Original Paper

Explainable AI to Guide the Design of Interpretable Prediction Models From High-Dimensional Data, Applied to Falls and Related Injuries in Adults Aged 50 Years or Older: Algorithm Development and Validation

Junyu Yan[1,2*], MSc; Damian Machlanski[1*], PhD; Kurt Butler[1], PhD; Panagiotis Dimitrakopoulos[1], PhD; Ewen M Harrison[3], PhD; Bruce Guthrie[2], PhD; Sotirios A Tsaftaris[1], PhD

[1]School of Engineering, University of Edinburgh, Edinburgh, United Kingdom

[2]Advanced Care Research Centre, Usher School of Population Health Sciences, University of Edinburgh, Edinburgh, United Kingdom

[3]Centre for Medical Informatics, Usher School of Population Health Sciences, University of Edinburgh, Edinburgh, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Damian Machlanski, PhD
School of Engineering
University of Edinburgh
Usher Building
5 Little France Rd
Edinburgh, Scotland, EH16 4UX
United Kingdom
Phone: 44 7563942201
Email: d.machlanski@ed.ac.uk

Abstract

Background: Transparent prediction models, such as Cox regression, remain the standard for clinical risk prediction. However, they are challenging to design optimally by hand when tens or even hundreds of features require selection, transformation, or interaction modeling. While complex machine learning models offer high performance, their "black-box" nature limits the clinical trust, transparency, and interpretability required for decision-making.

Objective: This study aimed to develop an explainable AI-based recommender that provides data-driven feature-related recommendations that, when incorporated, improve the predictive performance of existing interpretable statistical models on high-dimensional data, and to demonstrate it by developing and internally validating a Cox model for the 1-year risk of falls or related injuries (FRIs).

Methods: We used linked primary and secondary care records from DataLoch (Lothian, Scotland) for adults aged 50 years or older registered with a contributing general practice between January 2015 and February 2020. The outcome was the first FRI within 1 year of the index date. A random survival forest was explained with Shapley Additive Explanations (SHAP), and its feature attributions were translated into recommendations for a Cox model: feature exclusions, nonlinear terms, and interactions. Data were split 80/20 into training and test sets. We compared the recommendations-augmented Cox model with the baseline Cox model, a Cox model with restricted cubic splines, and a penalized Cox model with all pairwise interactions. The metrics were Harrell concordance index (C-index), Uno C, calibration slope, and integrated Brier score. We repeated the analysis with all 409 available features and in 3 public datasets (German Breast Cancer Study Group 2 [GBSG2], AIDS Clinical Trial [ACT], and PeakVO2).

Results: Of 220,341 patients (mean age 64, SD 12 years), 8489 (3.85%) had an FRI within 1 year. From 59 candidate features, the recommender excluded 15 (8 medication classes and 7 comorbidities). The final model retained 44 predictors (demographic and lifestyle factors, frailty, 21 comorbidities, and 15 medication classes), with nonlinear terms for age, frailty, BMI, and area deprivation, and interactions of age with stroke, osteoporosis, loop diuretics, and frailty. The augmented model had a C-index of 0.719 (95% CI 0.708-0.731) compared with 0.715 (95% CI 0.702-0.727) for the baseline model (difference 0.004, 95% CI 0.001-0.007), with 70 vs 73 parameters, and similar calibration (slope 0.962 vs 0.961). With all 409 features, the recommender recommended 304 exclusions, 4 nonlinear terms, and 2 interactions. The augmented model had a C-index of 0.722 compared

XSL•FO
RenderX

with 0.713, with 129 vs 423 parameters, and better calibration (slope 0.952 vs 0.846). The method also proved effective on 3 additional public datasets, demonstrating wider applicability.

Conclusions: The proposed exploratory AI recommender demonstrates the potential of explainable AI to guide the design of transparent predictive models.

(J Med Internet Res 2026;28:e105364) doi: 10.2196/105364

KEYWORDS

explainable artificial intelligence; explainability; machine learning; risk prediction; prediction models; feature selection; feature interactions; feature attribution; study design; high-dimensional data; Cox proportional hazards; Shapley values

Introduction

Predictive modeling is important in health care as it enables proactive clinical decisions by forecasting future health outcomes from patient data. A prime example is models for predicting 10-year cardiovascular risk, including PREVENT [1] and QRISK [2], which support clinician and patient decision-making about the initiation of preventive medications. Furthermore, established tools such as QFracture [3] and FRAX [4] estimate the 10-year risk of fragility fracture, but falls, a more frequent and more immediate outcome, are not their target, and in primary care, frailty-based measures such as the electronic Frailty Index (eFI) [5] are commonly used to identify patients at risk of falls. Falls and fragility fractures are not evenly distributed across the population as their incidence rises steeply with age. Among older adults treated for fall injuries, women sustain fractures at about twice the rate of men [6], and in the United Kingdom, fracture rates are lowest in Black individuals (African, Caribbean, or Other Black), are almost 50% higher in Scotland than in London and South East England, and, for hip fracture in men, are higher in more deprived areas [7]. These differences matter for models built from routinely collected records, in which recording, care-seeking and the ascertainment of predictors may also differ between groups.

Historically, clinical predictive modeling has been the domain of standard statistical methods. Techniques such as the Cox proportional hazards (CPH) model have served medical research for decades [8,9]. They are computationally efficient, mathematically rigorous, and, crucially, intrinsically interpretable. However, as the dimensionality (ie, the number of variables) of health care data has increased (eg, electronic health care records), the limitations of these classical approaches have become increasingly apparent. Standard statistical methods, while trustworthy, often rely on linearity assumptions that may not reflect biological reality. In clinical practice, the risk factor-outcome relationship is rarely a straight line. For instance, the risk associated with BMI or blood pressure may follow a U-shaped curve (ie, nonlinearity), or the impact of a specific biomarker may be additionally dependent on a patient's age or genetic profile (ie, interaction). Capturing these nuances through "feature engineering," the manual process of transforming variables and specifying interactions, becomes intractable as the amount of data grows. For example, with only 100 features, the number of 2-way interactions reaches 4950 possibilities, making it nearly impossible for researchers to manually identify every clinically relevant synergy.

Modern AI predictors have demonstrated the capacity to outperform classical approaches in a variety of clinical tasks [10]. These models excel at "end-to-end learning," in which the model learns expressive feature representations directly from raw data without the need for manual feature engineering [11,12]. This allows AI models to capture nonlinearities and high-order interactions that would be invisible to standard techniques. However, this performance gain comes at a significant cost: the loss of clinical "sense-checking." When a model functions as a "black-box," clinicians cannot verify whether a prediction is based on sound pathophysiology or mere statistical noise and "shortcuts" in the dataset [13]. While some argue that high predictive performance should take precedence over interpretability in clinical AI [14,15], this lack of transparency prevents clinicians from explaining why a patient is at risk, hindering their ability to identify high-risk profiles that warrant further clinical investigation and undermining the trust required for clinical adoption in high-stakes environments [16-18].

To address the black-box problem, the field of explainable AI has emerged [19]. One line of explainable AI research focuses on feature attribution (FA), the process of assigning "credit" to input features for a model's prediction using methods like Shapley Additive Explanations (SHAP) [20]. Explainable AI methods have been successfully used to analyze risk predictors [21-24], identify disease biomarkers [25,26], and identify factors in cardiovascular risk [27,28].

This study explores the potential of explainable AI not merely as a tool for post hoc justification of black-box predictions, but as a mechanism to design better white-box models, a process we refer to as data-driven study design. While previous work investigated automated feature engineering with simple models [29] and augmented Cox models with explainable AI-derived new predictors [30], it remains unclear whether explainable AI can automate all 3 feature engineering tasks (feature selection, nonlinear terms, and interaction terms) to improve the performance of conventional clinical models on high-dimensional health data. Our method aims to bridge this gap by gaining the predictive strength of complex AI models while retaining the simplicity of conventional predictors. The recommender is intended for researchers and analysts designing prediction studies. We demonstrate it with a prognostic model for falls or related injuries (FRIs) in adults aged 50 years or older.

This study aimed to (1) develop an explainable AI-based recommender that suggests feature exclusions, nonlinear terms,

XSL•FO
RenderX

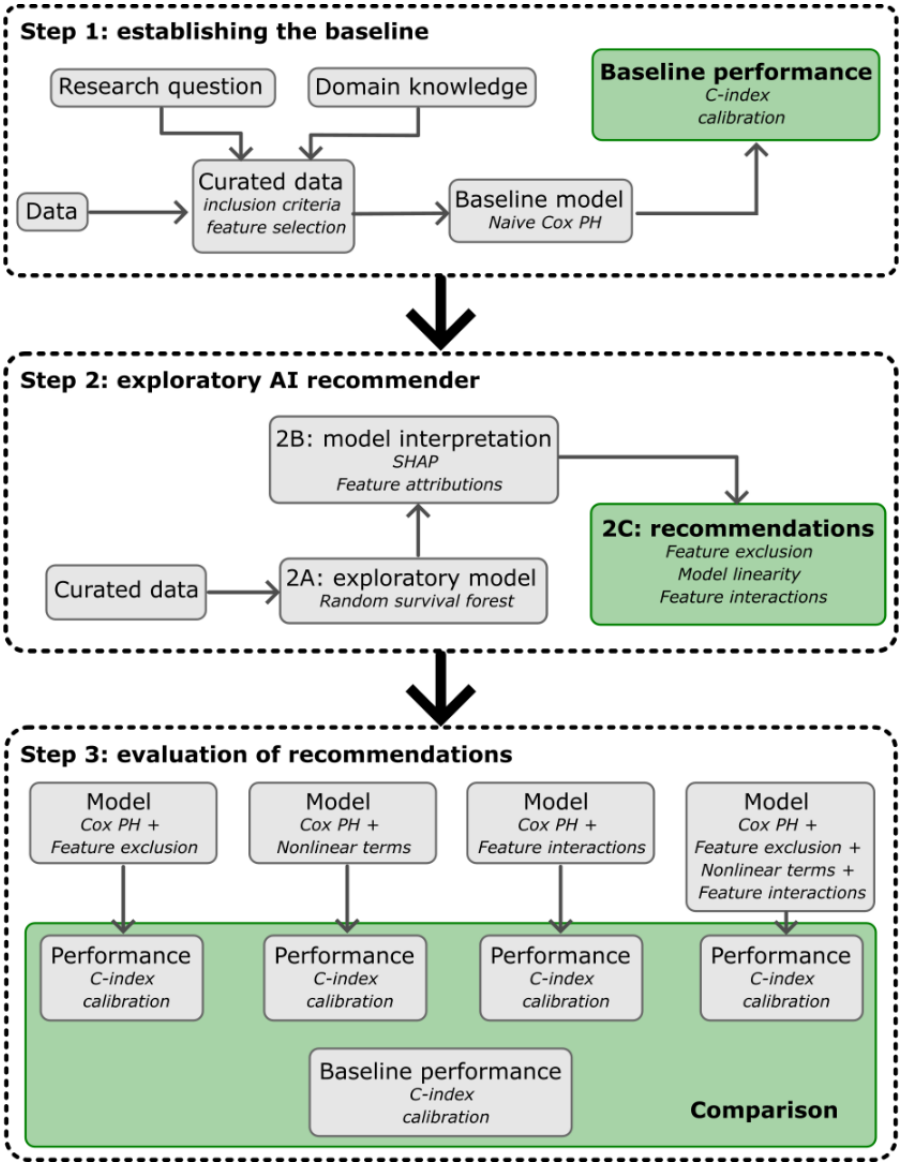
and interactions for transparent statistical prediction models, specifically Cox models; (2) demonstrate the recommender by developing and internally validating a model for the 1-year risk of FRIs in DataLoch, using both 59 selected features and all 409 available features; and (3) assess its applicability in 3 additional public datasets.

