Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Generalizing Major Adverse Cardiovascular Events (MACE) Prediction Across Study Designs Using Ensemble Transfer Learning in 4.2 million US Veterans

Sayera Dhaubhadel
Los Alamos National Laboratory
Ruy M. Ribeiro
Los Alamos National Laboratory
Kumkum Ganguly
Los Alamos National Laboratory
Nicolas W. Hengartner
Los Alamos National Laboratory
Tanmoy Bhattacharya
Los Alamos National Laboratory
Beauty Kolade
Los Alamos National Laboratory
Judith D. Cohn
Los Alamos National Laboratory
James M. Hyman
Los Alamos National Laboratory
Khushbu Agarwal
Pacific Northwest National Laboratory
Kelly Cho
VA Boston Healthcare System
Lauren Costa
VA Boston Healthcare System
Yuk-Lam Ho
VA Boston Healthcare System
Jin Zhou
University of California, Los Angeles
Allison E. Murata
VA Cooperative Studies Program Clinical Research Pharmacy Coordinating Center
Glen H. Murata
New Mexico VA Health Care System

Jason L. Vassy
VA Boston Healthcare System
Daniel C. Posner
VA Boston Healthcare System
J. Michael Gaziano
VA Boston Healthcare System
Yan V. Sun
Atlanta VA Health Care System
Peter W. Wilson
Atlanta VA Health Care System
Ravi Madduri
Argonne National Laboratory
Amy C. Justice
VA Connecticut Healthcare System
Phil Tsao
Stanford University
Christopher J. O'Donnell
VA Boston Healthcare System
Scott Damrauer
Philadelphia VA Medical Center
Benjamin H. McMahon
mcmahon@lanl.gov

Los Alamos National Laboratory

Article

Keywords:

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10395960/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Christopher O'Donnell is an employee of Novartis Pharmaceutical Company.

Generalizing Major Adverse Cardiovascular Events (MACE) Prediction Across Study Designs Using Ensemble Transfer Learning in 4.2 million US Veterans

Sayera Dhaubhadel[1,*], Ruy M. Ribeiro[1], Kumkum Ganguly[1], Nicolas W. Hengartner[1], Tanmoy Bhattacharya[1], Beauty Kolade[1], Judith D. Cohn[1], James M. Hyman[1], Khushbu Agarwal[2], Kelly Cho[3,4], Lauren Costa[4], Yuk-Lam Ho[4], Jin Zhou[5], Allison E. Murata[6], Glen H. Murata[7], Jason L. Vassy[3,4], Daniel C. Posner[4], J. Michael Gaziano[3,4], Yan V. Sun[8,9], Peter W. Wilson[8,9], Ravi Madduri[10], Amy C. Justice[11,12], Phil Tsao[13,14], Christopher J. O'Donnell[3,4], Scott Damrauer[15,16], and Benjamin H. McMahon[1,*]

[1]Los Alamos National Laboratory, Los Alamos, New Mexico, USA.
[2]Pacific Northwest National Laboratory, Richland, Washington, USA.
[3]Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
[4]VA Boston Health Care System, Boston, Massachusetts, USA.
[5]Fielding School of Public Health, University of California at Los Angeles, Los Angeles, California, USA.
[6]VA Cooperative Studies Program Clinical Research Pharmacy Coordinating Center, Albuquerque, New Mexico, USA.
[7]New Mexico VA Health Care System, Albuquerque, New Mexico, USA.
[8]Veterans Affairs Atlanta Health Care System, Decatur, Georgia, USA.
[9]Emory University, Atlanta, Georgia, USA.
[10]Argonne National Laboratory, Lemont, Illinois, USA
[11]Yale University School of Medicine, New Haven, Connecticut, USA.
[12]Veterans Administration Connecticut Healthcare System, West Haven, Connecticut, USA.
[13]Stanford University School of Medicine, Stanford, California, USA.
[14]VA Palo Alto Health Care System, Palo Alto, California, USA.
[15]University of Pennsylvania, Philadelphia, Pennsylvania, USA.
[16]Corporal Michael Crescenz VA Medical Center, Philadelphia, Pennsylvania, USA.
[*]mcmahon@lanl.gov; sayeradbl@lanl.gov

ABSTRACT

Accurate prediction of major adverse cardiovascular events (MACE) from electronic medical records (EMRs) to identify high-risk patients is fundamental for informing patientcare guidelines and policies. Yet, existing prediction models often fail to generalize because different epidemiological study designs capture distinct aspects of disease biology and observational biases. We develop an ensemble transfer learning (EnsTL) framework that leverages complementary information from nested retrospective case control (Rcc) and prospective screening cohorts (PCs) to improve cardiovascular risk prediction. Using longitudinal EMR data from 4.2 million U.S. Veterans, we trained multiple machine learning (ML) models on 247 clinically curated predictors spanning upto 7.5 years before prediction. Rcc cohort captured biological drivers of cardiovascular disease (CVD) more effectively, while PCs better reflected real-world screening populations. Our EnsTL model, combining and finetuning these complementary models, produced robust, well-calibrated predictions in the prospective calendar 2017 (C17) cohort with C-statistics 0.745–0.754, while preserving clinically meaningful associations. This framework enables more reliable population screening, facilitates adaptation of prediction models across healthcare settings, and provides a general strategy for developing clinically deployable EMR-based risk prediction models.

Introduction

Cardiovascular disease (CVD) is implicated in ∼25% of deaths, both in the US[1] and worldwide[2]. Epidemiological studies of CVD consistently associate it with non-modifiable risk factors such as age and sex, modifiable biomarkers for low-density lipoprotein (LDL), cholesterol, hypertension, and diabetes, and modifiable risk factors such as obesity, smoking, and lack of physical activity[2–6]. Many of these pathophysiological CVD risk factors co-occur, interact with one another, and some clusters are defined as a metabolic syndrome[7]. There is substantial heterogeneity in how these common risk factors impact patients' morbidity and mortality, so a better understanding of their interplay may have important clinical implications.[8]

Several studies have assessed the value of ML/AI in training and prediction of CVD outcomes in the prospective screening cohorts (PCs). Singh *et al.*[9] predicted deterioration of renal function in a cohort of 6,435 individuals from EMRs with a variety of longitudinal ML models. They concluded that longitudinal, multi-task predictors showed great promise, but emphasized that the details of problem

formulation significantly impacted results. Zhao, *et al.*[10] compared multiple ML techniques applied to several data sets in analyses of 109,490 CVD patients and concluded there was a 6 percentage point accuracy improvement between the re-weighted American Heart Association - Pooled Cohort Equations (AHA-PCE)-like model[11] and the best high-dimensional ML model. Steele, *et al.*[12] predicted coronary artery disease (CAD) mortality in a cohort of 80,000 people, concluding that ML models worked well without imputation, that random forest models did not outperform Cox models, and that variable selection techniques could identify important, novel, predictors in a data-driven manner. More recently, a US Department of Veterans Healthcare Administration (VHA) data set of 1.7 million patients with 66,602 atherosclerotic cardiovascular disease (ASCVD) events during 5-year follow-up was used to assess the impact of statins on both calculated CVD risk and outcomes[13]. Other studies comparing anti-hyperglycemic agents such as SGLT2 inhibitors[14] or GLP-1 receptor agonists[15] have appeared recently, relying on linear models and well-established study designs. Using the Million Veteran Program (MVP) genetics data from 79,151 patients, Vassy *et al.* examined the added value of polygenic risk scores to existing risk calculations.[16] Hippisley-Cox identified seven mental health and cancer-based subgroups in the QRISK4 prediction of CVD risk in 17 million adults from the UK[17].

Advances in machine learning (ML) and artificial intelligence (AI) have been applied to EMRs in a variety of ways,[18] including detailed neural network architectures proposed to shape the logic of analysis[19]. Contrary to expectations, these more sophisticated methods have often produced results only comparable to or even lagging behind traditional methodologies in providing robust healthcare solutions[20]. Many of the challenges in making practical use of AI and ML for medical outcomes prediction could be anticipated from the general difficulties associated with deep learning methodologies, as described by Marcus[21]. In particular, these methods are data-hungry, opaque to feature importance, and difficult to integrate with prior knowledge. Moreover, they work poorly with hierarchical dependencies, assume stable relationships among variables and thus will be of questionable generalizability[22]. Sculley, *et al.*[23] argued that these aspects of deep learning models lead to substantial hidden costs in deployment, as they are brittle and often need to be re-trained with fairly innocuous changes in details of the input data stream. On the other hand, numerous statistical techniques have been developed to create models that inform clinical practice, including logistic regression[5], Kaplan-Meier analysis[24] and Cox proportional hazard models[4,25]. Researchers have noted the critical importance of interaction terms[26], variable definitions[27,28], cohort composition[5], combined outcomes[29], study design[30], case-control matching,[31,32] and the efficiency of cross-sectional pooling[25].

In this work, we develop an ensemble transfer learning (EnsTL) framework that combines complementary epidemiological study designs/cohorts to improve both the accuracy and clinical applicability of EMR-based CVD risk prediction. This framework integrates the ML methods of ensemble modeling[33–35] and transfer learning (TL)[36,37] in a way that utilizes many of the above-mentioned statistical techniques and study designs/cohorts in clinical epidemiology to navigate the high-dimensional CVD risk prediction problem using observational data. Leveraging the scale and longitudinal richness of the VHA electronic medical records (EMRs)[38], we integrate ensembles of linear and nonlinear ML models trained across nested retrospective case control (Rcc) and prospective screening cohorts (PCs) to separate stable biological drivers of cardiovascular disease (CVD) from study-design-specific healthcare utilization patterns. By transferring complementary information between multiple study designs/cohorts and aggregating different model representations, our framework aims to produce prediction models that are both clinically interpretable and generalizable across patient populations. Beyond improving prediction of MACE, this work provides a practical methodology for developing robust clinical decision-support models that can be adapted to new healthcare systems, evolving patient populations, and future cardiovascular applications.

