heart Fail 2017;19:843–5.

18. Schmid FA, Schlager O, Keller P, et al. Prognostic value of long-term blood pressure changes in patients with chronic heart failure. Eur J Heart Fail 2017;19:837–42.

19. Klapholz M. β-Blocker use for the stages of heart failure. Mayo Clin Proc 2009;84:718–29.

20. Cheng JWM, Nayar M. A review of heart failure management in the elderly population. Am J Geriatr Pharmacother 2009;7:233–49.

21. Fonarow GC. A review of evidence-based beta-blockers in special populations with heart failure. Rev Cardiovasc Med 2008;9:84–95.

KEY WORDS data science, electronic health records, population health

APPENDIX For an expanded Methods section and supplemental tables and figures, please see the online version of this paper.