Methods

Overview

Our study proposes an explainable AI-based exploratory AI recommender (recommender henceforth) designed to systematically uncover inherent data relationships and optimize the design of predictive cohort studies. The methodology follows three stages (Figure 1): (1) establishing a baseline predictive model, representing the initial analytic stage of a research team; (2) deploying our recommender to generate actionable insights and recommendations for refining study design; and (3) evaluating the recommender's utility by quantifying the performance gains of models augmented with these recommendations. We apply this methodology to a clinical case study: a prognostic model for the 1-year risk of FRI, developed using DataLoch data.

Figure 1. Overview of the incorporated methodology. The incorporated methodology consists of three main steps. Step 1 intends to mimic the usual knowledge-driven practice without any AI-based tools. The raw data is curated by the research team in accordance with their research question and using their domain knowledge. The resulting curated data is then used with a naive Cox model to obtain the baseline performance. Step 2 demonstrates the use of our proposed exploratory AI recommender, a data-driven framework to obtain feature-related recommendations. Curated data, prepared in the previous step, is fed into the exploratory model (step 2a) that uses flexible AI methods (eg, random survival forest) to capture complex data patterns. Model interpretation techniques (step 2b), such as feature attributions, are then applied to extract those patterns from the exploratory model. Next, feature attributions are processed with a predefined set of rules (see Figure 2) to obtain feature-related recommendations (step 2c). In step 3, we evaluate the effectiveness of the provided recommendations. To achieve this, we take the baseline model from step 1 and extend it with recommendations from step 2c. The performance of augmented models is then compared to the performance of the baseline model in terms of both discrimination and calibration. PH: proportional hazards; SHAP: Shapley Additive Explanations.



XSL•FO
RenderX

Data Source and Study Population

The analysis was conducted on a high-dimensional clinical dataset provided by Lothian DataLoch, a secure, regional data repository linking primary and secondary care data in South East Scotland [31]. Because DataLoch holds the routinely collected records of all patients registered with contributing general practices, the cohort is population-based rather than a selected sample. It represents adults aged 50 years or older in one Scottish health board region. This study included all individuals aged 50 years or older, permanently registered for at least 1 year with a Lothian general medical practice (GP) that contributed to the DataLoch dataset between January 1, 2015, and February 28, 2020. This pre–COVID-19 period was specifically chosen to mitigate potential biases on FRI that might arise from pandemic-related changes in health care–seeking behavior. The index date was defined as the latest of January 1, 2015, the patient's 50th birthday, or the date on which 1 year of continuous registration was completed. Each patient entered the cohort once. Follow-up began on the index date and ended at the earliest of the first subsequent FRI, death, GP deregistration, 1 year after the index date, or February 28, 2020. Individuals were excluded if data for age, sex, or the Scottish Index of Multiple Deprivation (SIMD) were missing. The dataset was randomly split into 80% training (176,272 rows) and 20% held-out test data (44,069 rows). This study follows the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD+AI) reporting guidelines [32] (checklist in Multimedia Appendix 1). No study protocol was published, and the study was not registered.

Outcomes and Predictors Ascertainment

The primary outcome was the time from cohort entry to the first occurrence of FRI, with related injuries defined as low-energy fractures of hip, wrist, and proximal humerus. Diagnosis of FRI was defined by clinical codes from primary care (Read version 2), hospital admissions (*International Statistical Classification of Diseases and Related Health Problems 10th Revision* [*ICD-10*]), and emergency department records (Table S1 in Multimedia Appendix 2). Death before FRI was treated as a censoring event, so the analysis modeled the cause-specific hazard of FRI.

The primary analysis used 59 features selected from 409 available features based on established literature following the knowledge-driven approach. These include 4 patient demographic characteristics, 3 lifestyle risk factors, 28 comorbidities, the eFI [5], and exposure to 23 Fall-Risk-Increasing Drugs (list of features in Table S4 in Multimedia Appendix 2). Comorbidities were identified from all available primary care and hospital records dated before the index date. Exposure to each medication class was defined as at least 1 recorded dispensing during the 84 days before the index date. The same ascertainment window was applied to all medication classes. BMI, eFI, smoking status, and alcohol consumption were obtained from the most recent record before the index date. Age was calculated at the index date. The SIMD quintile was derived from the patient's residential postcode. For patients with more than 1 SIMD record, the most frequent quintile was used. Age and eFI were continuous variables, and BMI category and SIMD quintile were coded using ordered integer codes. Ethnicity, smoking status, and alcohol consumption were one-hot encoded, with missing values treated as separate categories. We chose this approach because missingness in routine records can itself be informative, and a model applied to routine data will encounter the same missing values [33]. The remaining predictors were binary. Full predictor definitions and preprocessing are provided in Note S1 in Multimedia Appendix 3.

Design, Implementation, and Evaluation of the Exploratory AI Recommender

First, we established a baseline predictive model using the standard multivariable CPH model. This model represents the starting point of a conventional, literature-driven predictive modeling study [28]. The baseline model was fitted on the training data using 59 selected candidate predictors, deliberately omitting advanced techniques such as feature selection, nonlinear terms, or interaction terms. Its performance on the held-out test set served as the baseline for evaluating recommender-driven refinements.

Second, we deployed the recommender to systematically explore underlying data relationships and generate concrete recommendations for refining the above knowledge-driven predictive study design. This began with fitting an advanced exploratory AI model to the training data to capture complex, nonlinear relationships and potential interactions between all features and outcomes. For this purpose, we used random survival forest (RSF) [34] as the exploratory AI model. Unlike typical machine learning applications, the RSF was used purely to capture high-dimensional data structures rather than for final clinical prediction. RSF was fitted with 300 trees, with min samples leaf as 100, and random seeds as 20 (Table S5 in Multimedia Appendix 2). The RSF returns a risk score, calculated as the sum of the predicted cumulative hazards over the distinct event times in the training data. We used the logarithm of this score as the model output to be explained. Next, the RSF model was interpreted using explainable AI techniques, such as SHAP [20], to translate "black-box" patterns into actionable insights. SHAP quantifies each feature's contribution (termed FAs) to the RSF's prediction for each patient (see Note S2 in Multimedia Appendix 3 for further details on FAs). FAs are expressed on the log scale of the risk score. On this scale, an additive effect of a feature yields an FA that depends only on that feature, and multiplying the score by a constant adds the same constant to every prediction. The FAs therefore do not depend on the arbitrary scale of the score. Standard FA analysis often relies on population-level averaging, which can obscure risk factors that are clinically significant only within specific subgroups. We therefore analyzed FAs in 2 subcohorts targeted at different levels of predicted risk: a low-risk and high-risk subcohort. Each subcohort had a reference patient: the mean feature vector of all training patients without an FRI (low-risk subcohort) or with an FRI (high-risk subcohort). The low-risk subcohort included training patients whose RSF log risk score differed from that of the low-risk reference patient by less than a margin, measured in SDs of the log risk score among training patients without an FRI. The high-risk subcohort was defined in the same way, relative to

XSL•FO
RenderX

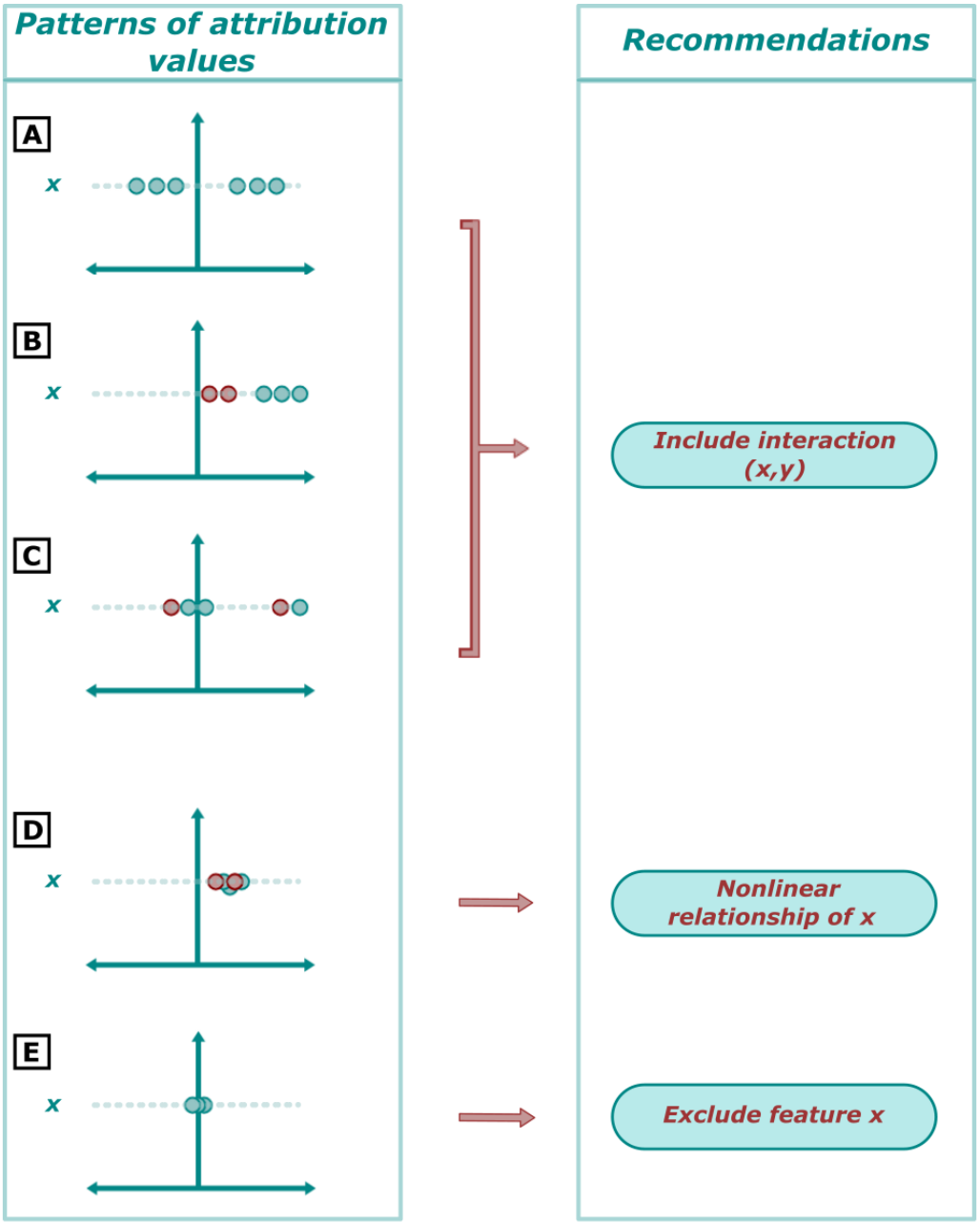
the high-risk reference patient and training patients with an FRI. The margin was chosen on the training data from a prespecified grid (0.025 to 1.0 SD). We selected the smallest value at which the recommendations did not change when the margin was widened further, because a narrower margin keeps each subcohort close to its reference patient (Note S4 in Multimedia Appendix 3). FAs in each subcohort were computed relative to the same reference patient and therefore describe deviations from a typical patient of that group (Note S2 in Multimedia Appendix 3). The FAs from these cohorts were then processed to derive modeling recommendations. While computing FAs might be more intensive on average than training a baseline Cox model, this exploratory phase is a one-time cost that precedes the deployment of the more efficient augmented model.