Results

Cohorts and study designs

Rcc cohorts are valuable for identifying biological drivers of disease, however both screening of patients for disease incidence and comparison of model performance across studies requires a prospective cohort[30], such as our C17. Figure 1 shows important attributes of our Rcc and prospective C17 study designs. For both study designs, the longitudinal variables spanning 7.5 years before the time of prediction are recorded in eight time-bins of increasing width as the time before prediction increases. Diagnoses and treatments (procedures and medications) are recorded as either present or absent in each time bin, while observations (labs and numeric vital signs) use the median value in each time bin, which is itself assigned to a physiological category and referenced to normal. Missing observations for the labs and vitals predictors were imputed (see Methods) with additional variables indicating whether or not the variable was missing altogether.

Table S3 compares subject characteristics across the four cohorts of individuals we created. Motivation and detailed descriptions for each are provided in the Methods section. We have between 210,084 and 1,078,498 cases across study designs (top row). As is common in the VA systems, Females, African-Americans, and Hispanics have age distributions that differ from the overall average, where white males are predominant. Details characterizing these distributions and the breakdown of cases across the component outcomes (acute myocardial infarction (AMI), stroke, and ASCVD death) and demographics are also provided in table S3.

Model scores and ensemble transfer learning

Figure 2 illustrates our ensemble transfer learning process, evaluated on the prospective C17 cohort. Accuracy or C-statistic comparison is meaningful only when assessed on a single prospective cohort that matches the natural distribution of outcomes[39] that is used when screening patients[30], such as our C17 for MACE risk prediction. The left plot shows C-statistic scores from fourteen models predicting MACE outcome **on the C17 cohort** with the same set of 247 variables (including 4 static and 243 longitudinal variables). The blue bars shows the performance of the models as trained, while the orange bars on top show improvement in performance when the models are finetuned on the C17 cohort. Our baseline model, shown at far left, approximates the commonly used AHA-PCE model and achieves a

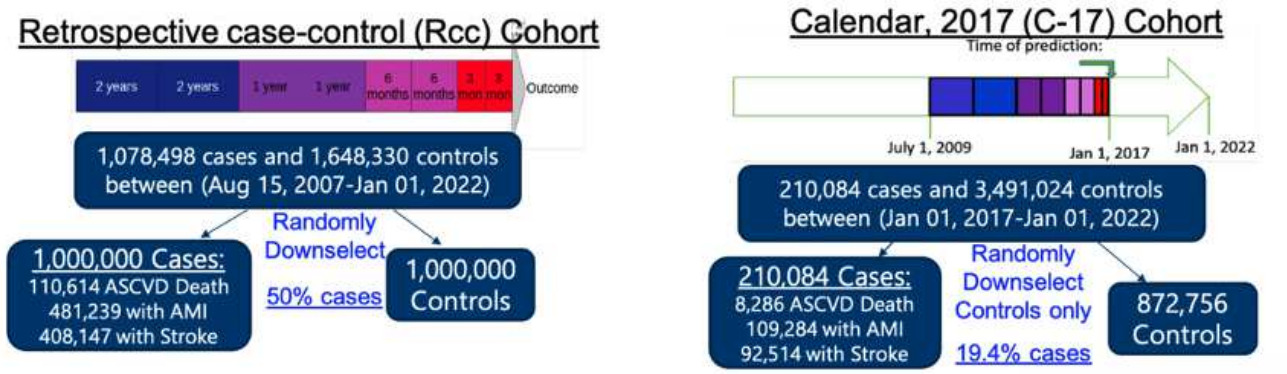


Figure 1. (**left**) Rcc study design – cases were defined as first occurrence (after August 15 of 2007) of stroke or heart attacks (AMI) from inpatient, outpatient, or fee-for-service diagnostic codes or death reported in the NDI with underlying or contributing cause ascribed to atherosclerotic cardiovascular disease (ASCVD) death. Controls were matched on year of birth and were selected by randomly choosing a visit among patients who were not cases and not excluded by having an outcome prior to July 1, 2007. Choosing a visit created a similar overall level of healthcare utilization among cases and controls. The outcome date was chosen as the date of the MACE event or the randomly selected visit, and the time of prediction was set to 45 days prior to the event. (**right**) Prospective C17 study design – all veterans with an inpatient, outpatient, or fee-for-service diagnostic code recorded after January 8, 2017 were included in the cohort, the time of prediction was set to January 1, 2017, and all patients with a recorded MACE event prior to the time of prediction were excluded from consideration. Controls were down-sampled by a factor of four to created a nested cohort. Note that the time bins and variables were defined consistently so that models trained on one design can be used on another. Descriptions of two prospective time-to-event cohorts (Cox-LDL and Cox-Visit) and more details on our study designs are provided in the Methods section.

C-statistic of 0.701. The next four models were trained with our Rcc cohort, matched on age and visits, but using four different prediction methodologies: a fully connected neural network (NN), random forest (RF), TabNet (a neural net model with an attention scheme), and a logistic regression with LASSO model selection (GLM-LASSO). Because patient age and level of healthcare utilization are confounded[40] and strongly interacting[26] with all variables, matching on age and visits forces these models to rely more heavily on biological drivers of CVD[30] and the nonlinear models more accurately capture interactions among variables. Consequently, we also trained two Cox models on up to twelve years of follow-up and our same four prediction methodologies directly on our C17 cohort, predicting five-year risk.

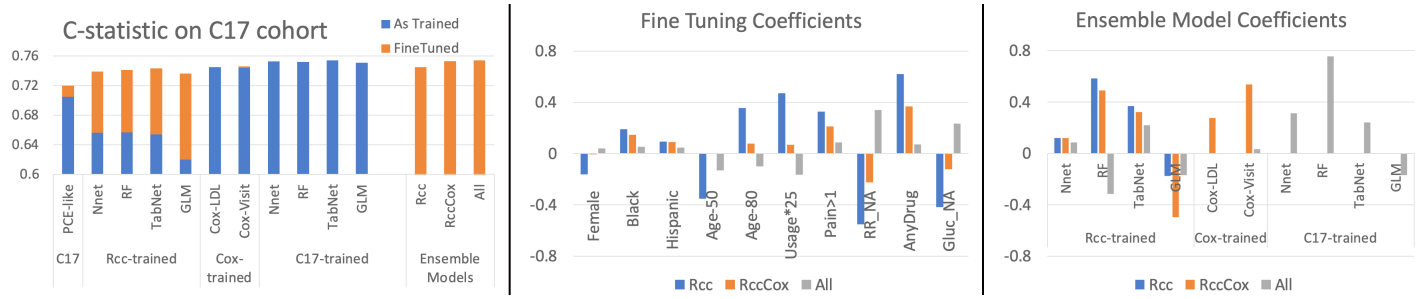


Figure 2. **(left)** C-statistics for ten models predicting MACE outcome, compared to our baseline PCE-like model shown at left. The models were created with diverse cohorts and prediction methodologies, described in text. **(middle)** Logistic regression coefficients for ten variables used in the ensemble transfer learning process. These same variables were also used to fine-tune models before being evaluated for accuracy on the C17 cohort, with blue bars in the left figure indicating accuracy before fine tuning and orange bars after. **(right)** Coefficients of each of the model scores, used together with the fine tuning coefficients in the middle panel, defining our three ensemble transfer-learned models: EnsRcc, EnsRccCox, and EnsAll.

Comparison of the blue bars shows that models trained on the matched Rcc cohort perform quite poorly when directly predicting on C17. Finetuning the models using only ten variables that are chosen to enable optimized treatment of age and healthcare utilization patterns, and recalibration for demographic groups resulted substantial improvements in the C-statistic scores. The orange bars in the left-most panel of Figure 2 show that this finetuning, or transfer learning process, enables the Rcc-trained models to predict outcomes in the C17 cohort nearly as well as the models directly trained on C17. Importantly, we see from Supplementary Table S4 that while Rcc-trained models can be fine-tuned to predict on C17 (C-statistic of 0.73 vs 0.75), the reverse process does not work as less well (C-statistic of 0.72 vs. 0.66).

We constructed three ensemble models from different sets of the base models. The first, *EnsRcc*, was constructed from the four Rcc-trained models. The second ensemble model (*EnsRccCox*) adds the two Cox models, and the third ensemble model (*EnsAll*) uses all of the base models except AHA-PCE-like. Each model was finetuned with the same ten variables described above, and the specific composition of each model is characterized in Figure 2, with the model coefficients in the right panel and the fine-tuning coefficients in the middle panel. While the same ten variables shown in the middle panel of Figure 2 were used in all cases of finetuning, the values

of the coefficients vary for the three ensemble models. All three ensemble models have significant weights on the Rcc-trained models, although *EnsAll* relies most heavily on the C17-RF model. Not surprisingly, the middle panel of Figure 2, shows significant dependence of our finetuned ensemble models on the two matched variables, age and healthcare utilization (represented by Usage, NAs, Any drug prescription). For completeness, we provide in Figure S2 the age-dependence for the four GLM-LASSO base models (for Rcc, C17, Visit-DS, LDL-DS cohorts).

These fourteen ML models show relatively similar performance in terms of C-statistic (after finetuning on the target C17 cohort), yet come from very different study designs and prediction methodologies. For the remainder of this work, we examine the properties of the models in detail to better understand how to extract biological mechanisms, understand biases, and characterize how our ensemble transfer learning approach enables latent spaces from multiple MACE models to improve predictions on related problems/cohorts. In particular, we examine the logistic regression coefficients from the GLM-LASSO model trained on Rcc cohort and compare them to the coefficients from the GLM-LASSO model trained on C17 cohort. We then examine how calibration and subgroup performance varies across four selected models: 1) baseline AHA-PCE-like model, 2) Rcc GLM-LASSO, 3) *EnsRcc*, consisting entirely of models trained on the Rcc cohort, and 4) NN trained on the C17 prospective cohort.

Model coefficients and biological drivers

Figure 3 shows logistic regression coefficients from the Rcc-trained model with *LASSO* model selection (GLM-LASSO) for the labs, diagnoses, and medications associated with metabolic syndrome and expected to be biological drivers of our MACE outcome. The other half of the coefficients are shown in Figure S1 and will be discussed below. The eight time bins are computed independently and are color-coded according to the study design shown in Figure 1 with the red colors more proximal to the time of prediction. The bottom graph provides the same amplitudes but has a width proportional to the prevalence among cases. Because the variables shown are binary-encoded (0/1) with normal as reference, the amplitudes can be directly compared as MACE risk, and the areas of the bars in the lower panel can be interpreted as the population-attributable risk.