Our recommender offers the following 3 types of recommendations (Figure 2): (1) Feature exclusion (Figure 2E). A feature was recommended for exclusion if the upper bound of the 95% CI of its mean absolute FA was below a data-driven significance threshold. The threshold was defined as 5% of the standard deviation of mean absolute FA values across all candidate features, a heuristic to distinguish features with negligible contributions. (2) Nonlinear relationships (Figure 2D). We screened continuous and ordinal features for potential nonlinear effects. If a feature has a linear effect in the Cox model, its FAs are a linear function of its values on the log-risk scale. For each feature, we therefore regressed its FAs on its values using both a straight line and a flexible function. For ordinal features, the flexible function estimated each category's mean FA by using 0/1 indicator variables, allowing changes between successive levels to differ. For continuous features, it was a natural cubic spline with 4 degrees of freedom. The gain in the coefficient of determination (R^2) from the straight line to the flexible function, that is, the additional variation in the FAs explained by allowing a nonlinear shape, was tested with a nested F test. A feature was recommended if its adjusted P value (see Statistical Analysis) was below .05 and its gain in R^2 exceeded 0.02 [35]. The validation of the nonlinearity rule is through a simulation considering different cases of linear and nonlinear data generation processes (Table S8 in Multimedia Appendix 2); (3) Feature interactions (Figure 2; A-C). Interactions were screened in 4 steps using training data only (Note S3 in Multimedia Appendix 3). First, features whose FAs were not explained by their own values (lower limit of the 95% bootstrap interval of the normalized residual SD ≥0.10) were selected as focal features. Second, each focal feature was paired with every other candidate feature, and we tested whether the second feature's within-stratum FA contrast (binary features) or product-term coefficient (continuous or ordinal features) varied with the focal feature, using Wald tests with robust standard errors. Under an additive model with a fixed reference, these quantities do not change with differences between strata, so a pair is flagged only if the model is nonadditive on the explained scale. Pairs with an adjusted P<.05 (Bonferroni within pair, then Benjamini-Hochberg across pairs) and an effect of at least 10% of the mean absolute FA advanced. The interaction screening rule is validated through simulation on only additive null interaction, weak interaction (0.1), and strong enough interaction with type I error below 10%, and interaction detection rate above 90% in both interaction cases (Table S9 in Multimedia Appendix 2). Third, pairs were retained if a Cox score test against a main-effects reference model had a Benjamini-Hochberg-adjusted P<.05. Fourth, the number of added parameters was capped using the sample-size criterion of Riley et al [35].

Third, we evaluated the effectiveness of the recommender. We constructed augmented Cox models that integrated these recommendations (excluding features, adding nonlinear terms and interactions) to predict the same 1-year risk of FRI. For nonlinear terms, continuous features entered the model with quadratic terms, and ordinal features as the indicator variables described above. All recommendations were derived in one pass from the FAs of the same exploratory model. Feature exclusion was applied first, and the nonlinearity and interaction rules were applied only to the retained features. To preserve the interaction hierarchy, an interaction was included only when both main effects were retained. Therefore, the order in which the recommendations are added does not change the final model. We then compared the predictive performance of these augmented models against the baseline on the held-out test set. We also compared the augmented Cox models with 2 approaches commonly used in routine prediction modeling. These were a Cox model with restricted cubic splines for continuous predictors (Cox-RCS) and a penalized Cox model including all main effects and all pairwise interactions (Coxnet; least absolute shrinkage and selection operator [LASSO], with the penalty selected by 10-fold cross-validation on the training data). Furthermore, the clinical plausibility of generated recommendations was discussed in light of the published literature (See Discussion).

XSL•FO
RenderX

Figure 2. Step 2 of the proposed method: the 3 types of recommendations derived from the feature attributions of the exploratory model. Each panel plots a feature's attributions against its values in one subcohort. Feature attributions are computed against one reference point per subcohort: the mean feature vector of patients without (low-risk) or with (high-risk) the event. (A-C) Interaction: at the same value of a feature, its feature attributions vary, which suggests that another feature modifies its contribution. Candidate pairs are screened on the feature attributions, checked with a score test in a Cox model, and limited by the sample-size criterion. (D) Nonlinear term: the feature attributions follow a curve rather than a straight line. A nonlinear term is recommended if, in either subcohort, a flexible function fits the FAs significantly better than a straight line (false discovery rate (FDR)-adjusted $P<.05$) and the gain in R^2 exceeds 0.02. (E) Exclusion: the feature attributions are close to zero. A feature is recommended for exclusion if, in both subcohorts, the upper limit of the 95% CI of its mean absolute feature attribution is below 5% of the SD across all features.



Application to the Full Feature Set

To examine the Recommender in a very high-dimensional setting, we applied it to all 409 available features, without literature-based selection. The pipeline was unchanged: the RSF was fitted on all 409 features, FAs were computed in newly selected low-risk and high-risk subcohorts, and the 3 recommendation rules were applied.

Application to Other Datasets

To assess the applicability of the Recommender across datasets, outcomes, and clinical settings, we applied the Recommender on 3 additional, publicly available time-to-event datasets: German Breast Cancer Study Group 2 (GBSG2; breast cancer), AIDS Clinical Trial (ACT; human immunodeficiency virus), and PeakVO2 (systolic heart failure) [36-38].

GBSG2

The data were obtained from a randomized clinical trial that compared different chemotherapy treatment regimens and their effects on survival in patients with breast cancer. This study group consists of 686 patients and 8 features. The outcome was recurrence-free survival, defined as the time to breast cancer

XSL•FO
RenderX

recurrence or death. Overall, 299 (43.6%) patients experienced an event, and 387 (56.4%) were right-censored. The features were age, estrogen receptor level (estrec), the number of positive (involved) lymph nodes (pnodes), progesterone receptor level (progrec), hormonal therapy (horTh), tumor grade (tgrade), menopausal status (menostat), and tumor size (tsize).

ACT

This dataset contains 1151 patients with HIV from a randomized controlled trial that compared different drug regimens. The dataset includes 11 features: treatment indicator (tx), treatment group indicator (txgrp), CD4 stratum at screening (strat2), sex, race/ethnicity (raceth), IV drug use history (ivdrug), hemophiliac (hemophil), Karnofsky Performance Scale (karnof), baseline CD4 count (cd4), months of prior ZDV use (priorzdv), and age. The endpoint was time to an AIDS-defining event or death, with a censoring rate of 89.4%.

PeakVO2 (Systolic Heart Failure)

This dataset includes cardiopulmonary exercise test records from 2231 patients with systolic heart failure. It contains 39 features covering demographic characteristics, comorbidities, and exercise test measurements. This dataset was used to assess the recommender in a higher-dimensional setting. The endpoint was all-cause mortality, with an event rate of 32.5%.

For each dataset, the cohort was randomly split into a training set (80%) and a test set (20%). The complete explainable AI-based exploratory AI recommender, including fitting a baseline Cox model, training an exploratory RSF model with 1000 trees (hyperparameter details in Table S5 in Multimedia Appendix 2), and applying SHAP to derive recommendations, was performed in the same way as for the primary dataset.

Statistical Analysis

Baseline characteristics of the study samples were described using frequencies and percentages for categorical variables and means (SDs) for continuous variables. Statistical analyses, model development, and testing were conducted in Python (version 3.12.5; Python Software Foundation), except for the proportional hazards tests, which were conducted in R (version 4.5.3) with the survival package (version 3.8.6; R Core Team). All preprocessing steps (encoding of categorical features, standardization of continuous features, and the centering used for quadratic terms) were estimated on the training set only and then applied unchanged to the test set. Discrimination was measured with Harrell concordance index (C-index) and Uno C (less sensitive to the censoring distribution, truncated at the 90th percentile of test-set follow-up times) [39]. Calibration was assessed with the conventional calibration slope. This is the coefficient of the complementary log-log transformation of the predicted survival probability at a fixed horizon in a Cox model fitted to the test set, with a Wald 95% CI. The horizons were 365 days for DataLoch, 1500 days for GBSG2, 320 days for ACT, and 1095 days for PeakVO2. Calibration plots compared the mean predicted survival with the Kaplan-Meier estimate in 5 quantile-based groups of predicted survival. Overall prediction error was summarized by the integrated Brier score (IBS). Unless stated otherwise, 95% CIs are percentile intervals from 1000 bootstrap resamples of the test set. Differences in C-index relative to the baseline model were computed on the same resamples. Intervals in the calibration plots were 95% percentile CIs from 500 bootstrap resamples of the test set.

P values in the recommendation rules were adjusted with the Benjamini-Hochberg procedure, with the false discovery rate (FDR) controlled at 5%. For the nonlinearity rule, the adjustment was across all candidate features within each subcohort. For the interaction screen, the repeated tests of each pair were first combined by Bonferroni adjustment, and then adjusted across all unique feature pairs. For the Cox score test, the adjustment was across the pairs that passed the screen. Proportional hazards were assessed for the augmented Cox model fitted to the training data, with global and per-term Schoenfeld residual tests. All tests were 2-tailed, and $P<.05$ was considered significant.

Ethical Considerations

All data from DataLoch were linked and deidentified by the DataLoch service (Edinburgh, United Kingdom) and analyzed within a Secure Data Environment (reference: DL_2024_030). DataLoch enables access to deidentified extracts of health care data from the South-East Scotland region to approved applicants [40]. The study was reviewed and received approval under delegated authority from a regional National Health Service Research Ethics Committee (REC 22/NS/0093) and Caldicott Guardian. Individual patient consent was not required. Clinical trial number: not applicable. The public datasets are de-identified and openly available. Patients and the public were not involved in the design, conduct, reporting, or dissemination of this study.

Results

Overview

The DataLoch cohort included 220,341 patients (mean age 64, SD 12 years), of whom 119,026 (54%) were females. Ethnicity was self-identified as White (Gypsy/Traveller, Irish, Other British, Other White, Polish, Scottish) in 187,254 (85%), Asian (Bangladeshi, Chinese, Indian, Other Asian, Pakistani) in 3093 (1.4%), and Black (African, Caribbean, Other Black) in 757 (0.3%) patients. The most prescribed medications were angiotensin-converting enzyme inhibitors (46,438 patients, 21.1%), beta-adrenoceptor blocking drugs (34,809 patients, 15.8%), and calcium-channel blockers (33,054 patients, 15%). The most prevalent comorbidities were hypertension (94,477 patients, 42.9%) and depression (58,260 patients, 26.4%). During follow-up, 8489 (3.85%) patients experienced an FRI. Median observed follow-up was 1 (IQR 1-1) year, with a total of 207,278 person-years of observation. The crude FRI incidence rate was 40.95 per 1000 person-years. The crude FRI incidence rate was 51.10 per 1000 person-years in women and 29.14 per 1000 person-years in men. In the time-to-FRI analysis, 211,852 patients (96.15%) were censored, including 4040 (1.83% of the cohort) who died before experiencing an FRI. The training set included 176,272 patients with 6791 FRIs, and the test set included 44,069 patients with 1698 FRIs. Table S4 in Multimedia Appendix 2 contains the patients' characteristics of the study.

XSL•FO
RenderX

Baseline Performance

As shown in Table 1, the baseline CPH achieved a C-index of 0.715 (95% CI 0.702-0.727) and a conventional calibration slope at 365 days of 0.961 (95% CI 0.909-1.013). The exploratory RSF model achieved a C-index of 0.716 (95% CI 0.703-0.727). The forest discriminated similarly to the linear, additive Cox model, indicating that most of the predictive information in the candidate features was captured by their main effects.

Table 1. Performance of the baseline, exploratory, augmented, and comparator models of the 59 selected features on the test set.

Models	Parameters, n	C-index (95% CI)	Uno C	IBS[a]
Baseline models				
Baseline Cox PH[b]	73	0.715 (0.702-0.727)	0.714	0.0229
RSF[c]	—[d]	0.716 (0.703-0.727)	0.715	—
Augmented Cox PH models				
+ Exclusion	58	0.715 (0.702-0.727)	0.714	0.0229
+ Nonlinear terms	81	0.718 (0.707-0.730)	0.717	0.0229
+ Interactions	77	0.715 (0.702-0.727)	0.713	0.0229
+ All recommendations	70	0.719 (0.708-0.731)	0.718	0.0228
Comparators				
Cox-RCS	79	0.718 (0.707-0.731)	0.717	0.0230
Coxnet	2624[e]	0.706 (0.694-0.719)	0.705	0.0229

[a]IBS: integrated Brier score.

[b]PH: proportional hazards.

[c]RSF: random survival forest.

[d]Not applicable.

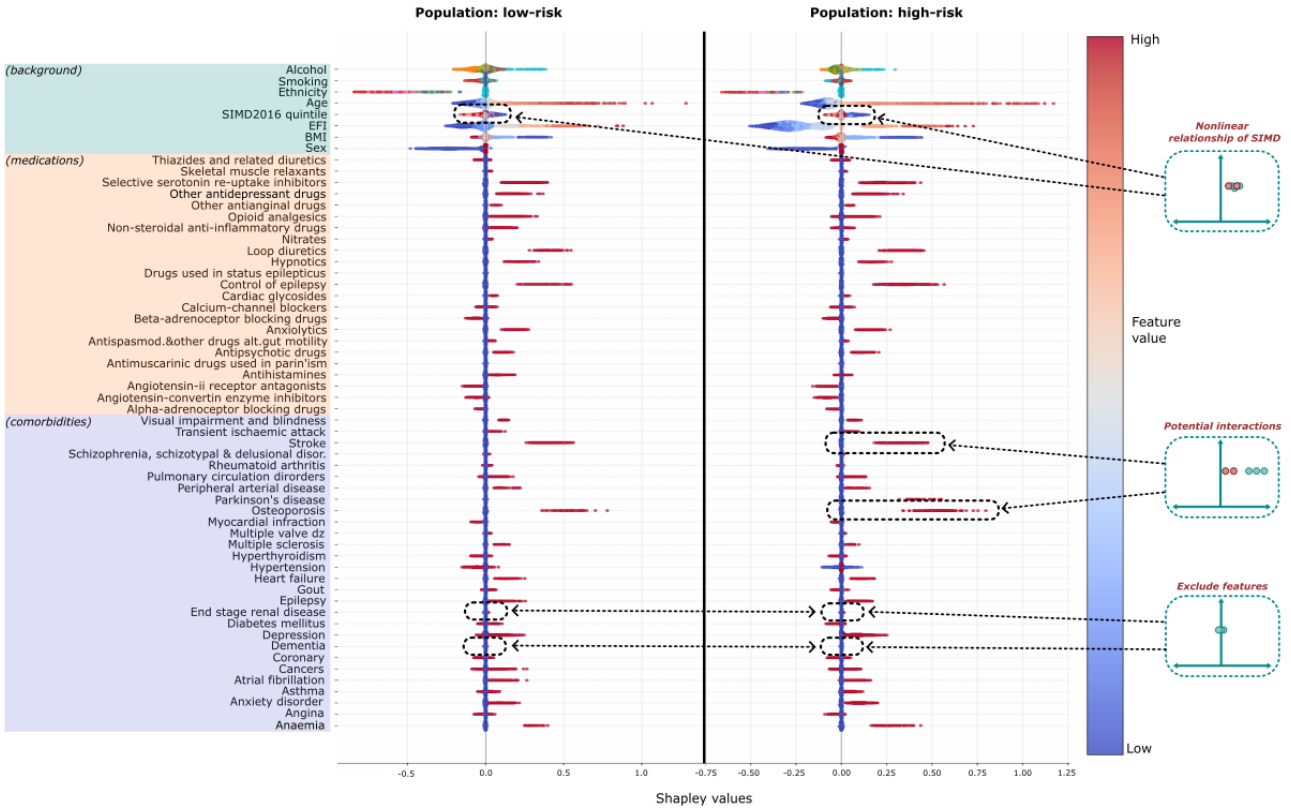
[e]This is the estimated parameter numbers for Coxnet, the actual coefficients for Coxnet are 23 for main effects and 18 for interactions.

Generated Recommendations

Figure 3 details the process for conducting FA analysis and deriving recommendations. The subcohort margin, chosen on the training data, was 0.05 SD of the log risk score. It selected 8208 patients into the low-risk subcohort and 5608 patients into the high-risk subcohort from the training set (n=176,272). The recommendations did not change when the margin was widened to 0.1 SD. The FA analysis was conducted in the 2 subcohorts separately, and the resulting recommendations are summarized in Table S6 in Multimedia Appendix 2. First, it recommended 15 features for exclusion, containing 8 medication classes and 7 comorbidities. Second, among the remaining 44 features, 4 features were recommended for nonlinear terms. Age and the eFI entered with an added squared term, and BMI category and SIMD quintile entered as level indicators (Table S15 in Multimedia Appendix 2). Finally, the interaction screen performed 1807 tests assessing 946 unique pairs among the 44 remaining features. Of these pairs, 370 passed the FA screen (step 2), and 4 passed the Cox score test (step 3). All 4 were within the allowance of 540 parameters set by the sample-size criterion (step 4). The 4 recommended interactions were age with stroke, age with osteoporosis, age with loop diuretics, and age with the eFI.

XSL•FO
RenderX

Figure 3. Feature attributions from a random survival forest estimating falls or related injuries risks among individuals within low-risk and high-risk cohorts. Feature attributions for low-risk (8208 patients) and high-risk (5608 patients) groups. Each dot corresponds to a single patient observation, and its position along the x-axis indicates its influence on the risk of a fall or related injury event. Positive feature attributions (right of 0) indicate that a feature increased the falls or related injuries risk, whereas negative features attributions (left of 0) indicate that a feature is correlated with decreased falls or related injuries risk. Color is used to indicate the value of included feature characteristics. Red and blue indicate higher and lower feature values, respectively. Medications and comorbidities are binary features; hence, red indicates the presence of the listed characteristic (eg, diagnosis of angina) and blue its absence (eg, no diagnosis of angina). Marked data points (dashed lines) showcase examples of how the predefined patterns (Figure 2) are being processed into recommendations in practice. SIMD: Scottish Index of Multiple Deprivation.