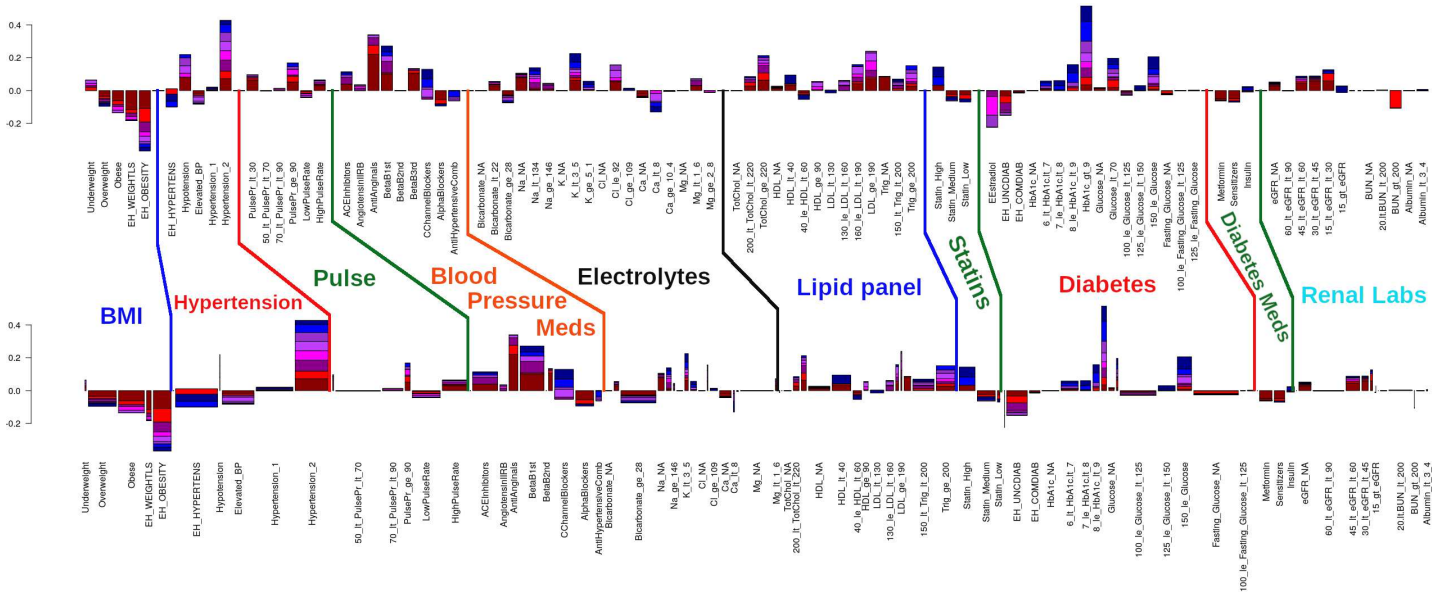


Figure 3. Bar plots of our metabolic syndrome-related GLM model coefficients from the Rcc cohort, with the 8 time-bins stacked for each predictor. The time bins proximal to the event at the base and colored red, while distal predictors (out to 7.5 years) are in bluer colors, corresponding to those shown in the arrow of Figure 1. Upward bars are predictive of a MACE event (positive coefficients), while downward bars indicate protectionb(negative coefficients). The reference categories for each lab and vital signs are not shown, which are taken to be the normal ranges. For most labs, lab_*NA* is included as a category indicating the patient did not have available data. Analogous plots for C17, LDL-DS, and Visit-DS, can be found in Supplemental Figures S3, S4 and S5.

It is important to note that this type of high-dimensional calculations may contain many parallel or co-linear variables which must be considered in the context of all of the other predictors. For example, for BMI, at the left side of Figure 3, both diagnosis codes and vital sign variables are used, and the diagnosis code for obesity having more than double the protective association versus information derived from the vital signs. To interpret this, one must know how the variables are constructed, and how frequently these variables overlap. From Supplementary Table S2, we see that 59% of the patients identified as obese by vital signs have a diagnosis of obesity, while 85% of the patients diagnosed with obesity have a vital sign indicating this. For the case of hypertension, we see that the diagnosis code is both common and slightly protective, while stage 2 hypertension ($>$ 150 mmHg) is more strongly associated with MACE outcome than stage 1 hypertension (between 130 and 140 mmHg). Fifty one percent of individuals with Stage 2 hypertension are on first-generation beta blockers, compared to 42% of all patients (Supplementary Table S2), and being on the medication is positively associated with MACE outcome.

While we have not attempted to disentangle the co-linearity of our predictor variables, we are both indicating how to understand the meaning of the variables and illustrating some of the strategies by which our non-linear models may be adding interaction terms to

gain their improved prediction accuracy. We provide a more comprehensive discussion of the biological drivers of lipid metabolism, hypertension, renal function, diabetes, and obesity and relate our findings to the literature in our Supplementary Discussion section.

Figure S1 shows additional model coefficients used in our calculations, including markers of progression, comorbidities, and additional laboratory results. Most notable are very strong positive associations of diagnoses of coronary artery disease (CAD), chest pains, and nicotine addiction, and a strong and highly prevalent protective effect of taking any medication during a particular time window. The markers of progression account for much of the risk that may have been originally due to the more chronic biomarkers, such as hypertension or lipid metabolism. Similarly, the protective effect of taking 'any medication' will co-occur with all the condition-specific medications, leading to an overall more protective effect than if one considered the medications alone.

Supplementary Figure S6 shows relative risk of stroke vs. AMI on the left and the diagnosis coded outcomes vs. ASCVD deaths obtained from National Death Index (NDI) on the right. For stroke vs. AMI, most variables, including all of the 'NAs' are identical for each component, while CAD, heart failure, anti-anginals, and high proBNP are elevated in AMI and TBI, dementia, and being black or female are enriched in stroke. For Dx-coded vs. non-outcomes, the codes are stratified according to enrichment in ASCVD death, lead by 'NAs' and vital signs, to less likely to be seen by ASCVD death, lead by treatments, CVD-related labs, and medications. Figure S7 shows the same pair of plots but for the C17 cohort. The main difference is how protective all of the 'NA' variables are in the C17 cohort, which is not matched on level of healthcare utilization or age. The preference of variables for stroke or AMI is much less pronounced in C17.

Model coefficients and methodological insights

We illustrate our observation of the model dependence on study design by providing similar bar plots for our other three linear models (C17, Cox-Visit-DS, and Cox-LDL-DS), in Supplemental Figures S3, S4, and S5. Generally, the biggest differences with Figure 3 for metabolic syndrome variables are changes in the level of protection (amplitude) of obesity diagnoses and vital signs. Likewise, the other notable changes in the associations across cohorts are reflected by CVD procedures and *AnyDrug* variable (which reflects encounter with the healthcare facility).

To examine these differences more closely, we plot the Rcc and C17 GLM-LASSO coefficients against each other in Figure 4, with symbol area proportional to the prevalence of the variable. The variables are split into three groups, explained in the inset legend and caption, to facilitate the examination of 243 separate predictors.

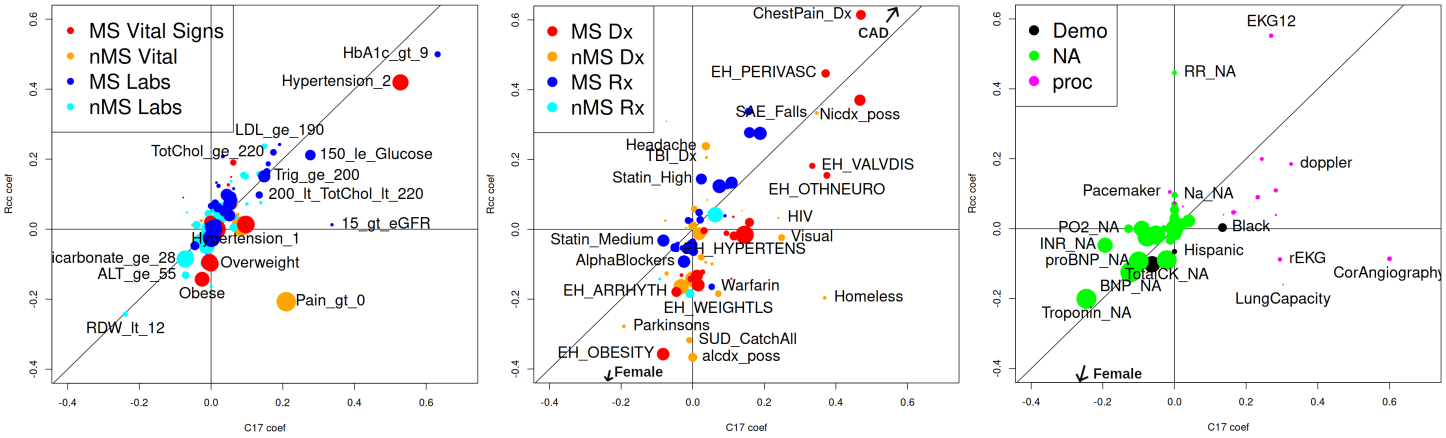


Figure 4. Comparison of GLM model coefficients between Rcc and C17 study cohorts for **(left)** vital signs and labs, color-coded according to whether related to metabolic syndrome or not, and **(middle)** diagnoses and medications, color-coded according to whether related to metabolic syndrome (MS) or not, and **(right)** demographics, procedures, and variables indicating the absence of labs or vital signs (NA). The x-axis indicates the coefficient for model trained with the C17 cohort, the y-axis indicates the coefficient for model trained with the Rcc cohort, the symbol size is proportional to the number of cases with each attribute, and the color code is indicated in the legend.

Three main effects are visible. First is the broad agreement of the biological drivers of CVD, with the diagnoses providing the largest coefficients, but together with medications and laboratory results showing rough agreement between Rcc and C17 study cohorts. Second, we see a number of variables along the x and y-axes, corresponding to variables selected only in the Rcc and C17 cohorts, respectively. Most of those not selected in Rcc are associated with missingness of data (green symbols), consistent with the matching on healthcare utilization in Rcc. Less important predictors such as *TBI*, *Arrhythmia*, and *alcohol addiction* may have been selected only in the Rcc cohort because of the greater statistical power, compared to C17. Finally, we see several variables, such as no recorded respiration rate (*RR_NA*), *homelessness*, not using any medications (*Med_NA*), and reported non-zero pain (*Pain.gt.0*), which are measures of healthcare utilization or data missingness, that are treated very differently between Rcc and C17 cohorts.