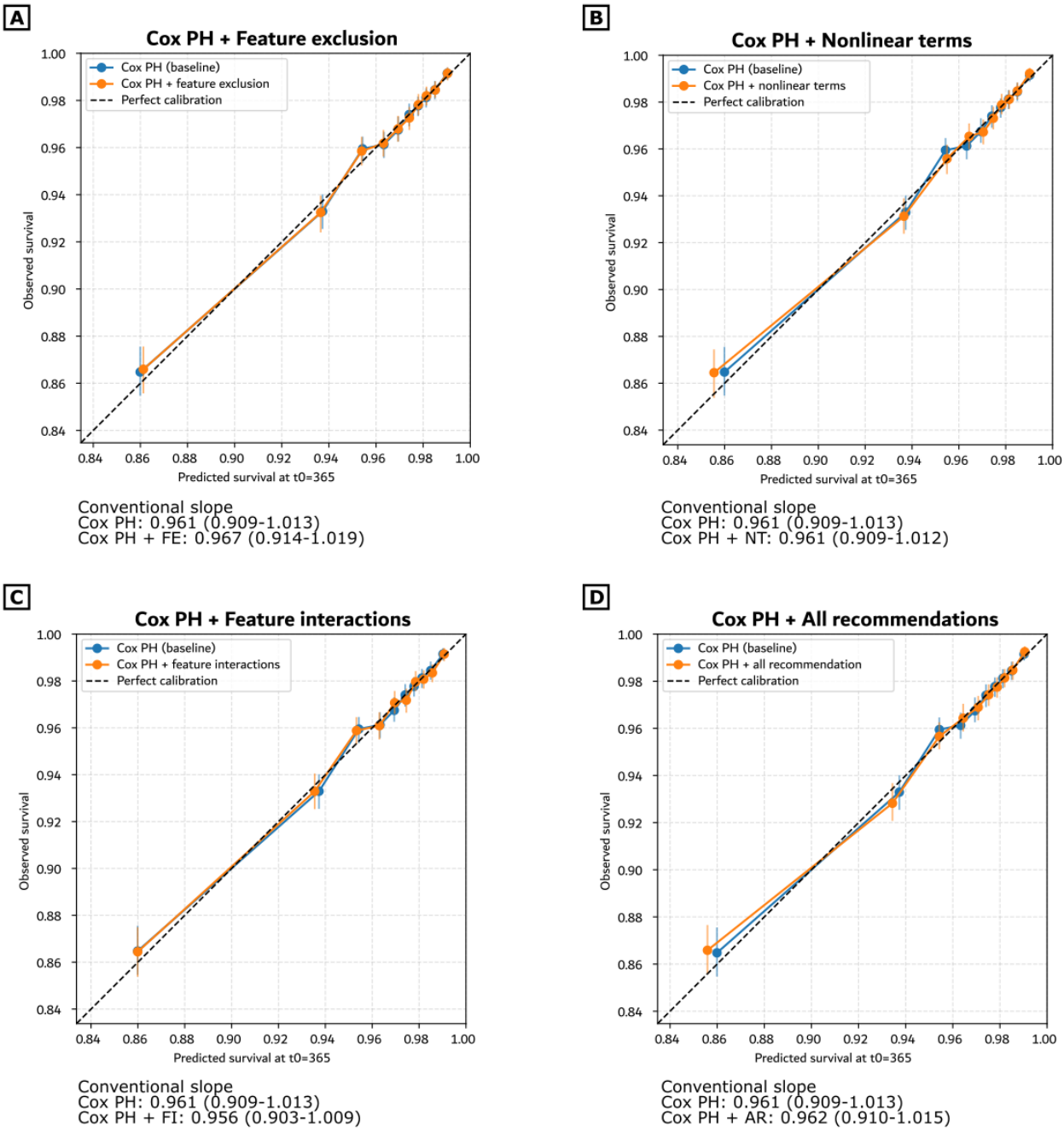


Quantitative Evaluation of Recommendations

We integrated the 3 types of recommendations into augmented Cox models independently and cumulatively to evaluate their individual and collective impact (Table 1 and Figure 4). In the independent assessment, an augmented Cox model incorporating only the feature exclusion recommendations (59 features reduced to 44) demonstrated no degradation in performance compared with the baseline Cox model (difference in C-index: 0.000, 95% CI –0.001 to 0.001; *P*=.87), while reducing the number of parameters from 73 to 58. Adding the nonlinear terms increased the C-index to 0.718 (95% CI 0.707-0.730; difference in C-index: 0.003, 95% CI 0.001-0.006; *P*=.02). Adding the 4 interactions did not improve discrimination (C-index 0.715, difference −0.001, 95% CI –0.002 to 0.000; *P*=.15). The model with all recommendations achieved the highest C-index, 0.719 (95% CI 0.708-0.731), an improvement of 0.004 (95% CI 0.001 to 0.007; *P*=.01) over the baseline model, with 70 parameters. Uno C followed the same pattern (0.713-0.718), and the IBS was almost identical across models (0.0228-0.0230), as expected with an event rate below 4%.

XSL•FO
RenderX

Figure 4. Evaluation of the recommendations provided by the proposed Recommender as compared to the baseline Cox model in the primary analysis (59 selected features). AR: all recommendations; FE: feature exclusion; FI: feature interactions; NT: nonlinear terms; PH: proportional hazards.



Compared with the 2 comparators, Cox-RCS achieved a C-index of 0.718 (95% CI 0.707-0.731) with 79 parameters, similar to the model with the recommended nonlinear terms. The model with all recommendations demonstrated slightly higher discrimination performance than Cox-RCS with fewer parameters (70 vs 79). Coxnet had lower discrimination performance than the baseline model (C-index 0.706, 95% CI 0.694-0.719). It retained 18 interactions, which included all 4 interactions recommended by the recommender.

Calibration at 365 days was close to ideal for all Cox models, with calibration slopes of 0.961 (95% CI 0.909-1.013) for the baseline model and 0.962 (95% CI 0.910-1.015) for the model with all recommendations (Figure 4). The final augmented model had 70 coefficients. The coefficient of each parameter in the augmented model is shown in Table S10 in Multimedia Appendix 2. With 6791 events in the training set, this corresponds to about 97 events per parameter. For the final augmented model, the global Schoenfeld residual test showed no evidence against the proportional hazards assumption (χ^2_{70}=64.44; *P*=.67), and 4 of the 70 terms had a *P* value of <.05 (Table S11 in Multimedia Appendix 2).

Evaluation of Recommendations on the Full Feature Set

When applied to all 409 available features, the recommender recommended 304 exclusions, 4 nonlinear terms, and 2 interactions (Table 2). The nonlinear terms were recommended for age, the eFI, BMI category, and SIMD quintile. The

XSL•FO
RenderX

interactions were between age and the eFI, and lipid-regulating drugs and drugs used in megaloblastic anemia. As in the primary analysis, the model with all recommendations achieved the highest C-index, 0.722 (95% CI 0.710-0.734). It showed an improvement of 0.009 (95% CI 0.004-0.014; *P*<.001) over the baseline model, with much fewer parameters (129 vs 423). Exclusion improved discrimination by 0.006 (95% CI 0.002-0.010; *P*<.001) and nonlinear terms by 0.003 (95% CI 0.001-0.005; *P*=.02). Adding the 2 interactions did not change discrimination (difference in C-index: 0.000, 95% CI −0.001 to 0.001; *P*=.30), as expected with so few terms. Uno's C increased from 0.712 to 0.720. The calibration slope improved from 0.846 (95% CI 0.797-0.894) for the baseline model to 0.952 (95% CI 0.901-1.003) for the model with all recommendations (Figure 5). The model with all recommendations had a significantly higher C-index than Cox-RCS (0.715; difference in C-index 0.006, 95% CI 0.002-0.010; *P*<.001) with about one-third of the parameters (129 vs 429). Coxnet could not be fitted because all pairwise interactions of 409 features give more than 80,000 candidate terms. For the final augmented model, the global Schoenfeld residual test showed no evidence against the proportional hazards assumption (χ^2_{129}=137.66; *P*=.28), and 10 of the 129 terms had a *P* value of <.05 (Table S12 in Multimedia Appendix 2).

Table 2. Performance of the baseline, exploratory, augmented, and comparator models of all 409 available features on the test set.

Models	Parameters, n	C-index (95% CI)	Uno C	IBS[a]
Baseline models				
Baseline Cox PH[b]	423	0.713 (0.701-0.725)	0.712	0.0230
RSF[c]	—[d]	0.717 (0.706-0.730)	0.716	—
Augmented Cox PH models				
+ Exclusion	119	0.719 (0.707-0.731)	0.718	0.0229
+ Nonlinear terms	435	0.715 (0.704-0.727)	0.714	0.0229
+ Interactions	425	0.713 (0.701-0.724)	0.712	0.0229
+ All recommendations	129	0.722 (0.710-0.734)	0.720	0.0228
Comparators				
Cox-RCS	429	0.715 (0.704-0.727)	0.714	0.0229
Coxnet	—	—	—	—

[a]IBS: integrated Brier score.

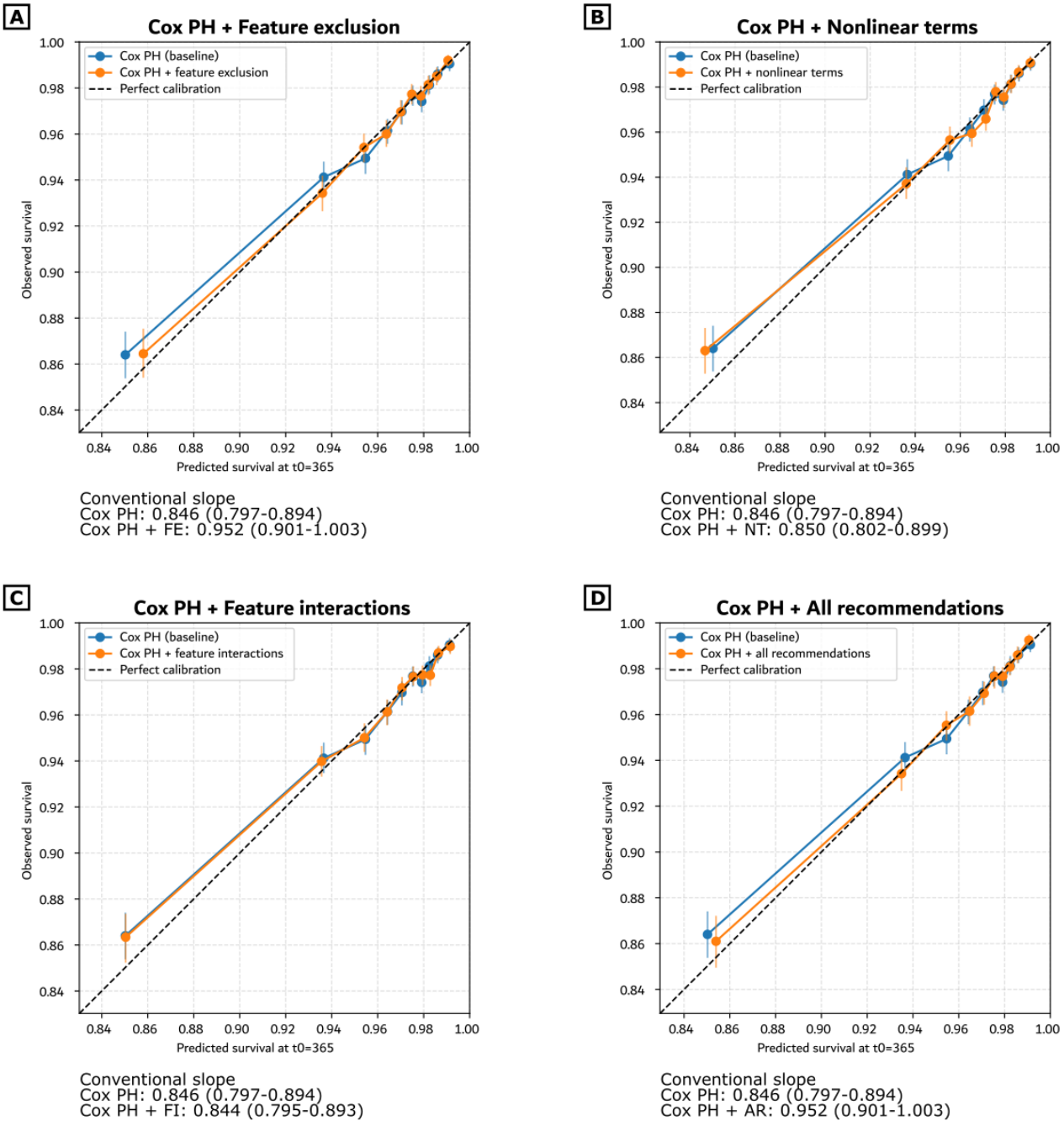
[b]PH: proportional hazards.