Further understanding of healthcare utilization biases can be found by comparing predictors of our sub-outcomes: specifically ASCVD death (which is reported from NDI, independent of the EMRs) to AMI and stroke (which are obtained from the EMR). The association of different classes of predictor variables to be associated with ASCVD death rather than Dx-encoded outcomes is strongest for missingness of laboratory results and weakest for CVD-specific procedures, with other modalities of data intermediate in the spectrum, visible in Supplementary Figure S6. Nevertheless, Figure S9 shows that even the likelihood of ASCVD death increases in proportion to greater

levels of healthcare utilization. Our interpretation of this finding is that the level of healthcare utilization predicts MACE outcomes both because of reporting biases and because patients with CVD have a greater level of healthcare utilization.

We chose to match on age (year of birth) and healthcare utilization (number of visits) because of the strongly confounded nature of both variables. Consequently, we illustrate the dependence of various predictor variables, sub-outcomes, and risk ratios in Figures S8 and S9. Figure S8 shows the incidence of stroke and AMI rising exponentially with age before leveling off near age 60. It also shows the risk ratio of its main risk factors rising strongly with decreasing age, especially diabetes and hypertension. Figure S9 is slightly more complex to interpret and shows the increasing instances of stroke and AMI (or heart attacks) with healthcare utilization - a similar slope as antidepressant prescriptions or alcohol use disorder. It is also evident that the variety of 'NA' variables associated with labs and vital signs can be used to characterize healthcare utilization patterns.

Comparing models through subgroup analysis

Further insight into the latent representation of the risk prediction problem by different models can be gained from calibration curves in important subgroups of patients[39], together with the associated normalized risk histograms of cases in these subgroups. Figure 5a shows these curves for four of our best performing models to identify the logic with which different represent comorbidities (Figures 5a, 5b, 5c, 5d).

Two deficiencies with the AHA-PCE-like model in predicting 5-year MACE events are readily apparent in Figure 5a (left). First, essentially no patients are given a logit risk score above zero, corresponding to an event probability of 50% (C17 cohort has 19% cases). Second, the model underestimates the risk by a factor of two for those patients with CAD diagnosis and for those with values of HbA1C of nine or above. These issues are partially addressed with the Rcc-trained GLM-LASSO model, which has a three unit increase in C-statistic. Half of the patients with a MACE event and an HbA1C ≥ 9.0 or a CAD diagnosis have a risk score above zero. While the calibration has improved, it still systematically underestimates risk at low scores and overestimates it for patients with a probability above 60%. This could be explained by the linear model not being able to accurately combine the multiple risk factors simultaneously present in high-risk patients. *EnsRcc* and the C17-trained neural network model employ two distinct strategies to create non-linear models of MACE risk; both improve on the other two models, with the C17 neural network somewhat better than *EnsRcc*. The CAD diagnosis, in particular, shifts to a higher score for *EnsRcc*, while the C17 neural network model better identifies the risk associated with HbA1c ≥ 9.0.

Comparison of these four models in the demographic subgroups (male, female, Black, and Hispanic) in Figure 5b shows little difference in treatment of these subgroups, beyond the general improvement in accuracy and calibration at higher risks that are evident in the nonlinear models. The lower risk distribution for females is partially explained by the women being fifteen years younger than the overall average (49.4 vs. 64.2 for the controls). Blacks and Hispanics have more amplitude in the risk histogram above a score of one than do males. The logistic regression coefficients show a protective coefficient associated with both women (-0.47), Hispanics (-0.93) and being younger (see Table S1), in the C17 cohort. Much greater differences are visible in Figure 5c, where the risk increases dramatically with age from under fifty to the 60-70 group, with the over eighty group narrowing slightly compared to other groups. The nonlinear models appear better able to identify a tail of patients at the high-risk end of the curve.

The largest differences among both models and subgroups are evident in the healthcare utilization subgroups, shown in Figure 5d. The low (red), overall distribution (black), and high (cyan) usage curves are each quite broad and displaced by a full unit (risk ratio of 2.718) between each group. Both measures of missingness (respiration rate not measured and no medications distributed) are strongly protective. However, the two non-linear models are able to identify significant subgroups at higher risk in both of these categories.

Discussion

Strengths of VA data

Predictive models of MACE outcome in the VHA have more than one million cases, multiple types of healthcare information, and extensive longitudinal data. We have shown that, with this rich data, models have sufficient statistical power to constrain 4 demographic and 243 longitudinal variables with LASSO model selection[41]. However, the observational nature of the data creates problems with confounding effects and selection biases that may be addressable with several strategies, including study design, careful variable definition, and matching and weighting.

Model construction and ensemble transfer learning

Much of medical outcomes research pertains to three distinct problems: identifying biological drivers of disease, screening for conditions, and comparing the effectiveness of proposed interventions. Each of these questions is associated with a type of study design. We concern ourselves with the first two questions here, which has been associated with our retrospective (Rcc) and prospective (C17) study designs, respectively[30]. We use the term "study design" to include outcome variable definition and decisions to use a combined outcome, exclusion criteria, and how to incorporate longitudinal aspects.

Matching and weighting, especially in case-control studies, can be an effective way to increase statistical power and improve treatment of confounding (such as level of healthcare utilization) and strongly interacting (such as age) variables, as described in other studies[31,32]. Our Rcc study, matched on year of birth and balanced on number of patient visits, is designed to identify biological drivers of disease, while our nested prospective cohort (C17) is designed to screen patients and identify associations with new incident disease, as described in Ernster *et al.*[30]. The similarity of the importance of metabolic syndrome-related variables in Figures 3 and S3 and in the left-most panel of Figure 4 shows the equivalency of the two approaches[31] to identifying risk coefficients: modeling and matching.

Most matching approaches are used in medical outcomes research to compare the effectiveness of mitigations or treatments, and so individuals are matched on the propensity score for treatment in order to separate the effect of treatments (typically reducing risk) from

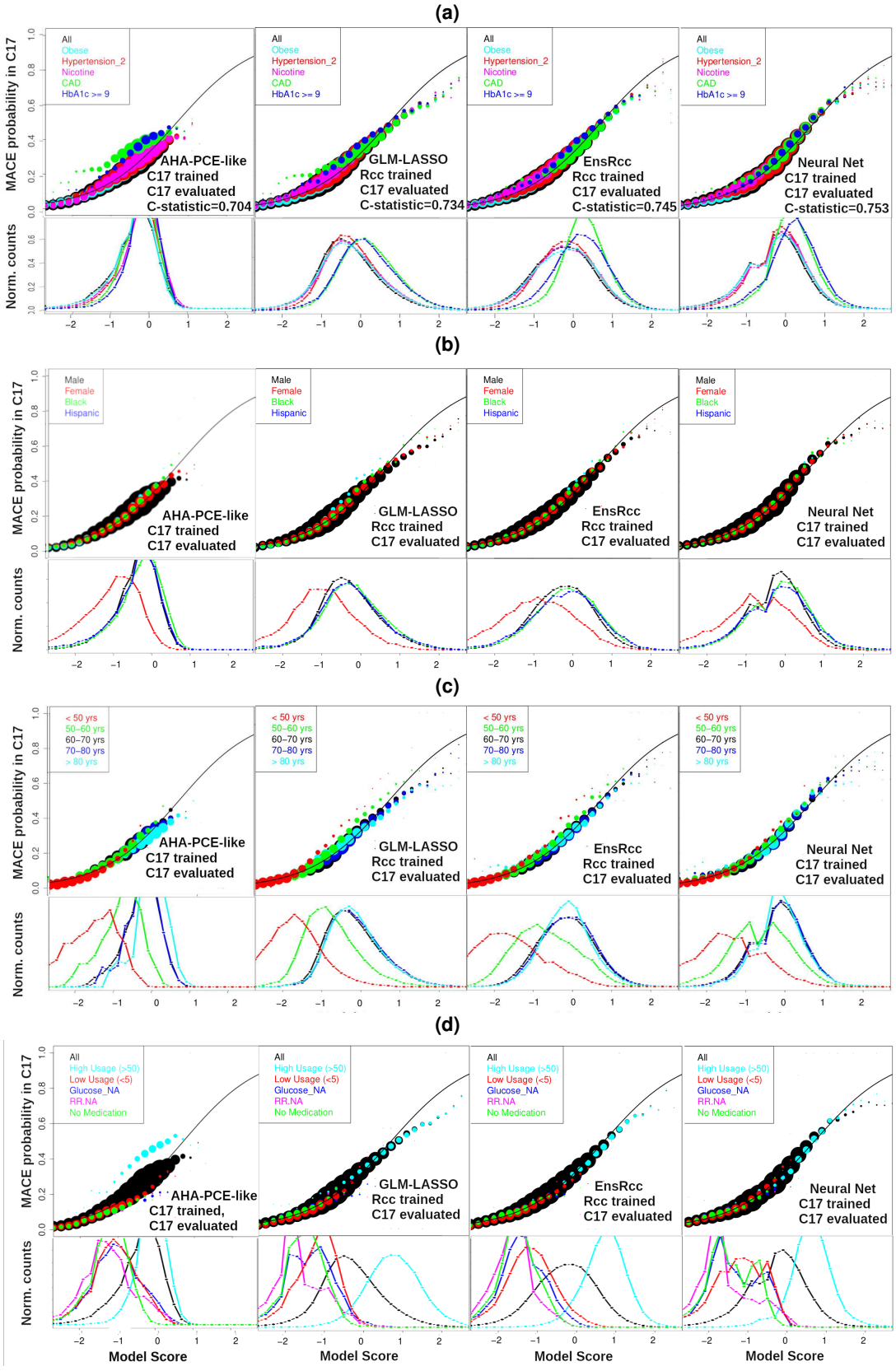


Figure 5. Calibration curves (top) and normalized histograms of scores (bottom) for four different categories of subgroups based on **comorbidities**, **race and ethnicity**, **age**, and **healthcare usage (from top to bottom rows)**. Symbol sizes in calibration curves are scaled to be proportional to the fraction of patients in each histogram with the indicated score. Models shown, from **left to right**, are results from the **AHA-PCE-like model, Rcc-trained logistic regression model, *EnsRcc*, and C17-trained neural network model**. Indicated C-statistics in (a) are computed on the C17 cohort after fine-tuning the Rcc-trained models on C17. Model probabilities, Pr, were transformed to model scores with the logit function, $\log(pr/(1-pr))$, before histograms were made (bottom panels of each row).

the selection bias (typically prescribed to sicker patients). In our case, we are concerned with the problem of separating the confounding effect that diagnoses-based outcomes are less frequent among people who visit the doctor less frequently.

Explicitly using ensemble transfer learning to relate Rcc cohort to C17 cohort provides additional insight into the prediction problem. The observation that Rcc-trained models can be effectively transferred to predict on C17, while the reverse does not result in similarly accurate predictions suggests the Rcc-trained models are, in this respect, more generalizable. But transfer learning (TL) serves multiple purposes. By fine tuning on the C17 cohort with age, race, ethnicity, and sex, we ensure not just that the model is calibrated on these variables, but ensure the overall algorithm is optimized as these groups change. Additionally, as new treatments or ICD coding systems come into use with time, the ensemble transfer learning can increase the weight ascribed to more recent models, with appropriate compensatory adjustments to the fine-tuned coefficients. This addresses part of the 'hidden costs' of AI/ML solutions.[23]

The broad range of predictive models providing similar discrimination and calibration suggests the use of ensemble modeling, that can be combined with transfer learning to provide more robust predictions than individual models, and we have found that examination of how the models change with methodology and study design can be informative, hopefully enabling improved generality and interpretability. One aspect we did not investigate in this work is to separate assessments of treatment benefits from the selection bias of the sicker patients being preferentially treated. This can be done with propensity scoring[31,32] or jointly modeling[42] treatments and associated biomarkers, and we expect the same strategies that improve outcome prediction can be useful in these models as well. Thus we arrive at a 2-stage strategy, where we first train models on an Rcc cohort matched on age and healthcare utilization, then create the best ensemble of nonlinear models to estimate treatment effect.

We note that the algorithm here is the core concept behind transfer learning (TL)[37] and foundational models[11]. The success of our ensemble transfer learning (EnsTL) approach is dependent on our separating biological drivers from those strongly dependent on age interactions and healthcare utilization patterns. In our recent work predicting suicide[43], we demonstrated that ensemble transfer learning could be used to finetune a model from a common, combined, outcome to a rarer and more serious outcome (suicide). We see no reason why this approach could not be used in other methodologies, such as comparative effectiveness analysis[44–46], that require specialized cohort construction.

Interpretation of variables

Choices of predictor variables and how to encode them can impact model performance and interpretability[28], sometimes in complex ways, as exemplified by the obesity paradox[47]. The VA EMRs allowed us to choose a wide range of variables, including demographics, characterization of disease state, including comorbidities, treatments, and a variety of metrics characterizing healthcare utilization patterns. We chose to divide variables into events, which are either present or absent in a particular time bin, and observations, such as weight or laboratory results, which are carried forward from other time bins in the event of missingness. Furthermore, we chose to encode everything except age as categorical variables, with presence/absence for events and a spanning, non-overlapping set of categories for laboratory results and most vital signs with the 'normal' category used as the reference and 'NA' included as a separate category. These encodings have three advantages. First, by having all 0/1 variables, the logistic regression coefficients can be compared without converting units. Second, the results are more easily related to clinical guidance by binning labs and vital signs into clinically meaningful categories (e.g., stage 1 vs. stage 2 hypertension). Finally, the model is less sensitive to outliers, and even logistic regression models will be able to account for 'U-shaped' risk curves without introducing spurious correlations.

A logical division of the variables emerged from this analysis, into biological drivers of CVD, markers of CVD progression, comorbidities, and markers of healthcare utilization patterns and missingness of data. The biological drivers are broadly inclusive of many of the common laboratory tests and well-known CVD risk factors. A detailed comparison with the literature is provided in the Supplementary Discussion section. Additionally, we find a plausible division of predictors that are enriched in AMI vs. stroke (Fig. S6, S7) and a strong age dependence of the relative risk of MACE for four markers of commonly associated risk factors (Fig. S8). Although providing broadly similar results to the literature, our calculation is self-consistently treating the full breadth of biological drivers, nested into well-defined cohorts of patients. This eliminates the need to weigh and understand distinct patient populations, exclusion criteria, and EMR reporting processes in reconciling results.

Model characterization

Interpretability is a well-known difficulty with AI/ML techniques[21]. In principle, different prediction methodologies are best-suited to distinct techniques, such as saliency maps[48], SHAP[49,50] for neural networks, combinatorial predictors for RF like iterative random forests[51], and parameters characterizing the attention scheme for TabNet[52]. One can structure deep learning architectures to capture the logic of longitudinal treatments, diagnostics, hospitalizations, and demographics in the hopes that the neural network captures each aspect of the logic in a distinct network location.[19].

We chose instead to characterize our models through comparison of linear coefficients, and calibration curves and risk histograms in important subgroups. This process has enabled us to understand the distinct role of parallel predictors and the need to separate confounding variables related to healthcare utilization patterns from those describing treatments, comorbidities, and disease states. For each of the models selected for subgroup analysis, we see progressive improvements from AHA-PCE-like baseline model to a prospectively trained fully connected neural network (NN).

Clinical implications

Much of clinically relevant screening and understanding of epidemiological risk factors derives from studies similar to the AHA-PCE model[5] that we used as a baseline, or the updated, PREVENT, study[53], which was not available when we performed our analysis. Both

the broad array of biologically relevant risk factors seen in Figure 3 and the five percentage point improvement in discrimination seen in Figure 2 suggest large-scale analyses such as ours may offer considerable improvements. Our subgroup analysis has shown that a single high dimensional model incorporating non-linearities and interactions among variables can predict accurately across race, ethnicity, and gender, across important risk factors, and across age groups and groups defined by healthcare utilization patterns.

Unfortunately, we have traded one form of complexity for another. The AHA-PCEs were defined with a very specific study design with strict inclusion/exclusion criteria, so that, for example, the decision to include new biomarkers validated on a different cohort of patients, can be a difficult process[5]. In particular, we have observed **much** greater dependence of MACE likelihood on age and healthcare utilization, which will vary from population to population, than on the biological drivers of disease or the subgroups defined by sex, race, and ethnicity. While forcing every study to use a well-defined relationship to a broadly defined prospective screening cohort (PC)[30] would ease the difficulty of comparison across studies, thoroughly exploring large-scale, observational studies such as this may ultimately prove more tractable. This is the reason we have provided extensive characterization of our cohort and model, with both univariate analyses in subgroups, dependence of the coefficients on diverse study designs, and even providing the co-occurrence of important predictor variables in both our Rcc and C17 cohorts. While tedious, it is at least possible to sort out how the diverse drivers of MACE outcomes interact with one another.

From a clinical perspective, our results suggest that robust cardiovascular risk prediction should not rely on a single study design. Matched nested retrospective case control (Rcc) are particularly effective for identifying stable biological drivers of disease, whereas prospective screening cohorts (PCs) better capture the characteristics of patients encountered in routine clinical practice. By integrating these complementary sources of information and modeling approaches, our ensemble transfer learning (EnsTL) framework provides a practical pathway for developing prediction models that are both biologically informed and well calibrated for prospective patient screening. Such models can be continuously and relatively easily updated as healthcare practices, coding systems, and patient populations evolve without requiring complete redevelopment of the underlying predictive framework.

More broadly, this framework provides a general strategy for translating large-scale observational EMR data into clinically deployable decision-support tools. Rather than constructing separate prediction models for every healthcare system or clinical application, robust mechanistic models can be learned from large, carefully designed cohorts and efficiently adapted to new institutions, populations, or outcomes through lightweight fine-tuning. Although demonstrated here for MACE prediction, the same principles are readily applicable to other chronic diseases and longitudinal clinical prediction problems, providing a scalable pathway toward more reliable and generalizable AI-enabled precision medicine. It will, of course, require a great deal of experience defining appropriate variables and processes for this to create models for robust practice, but this path appears to us to be tractable.

Limitations

The primary limitation of this study is the missingness of information, due to the observational nature of EMR-based outcomes analysis. In particular, we were unable to include Centers for Medicare and Medicaid Services (CMS) data in our analysis, leading to a decreased knowledge of healthcare for veterans over 65 years of age. However, Deardorff *et al.* report minimal impact on mortality prediction when adding CMS claims data to VHA EMRs[54], which suggests that our results can serve as a firm baseline for predicting CVD.

Our cohort was composed of US Veterans who are adults, 90% male, and have a median age of 65 years, (see Supplementary Table S1), and so the pathways towards MACE outcomes may not generalize to other populations. Our study has utilized only representative ML predictive models, with a focus on simpler and more generally understood methods, to more clearly elucidate the interplay between study design, ensemble modeling and transfer learning, and model performance. It appears likely that more elaborate base models could further improve model prediction accuracy.

Conclusions

We demonstrated that the rich and longitudinal VHA EMRs contain sufficient information to develop accurate and interpretable prediction models for major adverse cardiovascular events (MACE) while also providing meaningful insight into how epidemiological study design influences the information learned by the machine learning (ML) models. Although nested retrospective case control (Rcc) and prospective screening cohorts (PCs) identified largely consistent biological drivers of cardiovascular disease (CVD), they differed substantially in their representation of healthcare utilization patterns and observational biases. A similar set of biological drivers, covering the full breadth of metabolic syndrome variables was identified by either matching on age and health care utilization (in Rcc) or modeling the effects (in C17), albeit with more detail in the Rcc study design. Non-linear models were observed to improve the treatment of age, missingness of information, and combining chronic risk factors with markers of progression (such as a coronary artery disease and chest pains). These complementary strengths can be effectively combined through ensemble transfer learning (EnsTL).

Our results show that models trained on carefully matched Rcc cohorts capture stable biological mechanisms that can be successfully transferred to prospective screening populations through lightweight finetuning, producing well-calibrated predictions while preserving clinically meaningful associations. This approach enables prediction models that are more robust to differences in cohort construction and healthcare delivery than models developed from a single study design alone. These findings suggest that explicitly accounting for study designs can improve the robustness, interpretability, and transferability/generalizabilituy of EMR-based prediction models and provide a practical path towards developing reliable clinical prediction models from large observational healthcare databases.