[c]RSF: random survival forest.

[d]Not applicable.

XSL•FO
RenderX

Figure 5. Evaluation of the recommendations provided by the proposed Recommender as compared to the baseline Cox model in the full feature set (all 409 features). AR: all recommendations; FE: feature exclusion; FI: feature interactions; NT: nonlinear terms; PH: proportional hazards.



Evaluation of Recommendations on Other Datasets

Table 3 summarizes the results in the 3 public datasets. Detailed recommendations and model performances are provided in Table S13 and S14, respectively, in Multimedia Appendix 2. In each dataset, the augmented model with all recommendations had a higher C-index than the baseline Cox model (GBSG2: 0.701 vs 0.667; ACT: 0.749 vs 0.743; PeakVO2: 0.677 vs 0.670).

No interaction entered the model in any of the 3 datasets. The Recommender identified some candidate interactions, such as beta-blockers with resting heart rate in PeakVO2. However, in these smaller datasets, the main effects and nonlinear terms already used the parameter allowance of the sample-size criterion.

In GBSG2 (Figure 6), only nonlinear terms were recommended. The resulting model had a higher C-index than Cox-RCS (0.701 vs 0.687) with fewer parameters (14 vs 23), and the calibration slope moved from 1.24 to 0.94. Coxnet, which retained 2 main effects and 3 interactions, had a lower C-index than the baseline model (0.659). In ACT and PeakVO2 (see Figures 7 and 8, respectively), the augmented model performed similarly to the Cox-RCS and Coxnet. The Recommender recommended nonlinear terms for most continuous features in both datasets, so the augmented model was close to Cox-RCS. Coxnet retained few main effects (7 in ACT and 5 in PeakVO2) and many interactions (31 and 25). This suggests that some of its

XSL•FO
RenderX

interaction terms may approximate nonlinear effects rather than true interactions. Coxnet had the best calibration in these 2 datasets (slopes 0.73 and 1.00), probably because penalization shrinks the coefficients and so moderates extreme predictions. The baseline, augmented, and Cox-RCS models had calibration slopes of 0.61 to 0.72 in these 2 datasets, which shows some overfitting. Among the comparators, Coxnet had the best calibration (slopes 0.73 and 1.00), probably because penalization shrinks the coefficients and so moderates extreme predictions. The LASSO-augmented model applies the same penalization to the recommended terms. It had C-indexes of 0.702, 0.761, and 0.677 and calibration slopes of 1.03, 1.02, and 0.89 in GBSG2, ACT, and PeakVO2, respectively.

Across the 3 datasets, the models built from the recommendations had similar or higher discrimination than the comparators. With penalization, they also had good calibration.

Table 3. Performance of the baseline, augmented, and comparator models of three public datasets on the test set.

Datasets/models	Parameters, n[a]	C-index (95% CI)	ΔC-index vs baseline (95% CI)	Calibration slope (95% CI)	IBS[b]
GBSG2[c]					
Baseline Cox	8	0.667 (0.585-0.739)	—[d]	1.24 (0.70-1.78)	0.161
+ All recommendations	14	0.701 (0.630-0.763)	0.034 (−0.005 to 0.076)	0.94 (0.61-1.28)	0.154
Cox-RCS	23	0.687 (0.623-0.749)	0.020 (−0.032 to 0.078)	0.86 (0.52-1.20)	0.160
Coxnet	36	0.659 (0.573-0.731)	−0.008 (−0.032 to 0.018)	1.16 (0.68-1.63)	0.165
ACT[e]					
Baseline Cox	14	0.743 (0.655-0.827)	—	0.65 (0.36-0.94)	0.0570
+ All recommendations	18	0.749 (0.664-0.832)	0.006 (−0.020 to 0.033)	0.64 (0.36-0.92)	0.0574
Cox-RCS	23	0.744 (0.655-0.828)	0.001 (−0.019 to 0.019)	0.61 (0.33-0.90)	0.0569
Coxnet	98	0.744 (0.652-0.836)	0.001 (−0.040 to 0.038)	0.73 (0.42-1.03)	0.0536
PeakVO2					
Baseline Cox	39	0.670 (0.627-0.716)	—	0.69 (0.49-0.90)	0.1370
+ All recommendations	43	0.677 (0.634-0.723)	0.006 (−0.003 to 0.016)	0.72 (0.52-0.92)	0.1356
Cox-RCS	78	0.674 (0.630-0.721)	0.004 (–0.008 to 0.016)	0.70 (0.50-0.89)	0.1374
Coxnet	778	0.674 (0.628-0.719)	0.003 (–0.011 to 0.020)	1.00 (0.72-1.28)	0.1338

[a] Number of estimated coefficients. For Coxnet, the number of candidate coefficients (main effects and all pairwise interactions); the numbers retained after penalization were 2 main effects and 3 interactions (GBSG2), 7 and 31 (ACT), and 5 and 25 (PeakVO2).

[b]IBS: integrated Brier score.

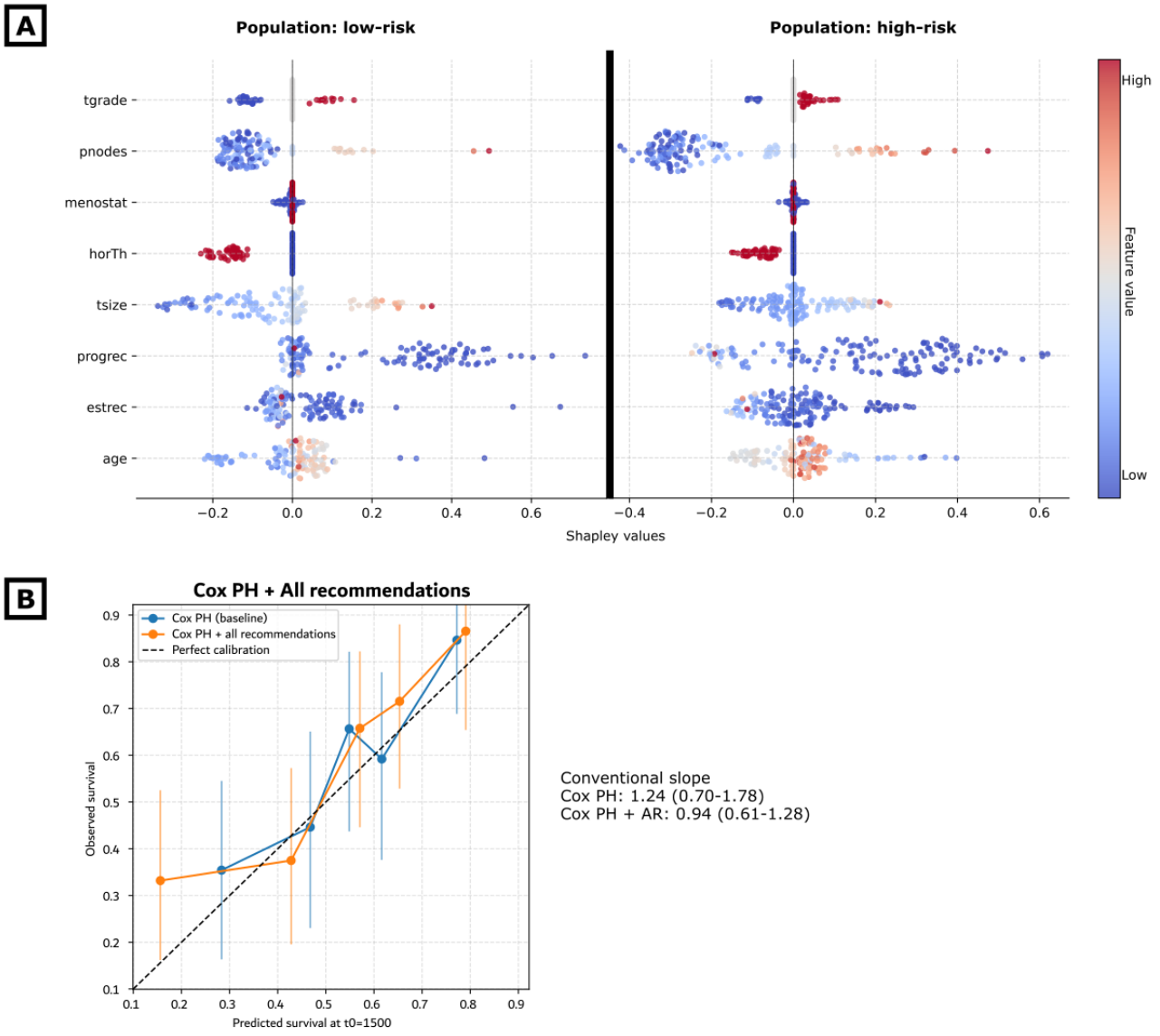
[c]GBSG2: German Breast Cancer Study Group 2.

[d]Not applicable.

[e]ACT: AIDS Clinical Trial.

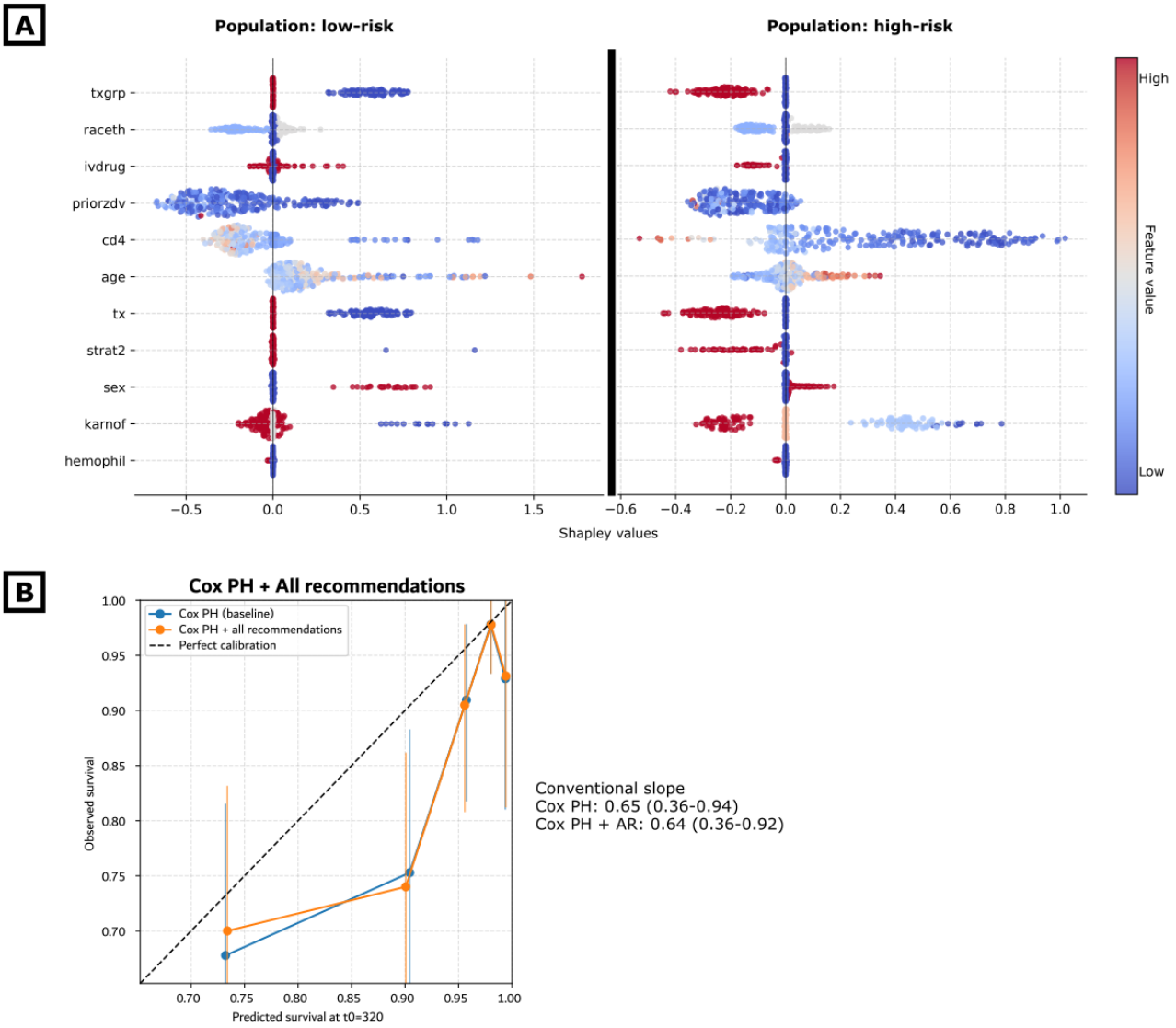
XSL•FO
RenderX

Figure 6. Evaluation of the Recommender on the GBSG2 public dataset. (A) Shapley values obtained from the RSF model. (B) Calibration of the Cox PH model augmented with all recommendations. AR: all recommendations; PH: proportional hazards.



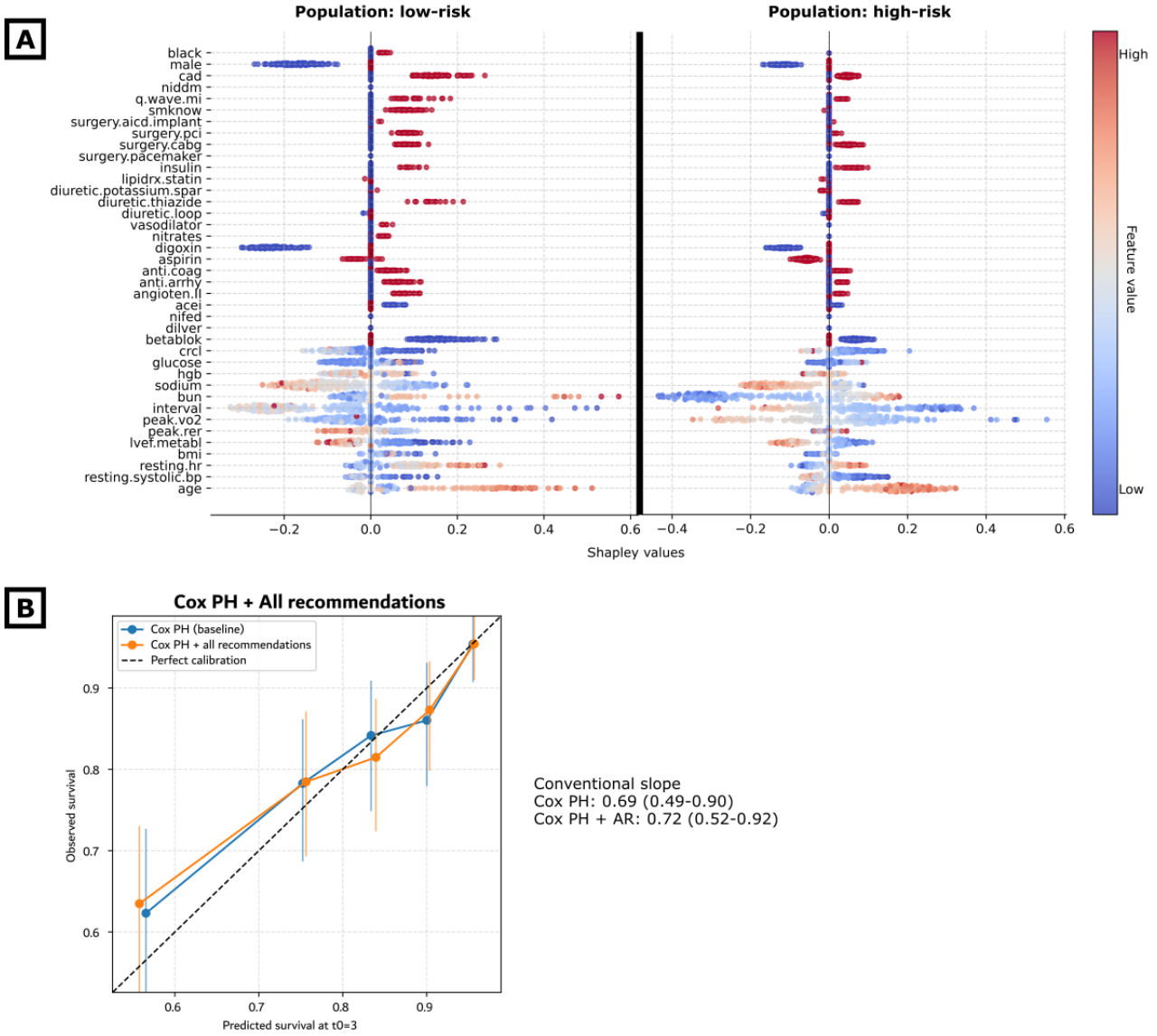
XSL•FO
RenderX

Figure 7. Evaluation of the Recommender on the ACT public dataset. (A) Shapley values obtained from the RSF model. (B) Calibration of the Cox proportional hazards model augmented with all recommendations. AR: all recommendations; PH: proportional hazards.



XSL•FO
RenderX

Figure 8. Evaluation of the recommender on the PeakVO2 public dataset. (A) Shapley values obtained from the RSF model. (B) Calibration of the Cox proportional hazards model augmented with all recommendations. AR: all recommendations; PH: proportional hazards.



Discussion

Principal Results

In this study, we developed and evaluated an exploratory AI recommender designed to support predictive study design in high-dimensional clinical datasets. Rather than using AI as an opaque final prediction model, the Recommender uses a flexible explainable-AI module as an exploratory engine to identify data-driven modeling insights that can then be translated into transparent statistical models. In a fall risk case study, the recommender recommended excluding 15 of 59 candidate features, adding nonlinear terms for 4 features, and adding 4 interactions from 946 candidate pairs. With all 409 available features, it recommended excluding 304 features, adding nonlinear terms for 4 features, and adding 2 interactions from 7021 candidate pairs. The augmented Cox model had a statistically significant gain in discrimination over the baseline model (C-index 0.719 vs 0.715), with 3 fewer parameters and similar calibration. With all 409 features, the gain was larger (0.722 vs 0.713), with about one-third of the parameters (129 vs 423), and the calibration slope moved closer to 1 (0.952 vs 0.846). The exploratory forest has similar discrimination as the baseline Cox model (0.716 vs 0.715), which indicates that the candidate features carry little nonlinear or interaction signal beyond their main effects. Large gains were therefore not expected. The main benefits of the Recommender are a simpler, transparent model and a short list of hypotheses for formal testing. Because AI is used only to explore the data, the final model keeps the auditability and transparency that are required in regulated clinical environments.

The Recommender's outputs agree with published evidence. Age and frailty are established predictors of falls and related injuries [41], and their associations with fall risks are known to be nonlinear [42]. The recommended interaction between age and stroke also agrees with a population-based study, in which the relative risk of hip fracture after stroke was higher in younger than in older patients [43]. Beyond confirming existing knowledge, our Recommender also identified some novel hypotheses that deserve further exploration. For example, with all 409 features, it recommended an interaction between

XSL•FO
RenderX

lipid-regulating drugs and drugs used in megaloblastic anemias. Statins alone are not consistently associated with falls in older adults [44], but they are associated with myopathy [45] and, less consistently, with peripheral neuropathy [46]. Vitamin B12 or folate deficiency can also cause neuropathy and impaired proprioception [47]. When these effects occur together, their combined effect on balance and gait may be larger than either effect alone. If this hypothesis is confirmed in other data, it may inform medication review in older adults who take both types of drugs. Crucially, the Recommender supports rather than replaces clinical judgment. The Recommender's transparent design allows researchers to critically evaluate such data-driven recommendations.

Comparison With Prior Work

This work extends previous research to predictive model designs in several ways. First, the Recommender is not simply a feature selection tool. Existing approaches, including robust variable selection and regularization-based methods for Cox models [48], provide rigorous strategies for deriving effective predictor sets, improving model stability, and reducing overfitting within prespecified modeling frameworks. However, these methods are primarily designed to determine which variables should be retained in a final model. In contrast, our Recommender combines a flexible nonlinear exploratory model with explainable AI to generate broader design-relevant insights, including candidate feature exclusions, potential nonlinear effects, interaction structures, and subgroup-specific predictor behavior. These outputs are intended not only to support model development, but also to help researchers understand why a feature may be informative, how its contribution may vary across patient subgroups, and which feature relationships may warrant formal statistical evaluation. In this sense, the Recommender complements rather than replaces established biostatistical methods by prioritizing clinically and statistically plausible signals from a large candidate space for targeted follow-up using conventional tools, including penalized Cox models, restricted cubic splines, interaction terms, and likelihood ratio tests. Closest to our work, Gosiewska et al [29] used a flexible surrogate model to derive feature transformations for an interpretable model, and Liu et al [30] used machine learning to identify new predictors that were then added to Cox models. However, each method addresses only one type of model change, and to our knowledge, neither screens for interactions or recommends feature exclusions. Our Recommender differs in proposing all 3 types of change from one exploratory model, with formal tests and a parameter limit, rather than adding or transforming features alone.