Methods

MACE outcomes and study designs

Our outcome (and its time) was defined as the first recorded instance (and that time) of any one of three types of MACE (Major Adverse Cardiac Events): acute myocardial infarction (AMI), ischemic stroke, and atherosclerotic cardiovascular disease (ASCVD) death. We used International Classification of Diseases (ICD)-9 and 10 codes for AMI and stroke from inpatient and outpatient records and hospital service fees. For ASCVD death, we linked our patient data to the underlying cause of death in the NDI database. For AMI, both 410.x (acute myocardial infarction) and 412.x (old myocardial infarction) were used for ICD-9 codes, while I21.x (acute myocardial infarction), and I22.x, and I25.2 (old myocardial infractions) were used for ICD-10. For stroke, 433.[012389]1 (occlusion and stenosis of cerebral arteries, with infarction, where square brackets indicate any of the enclosed digits can be matched) and 434.91 (occlusion of cerebral arteries with unspecified infarction) were used for ICD-9 codes, while I63.x (cerebral infarction) were used for ICD-10. In addition to the AMI and stroke conditions, the cause of death file was searched for any I25 (chronic ischemic heart disease), Z95.1 (presence of aortocoronary bypass graft), and Z98.61 (presence of coronary angioplasty implant and graft) codes.

We define 'time of prediction' as the date when the prediction of a given outcome, which occurs later, is made. For all study designs (see Figure 1), any record of the above-defined outcomes before the time of prediction resulted in excluding the patient as either a case or a control. We analyzed four cohorts of patients, each with its own additional criteria. For the Rcc cohort, the 'time of prediction' was required to be after July 1, 2007. For the LDL - decision support (LDL-DS) cohort, at least one cholesterol lab reported in the 3 yrs before July 1, 2010, was required, and the time of prediction was one week after the first such result. Similarly, for the Visit - decision support (Visit-DS) cohort, at least one outpatient visit was required in the 3 yrs before July 1, 2010, and the prediction time was one week after the first such visit. Finally, for the C17 cohort, at least one inpatient or outpatient visit was required in the three months before the 'time of prediction', January 1, 2017, for all patients in this cohort.

Additionally, for all designs: 1) we required a date of birth post January 1, 1901, and registration with the VHA by Jan 1, 2022, when the data tables were created, and 2) we included a 4-6 week gap between the 'time of prediction' and the time of the first 'allowed' outcome, to minimize predictor variables from 'leaking' information from the very same visit defining the outcome. A very small number (dozens) of patients were excluded based on failed self-consistency checks, such as recorded dates of death in the future, perhaps resulting from inconsistent information when patients registered at multiple VHA Medical Centers.

Our Rcc design used as cases all first MACE events occurring between Aug 15th, 2007 and Jan 1st, 2022, and prediction was made with data spanning 7.5 years before the event. Controls were selected by randomly choosing patients' visits among non-cases during the same time period as the matching outcome event and from the same date of birth cohort. This study design allows the greatest amount of information (both chronic and acute) to be used, utilizes all available cases, removes age as a confounding variable, and balances overall healthcare utilization between cases and controls, potentially providing more insight into the physiology and mechanism of disease. In this study cohort, we have 1 million cases and 1 million controls (Figure 1). For matching in the Rcc cohort, two criteria were used. First, controls were chosen from within one of 100 equally sized and time-ordered date-of-birth cohorts of roughly 230,000 patients each. Second, to balance the average healthcare utilization of controls and cases, controls were chosen randomly among visits of patients who were not excluded or cases. This approach increases the likelihood of a patient being chosen as a control in proportion to the number of visits.

Our second study design, LDL-DS, was designed as a time-to-event model with an opportunity for at least 12.5 years of follow-up, and the time of prediction was set to one week after LDL cholesterol labs were obtained. The LDL labs were required to be obtained during the three years from July 1, 2007 to July 1, 2010. No matching of cases to controls occurred, so there were approximately 40% as many cases but nearly twice the controls as in Rcc. Our third study design, Visit-DS, was similar to LDL-DS but required only a visit rather than an LDL lab. Consequently, there were nearly twice as many cases and controls as in LDL-DS.

Finally, we created a calendar study design, C17, that predicted MACE events on January 1, 2017, for the five-year interval after February 8, 2017, for patients with a visit during the three months before the time of prediction. This design has the lowest fraction of cases (5.7%) but reflects a recent sampling of patients utilizing the VA system, and we analyzed it, together with Rcc, in the most detail. To simplify the calculation and improve the machine learning model performance, we down-sampled the controls by a factor of four, computed randomly within each date-of-birth cohort. This kept the enrichment of cases in our actual predictive model calculations within a factor of two of each other (from 24% to 50%), which is helpful when comparing results between models.

Variable definitions

Our predictor variables include age, gender, race, ethnicity, and 243 time-dependent binary variables, which include categories of vital signs, lab results, diagnosis, procedure codes, and pharmacy fill data for the study individuals. For age, gender, race, and ethnicity, many patients were queried multiple times. In this case, the most frequent (mode) of the answers was retained. Gender was encoded as "male" (reference) and "not male". Race was encoded as Black or African American or "not" (reference), and ethnicity was encoded as Hispanic (reference) or "not".

Our time-dependent predictor variables came from several data domains, including demographics, ICD-9 and ICD-10 diagnosis codes, medication generic names, procedure codes (CPT), vital signs, hospital resources utilized (stop codes), and laboratory results. The time dependence was organized backward from the time of prediction in adjacent periods two each of 3 months, 6 months, 1 year, and 2 years, for a total of 7.5 years of data, as illustrated in Figure 1. This balanced the need for parsimony against the desire for acute information while retaining 7.5 years of data. We refer to these periods as the time windows of data collection.

Medications, procedures, and diagnoses

The drugs and drug categories for antidepressants, stimulants, opioid-for-pain, estradiol, levothyroxine, ezetimibe, bile acid sequestrants (BAS), aspirin, warfarin, and metformin were defined by searching the "DrugNameWithoutDose" field for a list of words or word stems, and manually checking the output for relevance. The statins were mapped into low, medium, and high intensity based on the drug name and dosage following Grundy *et al.*[55]. Cardiovascular drugs and diabetes drugs were identified by using the "VADrugClassification" scheme, such as HS502 for oral hypoglycemic agents, HS501 for insulin, and most of the categories of CV drugs: CV050 for digitalis glycosides, CV150 for alpha-blockers, CV200 for calcium channel blockers, etc. The VA drug classification scheme is available online at[56]. We also created the category "Any Drug" for any of the products distributed by the pharmacy, including non-drug products such as lancets. A spreadsheet with the most common medications comprising each group is provided in the Supplementary Information (SI). Medications were recorded as present in each time window of data collection if one or more instances were recorded and absent if none of the medications included for each variable was present.

Procedure codes were selected from among the most prevalent inpatient and outpatient CPT codes identified as pertinent to CVD, grouped manually by inspection, and included EKG12 (9300, 9305, 93010, and 93225), rEKG (93040, 93041, and 93272), doppler (93307, 93320, and 93325), dialysis (90935, 90937, 90999, 90989, 36145, and 36489), stress test (9301[5678] and 93350), artery scan (93875, 93880, and 9392[2346]), veinous study (93965, and 9397[0610]), ESRD service (9092[15]), left heart catheter (93510), coronary angiography (93539, 9354[035], and 9355[56]), pacemaker (9373[123]), cardiac rehabilitation (9379[78]), lung capacity test (93720), blood pressure monitoring (93784 and 93790), stent (92980), and defibrilator (93741), where square brackets indicate any of the enclosed digits can be matched. Procedure codes were recorded as present in each data collection time window if one or more instances were recorded and absent if none of the codes included for each variable was noted.

Variables defined from ICD-9 and ICD-10 codes were adapted from a list provided by the VA Office of Mental Health and Suicide Prevention, following from definitions used in Ref.[57]. Diagnosis codes were collected from inpatient, outpatient, and fee-for-service tables. ICD-9 and ICD-10 codes were lumped together, with letters for ICD-9 codes (used before October 1, 2015) changed to lowercase and all search terms required to match from the first character in the field. For example, using this process, the variable alcdx_poss characterizes alcohol use disorder and matches any ICD-9 codes starting with 291, 303, and 305.0 and ICD-10 codes starting with F10. Variables sometimes overlap in the codes they use and were included with the idea that Lasso model selection would identify the best overall for the MACE prediction problem. An important aspect is that variable definitions should ideally be consistent over time, including the ICD-9 to ICD-10 changeover in 2015, and also through the yearly revisions of the coding manuals. To this end, we plotted histograms (see the SI) of Dx-based variables together with the number of occurrences of the most prevalent component variables for a 3% sampling of representative aged patients. This allowed verification that there was consistent reporting during the study interval, and we attempted to harmonize the definitions as much as possible. Still, a certain degree of confounding is likely to occur. Diagnosis-based variables were recorded as present in each data collection time window if one or more instances were recorded and absent if none of the codes included for each variable was noted.

Vital signs and laboratory data

Vital signs were taken from both inpatient and outpatient visits and included systolic and diastolic blood pressure, weight, pulse, respiration rate, oxygen saturation, and self-reported pain (scale from 1 to 10). Laboratory results were collected from the "LabChemTestResultNumeric" field. For each of these variables, the median reported value was computed for each patient during each of the eight time windows of data collection. For all vital signs except pain and all laboratory results, values were imputed for time windows without recorded values by first carrying values forward in time to replace NAs, then backward in time to replace any remaining NAs. The reasoning is that recorded values are observations of a continuous process; as such, the most recent observation should be used in the case of missing information. By contrast, reported pain, diagnoses, procedures, and medications are considered events, and the presumption is that when they are not noted, it is because they did not occur. Weight was transformed into BMI using the median height for each patient, computed from all recorded measurements of height for each patient, and any values below 10 or above 100 were discarded.

Laboratory results and vital signs were categorized according to standard clinical nomenclature and referenced to the 'normal' category. Note that lack of measurement is included as an 'NA' category, which is potentially informative. Specifically, BMI (kg/m^2) was classified as (i) underweight if < 18.5, (ii) overweight if > 25 and < 30, and (iii) obese if > 30, with the reference state, 'normal', being between 18.5 and 25. Similarly, for systolic blood pressure ($mmHg$), (i) hypotension is < 90, (ii) elevated blood pressure is between 130 and 140, (iii) hypertension 1 is between 130 and 140, and (iv) hypertension 2 is greater than 140; with the reference state, 'normal', between 90 and 130. A similar process was used for the other vital signs. Pulse pressure was computed by subtracting the reported median diastolic from the reported median systolic for a given time period. Because of the way results were carried forward and backward across time bins, patients in the NA category will have NA recorded for all eight time windows of data collection.

Laboratory results in the VA medical records are assigned a unique identifier ("LabChemTestSID"), along with a location ("sta3n") and sample type ("topology") codes. On the basis of these fields, we have established lists of identifiers associated with 100 of the most common laboratory result types, and we selected a subset of these labs for use as MACE predictor variables. Laboratory results were processed in a similar manner to vital signs, except that all labs had an NA category included, so a patient not identified as NA, high, or low is a patient whose median result for that time window is within the normal range for that lab result. Second, histograms of median results across patients were inspected for each laboratory result type, and cutoff values were identified as likely separating values using other than those expected; values outside the 'reasonable' range were assigned to the NA category, as with missing results. For the case of creatinine, values >10 were divided by 88.2, to convert from μmol/L to g/dL before out-of-range values were identified, and the standard gender-dependent and race-dependent formula was used to convert to eGFR. If both eGFR and creatinine were reported in a given time

interval, the eGFR value was used. For hemoglobin levels and red blood cell counts, distinct mappings between values and categories were used for men and women, with 11-12 g/dL assigned 'low' for men, 10-11 g/dL assigned 'low' for women for hemoglobin level, and 'normal' range for RBC was between 4.5 and 5.5 for men and 3.8 and 4.8 for women. For differential WBC, no attempt was made to process absolute measures of WBC categories; if values were not in the range expected for percentage, they were assigned to the NA category.

Baseline AHA-PCE-like Model

We recreated the American Heart Association - Pooled Cohort Equations (AHA-PCE)[5] prediction model by matching each of the predictor variables in that model, as best as possible, with the variables in our data set containing the 247 predictor variables. Demographic information was taken as is from our data. Information on the cholesterol level was extracted from the HDL cholesterol, LDL cholesterol, and total cholesterol variables and encoded as present if it was present in any of the eight time bins for the three variables. For smoking, the Nicdx variable, which includes the diagnosis codes for nicotine use (ICD9 305.1, and ICD10 F12.2 and Z72.0) was used. Blood pressure information was extracted from the hypotension, elevated blood pressure, hypertension_1, and hypertension_2 variables. To encode whether or not a patient is prescribed blood pressure medications, a combination of ace-inhibitors, angiotensin receptor blockers, anti-anginal, alpha-blockers, beta-blockers, calcium channel blockers, loop diuretics, thiazide diuretics, and potassium-sparing diuretics variables was used and encoded as a yes if the patient was on any of these medications in any of the eight time bins. To encode whether a patient had diabetes, a combination of diagnosis of diabetes, labs with fasting glucose greater than 126 mg/dl or casual glucose >200 mg/dl or hemoglobin A1C (HbA1C) greater than 7%, and prescription of insulin or other diabetes drugs such as sensitizers, metformin, or thiazolidinediones was used.

Prediction methodologies

Six different modeling algorithms were used to predict the occurrence of a MACE event over the time window of prediction, given the 7.5 years of medical history of the patients before the time of prediction. The algorithms used are i) a American Heart Association - Pooled Cohort Equations (AHA-PCE)[5] like equation as a baseline, ii) logistic regression with LASSO model selection (GLM-LASSO)[24,41,58], iii) random forest (RF)[59], iv) fully connected neural network (NN)[60–62] with 3 hidden layers, v) a transformer based neural network (TabNet)[52], and vi) a Cox model[63].

For all methods except Cox, the problem was formulated as a binary outcome after the prediction baseline date. The logistic regression model was tuned with the Lasso penalty to enable model selection of the predictor variables. For random forest, we used 500 trees, the square root of the number of input features for the number of features to consider for the best split, and require a minimum of 100 samples at a leaf node. The fully connected neural networks had 3 hidden layers with 1024, 512, and 256 neurons respectively and ReLu non-linearity[64]. For TabNet, the Pytorch implementation was used[65] with default settings for the hyperparameters, including 8 for the width of the decision prediction layer, 8 for the width of attention embedding, 3 for number of steps in the architecture, and 1.3 for gamma, which is the coefficient for feature reusage in the masks.

For Cox models, the study was formulated as a time-to-event problem. The LDL-DS and Visit-DS cohorts were analyzed with a Cox regression time-to-event approach. To facilitate comparison with the predictions of the other algorithms, after fitting the Cox models, they were evaluated as interval predictions at the indicated times-to-event. The LDL-DS and Visit-DS cohorts were also analyzed with an AHA-PCE-like model by formulating them as an interval prediction problem.

We calculated C-statistics (or area under the receiver operating characteristic - ROC - curve) to compare the different models and study cohorts. As part of this comparison, we evaluated the c-statistic both with the trained model used directly on each study cohort and with a version of the trained model finetuned by computing a GLM-LASSO model on a subset of variables, including the model score described, together with Gender, Race, Ethnicity, Age, Age^2, and variables representative of healthcare utilization (Pain > 0, AnyDrug, no measured respiration rate, and no glucose measurement).

For all study designs and methodologies, we use half of the cases and controls to train the model and the other half to evaluate it.

Author Contributions

B.H.M, S.D., Sc.D., and C.J.O conceived the study; S.D., R.M.R., J.D.C., K.C., L.C., Y.H., and B.H.M. implemented data wrangling; S.D. and B.H.M. did most and model development and subsequent analyses. S.D., R.M.R, T.B., N.W.H, A.C.J., J.Z., K.A., and B.H.M. performed extensive model validation. B.K., K.G., A.E.M, G.H.M., A.C.J, C.J.O., and Sc.D served as the subject matter experts (SMEs) and contributed in relating our study to existing studies. A.C.J., P.W.W., J.L.V, D.C.P, J.M.G., Y.V.S, and C.J.O. provided clinical interpretation of the modeling results. S.D., R.M.R, K.G., and B.H.M. mapped the modeling results to the clinical concepts; S.D., R.M.R, and B.H.M. wrote the manuscript; and all authors reviewed the manuscript.

Funding Declaration

This work was funded by the Million Veteran Program (MVP), Office of Research and Development (ORD), US Department of Veterans Healthcare Administration (VHA) under award #MVP014.

Acknowledgements

This research was conducted with the approval of the VA Central IRB with project number MVP014. All methods were carried out in accordance with this approval. The research is based on data from the US Department of Veterans Healthcare Administration (VHA) and

was supported by the MVP-CHAMPION collaboration between the United States Department of Energy (DOE) and the US Department of Veterans Affairs (VA). This publication does not represent the views of the VA, DOE, or the United States Government.

We thank the many people involved in setting up the MVP-Champion infrastructure at Oak Ridge National Laboratory, including Sumitra Muralidahar, Dimitri Kusnezov, Edmon Begoli, and Franciel Linarus and team. We would also like to thank Jodie Trafton and her team at the VA Office of Mental Health and Suicide Prevention (OMHSP) for providing ICD-9 and ICD-10 codes for their 100 Dx-related variables definitions and much other pragmatic modeling advice. We thank numerous VA and DOE colleagues who developed these ideas over the past several years and for carefully reading many previous iterations of this manuscript.

Notice: This manuscript has been authored by Triad National Security, LLC under Contract No. 89233218CNA000001 with the U.S. Department of Energy/National Nuclear Security Administration. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript or allow others to do so, for United States Government purposes. The Government has also granted for itself and others acting on its behalf a nonexclusive, paid-up, irrevocable worldwide license in the code and data, within this manuscript, to reproduce, prepare derivative works, and perform publicly and display publicly, by or on behalf of the Government. NEITHER THE GOVERNMENT NOR THE CONTRACTOR MAKES ANY WARRANTY, EXPRESS OR IMPLIED, OR ASSUMES ANY LIABILITY FOR THE USE OF CODE WITHIN THIS MANUSCRIPT.

Competing interests

Christopher O'Donnell is an employee of Novartis Pharmaceutical Company.

Data availability

The data that support the findings of this study are available from US Department of Veterans Healthcare Administration (VHA), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the corresponding author upon reasonable request and with permission of the VHA.

Code availability

The overall framework code currently resides in a secure enclave that hosts personal healthcare information (PHI). The relevant portions of the code will be made available by the corresponding author upon reasonable request.

References

1. Ahmad, F. B. & Anderson, R. N. The leading causes of death in the US for 2020. *JAMA* **325**, 1829–3021 (2021).
2. Roth, G. A., Mensah, G. A. & Others. Global burden of cardiovascular diseases and risk factors, 1990-2019: Update from the GBD 2019 study. *J. Am. Coll. Cardiol.* **76**, 2982–3021 (2020).
3. Wilson, P. W. F. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998).
4. D'Agostino, R. B., Lee, M.-L. & Belanger, A. J. Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study. *Stat. Medicine* **9**, 1501–1515 (1990).
5. Goff, D. C. *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk. *Circulation* **129**, S49–S73, DOI: 10.1161/01.cir.0000437741.48606.98 (2014). https://www.ahajournals.org/doi/pdf/10.1161/01.cir.0000437741.48606.98.
6. Flora, G. D. & Nayak, M. K. A brief review of cardiovascular diseases, associated risk factors and current treatment regimes. *Curr. Pharm. Des.* **25**, 4063–4084 (2019).
7. Eckel, R. H., Grundy, S. M. & Zimmet, P. Z. The metabolic syndrome. *The Lancet* **365**, 1415–1428 (2005).
8. Kasper, D. L., Fauci, A. S., Hauser, S. L. & Longo, D. L. *Harrison's Principles of Internal Medicine, 20th Ed.* (McGraw-Hill, USA, 2018).
9. Singh, A. *et al.* Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *J. Biomed. Inform.* **53**, 220–228 (2015).
10. Zhao, J. *et al.* Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci. Rep.* **9**, 717 (2019).
11. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *ArXiv* (2021).
12. Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H. & Luscombe, N. M. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One* **13**, e0202344 (2018).
13. Vassy, J. L. *et al.* Estimation of atherosclerotic cardiovascular disease risk among patients in the veterans affairs health care system. *JAMA Netw Open* **3**, e16–23, DOI: 10.1001/jamanetworkopen.2020.8236 (2020).
14. Usman, M. *et al.* Effect of sglt2 inhibitors on cardiovascular outcomes across various patient populations. *J. Am. Coll. Cardiol.* **81**, 2377–2387 (2023).
15. Xie, Y., Choi, T. & Al-Aly, Z. Mapping the effectiveness and risks of glp-1 receptor agonists. *Nat. Med.* **31**, 951–962 (2025).
16. Vassy, J. *et al.* Cardiovascular disease risk assessment using traditional risk factors and polygenic risk scores in the million veteran program. *JAMA Cardiol.* **8**, 564–574 (2023).
17. Hippisley-Cox, J. *et al.* Development and validation of a new algorithm for improved cardiovascular risk prediction. *Nat. Med.* **30**, 1440–1447 (2024).
18. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record EHR analysis. *IEEE J. Biomed Heal. Inform.* **22**, 1589–1604 (2018).
19. Pham, T., Tran, T., Phung, D. & Venkatesh, S. Predicting healthcare trajectories from medical records: A deep learning approach. *J. Biomed. Inform.* **69**, 218–229 (2017).
20. McDermott, M. *et al.* Reproducibility in machine learning for health research: Still a ways to go. *Sci. Transl. Med* **24**, 586 (2021).
21. Marcus, G. Deep learning: A critical appraisal. *ArXiv* **abs/1801.00631** (2018).
22. Dhaubhadel, S. *et al.* Global explainability of a deep abstaining classifier for cancer pathology reports. *IEEE J. Biomed. Heal. Informatics* 1–15, DOI: 10.1109/JBHI.2026.3703343 (2026).
23. Sculley, D. *et al.* Hidden technical debt in machine learning systems. *NIPS Proc.* **2**, 2503–2511 (2015).
24. Efron, B. Logistic regression, survival analysis and the Kaplan-Meier curves. *J. Am. Stat. Assoc.* **83**, 414–425 (1988).
25. Ngwa, J. S. *et al.* A comparison of time dependent Cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the Framingham Heart Study. *BMC Med. Res. Methodol.* **16**, 148–160 (2016).
26. Vatcheva, K., Lee, M., McCormic, J. & Rahbar, M. The effect of ignoring statistical interactions in regression analyses conducted in epidemiologic studies: An example with survival analysis using cox proportional hazards regression model. *Epidemiology* **6** (2015).
27. Concato, J., Shah, N. & Horwitz, R. I. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New Engl. J. Medicine* **342**, 1887–1892, DOI: 10.1056/NEJM200006223422507 (2000). PMID: 10861325, https://doi.org/10.1056/NEJM200006223422507.
28. Concato, J. *et al.* Observational methods in comparative effectiveness research. *Am. J. Medicine* **123**, e16–23 (2010).
29. Cordoba, G., Schwartz, L., Woloshin, S., Bae, H. & Gøtzsche, P. Definition, reporting, and interpretation of composite outcomes in clinical trials: Systematic review. *BMC* **341**, C3920 (2010).

30. Ernster, V. L. Nested case-control studies. *Prev. Med.* **23**, 587–590 (1994).

31. Ho, D., Imai, K., King, G. & Stuart, E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Analysis* **15**, 199–236 (2007).

32. Stuart, E. Matching methods for causal inference: A review and a look forward. *Stat Sci.* **25**, 1–21 (2007).

33. Nguyen, D.-K., Lan, C.-H. & Chan, C.-L. Deep ensemble learning approaches in healthcare to enhance the prediction and diagnosing performance: The workflows, deployments, and surveys on the statistical, image-based, and sequential datasets. *Int. J. Environ. Res. Public Heal.* **18**, 1–19 (2021).

34. Ju, C., Bibaut, A. & van der Laan, M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J. Appl. Stat.* **45**, 2800–2818 (2018).

35. Liu, X., Liu, Z., Wang, G., Cai, Z. & Zhang, H. Ensemble transfer learning algorithm. *IEEE Access* **6**, 2389–2396 (2018).

36. Bozinovski, S. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica* **44**, 291–302, DOI: https://doi.org/10.31449/inf.v44i3.2828 (2020).

37. Zhuang, F. *et al.* A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76, DOI: 10.1109/JPROC.2020.3004555 (2021).

38. Honerlaw, J. *et al.* Centralized Interactive Phenomics Resource: an integrated online phenomics knowledgebase for health data users. *J. Am. Med. Informatics Assoc.* **31**, 1126–1134, DOI: 10.1093/jamia/ocae042 (2024). https://academic.oup.com/jamia/article-pdf/31/5/1126/57286246/ocae042.pdf.

39. Cook, N. R. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **117**, 928–935 (2007).

40. Jager, K., Zoccali, C., MacLead, A. & Dekker, F. Confounding: What it is and how to deal with it. *Kidney Int.* **73**, 256–260 (2008).

41. Tibshirani, R. Regression shrinkage and selection via the LASSO. *J. Royal Stat. Soc. Ser. B (Methodological)* **58**, 267–288, DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x (1996). https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x.

42. Rizopoulos, D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67**, DOI: 10.1111/j.1541-0420.2010.01546.x (2011).

43. Dhaubhadel, S., Ganguly, K., Ribeiro, R. & et al. High dimensional predictions of suicide risk in 4.2 million US Veterans using ensemble transfer learning. *Sci. Rep.* **14**, 1793, DOI: https://doi.org/10.1038/s41598-024-51762-9 (2024).

44. Ciminata, G. *et al.* Propensity score methods for comparative-effectiveness analysis: A case study of direct oral anticoagulants in the atrial fibrillation population. *PLoS One* **17**, e0262293 (2022).

45. Ford, J. *et al.* Controls, comparator arms, and designs for critical care comparative effectiveness research: It's complicated. *Clin. Trials* **21**, 124–135 (2024).

46. E.W., G. *et al.* Use of real-world data in population science to improve the prevention and care of diabetes-related outcomes. *Diabetes Care* **46**, 1316–1326 (2023).

47. Sperrin, M., Candish, J., Badrick, E., Renehan, A. & Buchan, I. Collider bias is only a partial explanation for the obesity paradox. *Epidemiology* **27**, 525–530 (2016).

48. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations* (2014).

49. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4768–4777 (Curran Associates Inc., Red Hook, NY, USA, 2017).

50. Martinez, C. *et al.* Deep sequential neural network models improve stratification of suicide attempt risk among US Veterans. *J. Am. Med. Assoc.* **31**, 220–230 (2024).

51. Basu, S., Kumbier, K., Brown, J. B. & Yu, B. Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci.* **115**, 1943–1948, DOI: 10.1073/pnas.1711236115 (2018). https://www.pnas.org/doi/pdf/10.1073/pnas.1711236115.

52. Arik, S. O. & Pfister, T. TabNet: Attentive interpretable tabular learning. *Proc. AAAI Conf. on Artif. Intell.* **35**, 6679–6687 (2021).

53. Khan, S. S. *et al.* Development and validation of the american heart association's prevent equations. *Circulation* **149**, 430–449, DOI: 10.1161/CIRCULATIONAHA.123.067626 (2024). https://www.ahajournals.org/doi/pdf/10.1161/CIRCULATIONAHA.123.067626.

54. Deardorff, W. J. *et al.* Do functional status and medicare claims data improve the predictive accuracy of an electronic health record mortality index? Findings from a national Veterans Affairs cohort. *BMC Geriatr.* **22**, 1–9 (2022).

55. Grundy, S. N. & *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **139**, e1082–e1143 (2019).

56. VA Drug Classification (1998).

57. McCarthy, J. F. *et al.* Predictive modeling and concentration of the risk of suicide: Implications for preventive interventions in the us department of veterans affairs. *Am. J. Public Heal.* **105**, 1935–1943 (2015).

58. Cox, D. R. The regression analysis of binary sequences. *J. Royal Stat. Soc. Ser. B (Methodological)* **20**, 215–232 (1958).

59. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32, DOI: https://doi.org/10.1023/A:1010933404324 (2001).

60. Dreyfus, S. E. Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. *J. Guid. Control. Dyn.* **13**, 926–928, DOI: 10.2514/3.25422 (1990). https://doi.org/10.2514/3.25422.

61. Gurney, K. *An Introduction to Neural Networks* (Taylor & Francis, Inc., USA, 1997).

62. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas imminent in nervous activity. *The Bull. Math. Biophys.* **5**, 115–133, DOI: 10.1007/BF02478259 (1943).

63. Cox, D. R. Regression models and life-tables. *J. Royal Stat. Soc. Ser. B (Methodological)* **34**, 187–220 (1972).

64. Fukushima, K. Cognitron: A self-organizing multilayered neural network. *Biol. Cybern.* **20**, 121–136, DOI: https://doi.org/10.1007/BF00342633 (1975).

65. DreamQuark. TabNet : Attentive interpretable tabular learning. https://github.com/dreamquark-ai/tabnet (2020).

Glossary

AHA-PCE American Heart Association - Pooled Cohort Equations
AI artificial intelligence
AMI acute myocardial infarction
ASCVD atherosclerotic cardiovascular disease
BMI body mass index
C17 calendar 2017
CAD coronary artery disease
CMS Centers for Medicare and Medicaid Services
CPT Current Procedural Terminology
CVD cardiovascular disease
DOE United States Department of Energy
EMR electronic medical record
EnsTL ensemble transfer learning
GLM-LASSO logistic regression with LASSO model selection
HbA1C hemoglobin A1C
ICD International Classification of Diseases
IRB Institutional Review Board
LDL low-density lipoprotein
LDL-DS LDL - decision support
MACE major adverse cardiovascular events
ML machine learning
MS metabolic syndrome
MVP Million Veteran Program
NDI National Death Index
NN fully connected neural network
OMHSP Office of Mental Health and Suicide Prevention
ORD Office of Research and Development
PC prospective screening cohort
Rcc nested retrospective case control
RF random forest
SI Supplementary Information
SME subject matter expert
TBI traumatic brain injury
TL transfer learning
VA US Department of Veterans Affairs
VHA US Department of Veterans Healthcare Administration
Visit-DS Visit - decision support

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MACEpredictionSI.pdf
- CVD.pdf