Second, we compared the Recommender with 2 standard approaches. Cox-RCS adds flexible nonlinear terms for all continuous features. In all datasets, the augmented model discriminated similarly to or better than Cox-RCS, with fewer parameters (eg, 70 vs 79 in DataLoch and 43 vs 78 in PeakVO2). With all 409 features, it had a higher C-index than Cox-RCS (difference 0.006, 95% CI 0.002-0.010) with 129 vs 429 parameters. Thus, the Recommender matched the discrimination of the more flexible Cox-RCS with a simpler model, which used squared terms instead of splines and fewer features. Coxnet selects terms from all main effects and all pairwise interactions by LASSO shrinkage. Its candidate set grows quadratically with the number of features (2624 terms with 59 features), so it could not be used with all 409 features. It also cannot separate nonlinear effects from interactions. In the public datasets, Coxnet retained few main effects but many interactions, so some of its interactions may approximate nonlinear main effects rather than reflecting true interactions. When the same LASSO penalty was applied to the recommended terms, the C-index was higher than that of Coxnet in all 3 datasets. The recommended terms may therefore be a better starting point for a penalized model than all pairwise interactions.

The public datasets also show how the Recommender behaves in smaller datasets with fewer events. In each dataset, the model before interactions already had more parameters than the sample-size criterion allowed, so the criterion correctly prevented additional interaction terms. The calibration slopes of all unpenalized Cox models in ACT and PeakVO2 were consistent with this risk of overfitting. In such small datasets, the recommended terms can be combined with penalization, which gave good calibration in all 3 datasets.

Third, the recommender provides a mechanism for exploring subgroup-specific predictor behavior that may be overlooked by conventional model development. Traditional models often emphasize high-risk populations, yet patients classified as low risk are not necessarily free from risk; rather, potential risk factors may be compensated by other protective factors [49]. Our subcohort analysis showed that nonsteroidal anti-inflammatory drugs (NSAIDs), although contributing minimally in the high-risk cohort, were a significant predictor specifically in the low-risk cohorts. Overall NSAID exposure was similar in patients with and without an FRI (6.5% vs 6.8%), so this pattern is not explained by a crude difference in prescribing. It may instead reflect how NSAIDs are prescribed or avoided in different groups of patients, that is, confounding by indication or contraindication. This provides a hypothesis for future causal studies of NSAID prescribing. Moreover, in the whole-cohort Cox model, NSAIDs were not associated with FRI (hazard ratio [HR] 1.06, 95% CI 0.96-1.17), which illustrates how a subgroup-specific pattern can be hidden in population-level estimates. The Recommender does not calculate fairness metrics, and we did not assess fairness. However, researchers could define strata by protected characteristics, such as sex, ethnicity, or socioeconomic deprivation, and compare FAs across them. By exposing these subgroup-specific deviations, the Recommender provides insights that might help researchers identify bias early in the study design.

Fourth, the recommender bridges the gap between AI power and clinical transparency. AI has demonstrated remarkable predictive capabilities with complex and large-scale biomedical data [50], but its opacity often prevents trust and clinical adoption [51]. By using AI models purely as an exploratory tool to inform conventional statistical models, we address the accuracy-interpretability trade-off without the fidelity issues of post hoc explainable AI [52].

Furthermore, our recommender demonstrates some utility in a "hard-to-improve" scenario, as the improvement in C-index

XSL•FO
RenderX

(0.004 with the 59 selected features and 0.009 with all 409 features) is statistically significant but numerically modest. Relatively simple models can perform reasonably well on high-quality, large-scale datasets, leading to great difficulty in achieving marginal gains. However, in large-scale population health, even marginal gains in discrimination and calibration can lead to significantly better resource allocation and risk stratification when applied to hundreds of thousands of patients. Our approach is scalable, as it rapidly explores thousands of pairwise interactions; efficient, as it avoids the computational demands of complex architectures like graph neural networks [53,54]; and adaptable, as it is applicable to different models (eg, DeepSurv [55] or diagnosis prediction) and clinical datasets. While SHAP is used here due to its theoretical foundations in game theory, the Recommender is attribution-agnostic. Researchers may consider other techniques like Integrated Gradients [56] or SurvSHAP(t) [57] for time-dependent explanations, depending on the computational budget and the specific exploratory model used.

The recommender offers practical implications for clinical research. It streamlines future data collection as it may reduce the number of variables that need to be collected without loss of predictive performance, can guide feature selection in further causal studies, and enables researchers to move beyond preconceived hypotheses [58]. More broadly, the Recommender encourages researchers to move beyond purely literature-driven or knowledge-based model specification by providing a systematic, data-driven mechanism for discovering important features, nonlinearities, interactions, and subgroup-specific effects. These insights may also inform subsequent causal studies, although they should not be interpreted as causal evidence without further causal design and validation.

The recommender is intended for analysts who develop prediction models, not for clinicians at the point of care. The released code automates all steps, from fitting the exploratory model and computing attributions to choosing the subcohort margin, applying the 3 recommendation rules, checking candidate interactions in a Cox model, and limiting the number of added parameters. The analyst only needs to specify the encoding of each feature (continuous, ordinal, nominal, or binary). Preselection of candidate features is optional, as the analysis of all 409 features shows, and the analyst can prespecify cut points for continuous features when a specific subgroup is of interest (eg, age 65 years according to a common older age threshold [59]). The analyst should review each recommendation for clinical plausibility before adopting it. In this study, all recommendations were applied as generated so that the method itself could be evaluated.

Limitations

This study has several limitations. First, our interaction terms were restricted to (already challenging) first-order pairs to ensure computational feasibility. Higher-order interactions were not explored, but are a potential focus for future research, where computational complexity is the main concern [60]. Second, the validity of the recommendations depends on the exploratory model's performance (eg, RSF). Suboptimal discrimination or calibration could introduce bias into FAs, necessitating careful model selection or ensemble approaches. Thus, we advise using the recommendations produced only with high-performing exploratory models (high C-index and stable calibration), and leave investigation into quality metrics of FAs to future work. Third, our analysis focused on large-scale populations rather than individual longitudinal trajectories, leaving temporal dependencies out of scope, though in principle applicable with extensions [57]. Fourth, we screened candidate interactions and checked them in a Cox model on the same training data, without a separate confirmation set. We chose this because a further split would reduce the events available to the exploratory forest and its subcohorts, and so reduce the power of the screen. Confirmation on separate data is a priority for future work in datasets with more events. Moreover, we adopted a random train/test split. The DataLoch extract has no practice identifier, so we could not test for clustering by practice. Patients from the same practice can be in both sets, and any practice-level effect on prescribing or care pathways would not be captured by this safeguard. A temporal split was also of limited use, because follow-up was 1 year and 78.7% of patients had the same index date (January 1, 2015). In our DataLoch analysis, death before an FRI was treated as a censoring event, so the model estimates the cause-specific hazard of FRI. Because competing deaths were uncommon (4040 patients, 1.83% of the cohort) and follow-up was limited to 1 year, we did not perform a Fine-Gray analysis. However, competing deaths mainly affect absolute risk estimates and are unlikely to change the comparison between the baseline and augmented models, which is the focus of this study. The cohort came from a single Scottish region, with few patients from ethnic minority groups (Table S4 in Multimedia Appendix 2) and higher fracture rates than some other United Kingdom regions [7]. The model, and the size of the gains from the recommender, may differ in other populations and health systems. Finally, we emphasize that FAs represent predictive associations, not causal relationships; they denote contributions to the model’s output rather than the clinical outcome itself. Any clinical interventions based on these findings would require subsequent causal discovery or randomized controlled trials.

Conclusions

We demonstrated that explainable AI can uncover complex feature relationships within high-dimensional clinical datasets. This data-driven exploratory approach showed the potential to improve the robustness of predictive cohort study design by augmenting traditional, expert-led methods. These findings also suggest that explainable AI-informed insights can generate novel hypotheses that may warrant further investigation.

Generative AI Use

Generative AI (Claude, Anthropic) was used to assist with code revision. All changes were reviewed by the authors. The authors take full responsibility for the published content.

XSL•FO
RenderX

Funding

We acknowledge the support of the UK Research and Innovation AI program and the Engineering and Physical Sciences Research Council (EPSRC) for the Causality in Healthcare AI Hub (grant number EP/Y028856/1). Junyu Yan is supported in part by the Advanced Care Research Centre through a PhD studentship. The funders had no role in study design, analysis, interpretation, or writing.

Data Availability

The data that support the findings of this study are available from DataLoch [40], but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. DataLoch data are available to researchers who complete governance training and obtain approvals. Access to DataLoch requires submission of a proposal, with investigator support and approval as needed. DataLoch data is stored in a National Health Service safe haven and can be accessed remotely. German Breast Cancer Study Group 2 (GBSG2), AIDS Clinical Trial (ACT), and PeakVO2 are publicly available [61,62]. The code to replicate the analysis on publicly available datasets (GBSG2, ACT, and peakVO2) is available on GitHub [63] and stored on Zenodo [64] (v1.0.0, commit cebb78c).

Authors' Contributions

Conceptualization: JY, DM, KB, PD
Data curation: JY, DM
Formal analysis: JY, DM, KB, PD
Investigation: JY, DM
Methodology: JY, DM, KB, PD
Software: JY, DM
Writing – original draft: JY, DM, KB, PD
Writing – review and editing: JY, DM, KB, PD, EMH, BG, SAT
Supervision: EMH, BG, SAT
Funding acquisition: SAT

Conflicts of Interest

None declared.

Multimedia Appendix 1

TRIPOD+AI checklist.
[DOCX File , 39 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Clinical code lists for falls and related injuries, long-term condition definitions, DataLoch cohort characteristics, model hyperparameters, Recommender outputs and interaction screening, simulation studies of the nonlinearity and interaction tests, the final augmented Cox model and its proportional hazards tests, and results in the 3 public datasets.
[DOCX File , 112 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Feature preprocessing, feature attribution with Shapley values, the 4-step feature interaction strategy, and selection of the subcohort margin.
[DOCX File , 113 KB-Multimedia Appendix 3]

References

1. Khan SS, Matsushita K, Sang Y, Ballew SH, Grams ME, Surapaneni A, et al. Chronic Kidney Disease Prognosis Consortium and the American Heart Association Cardiovascular-Kidney-Metabolic Science Advisory Group. Development and validation of the American Heart Association's PREVENT equations. Circulation. 2024;149(6):430-449. [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.123.067626] [Medline: 37947085]
2. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099. [FREE Full text] [doi: 10.1136/bmj.j2099] [Medline: 28536104]

XSL•FO
RenderX

3. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. BMJ. 2009;339:b4229. [FREE Full text] [doi: 10.1136/bmj.b4229] [Medline: 19926696]
4. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. FRAX and the assessment of fracture probability in men and women from the UK. Osteoporos Int. 2008;19(4):385-397. [FREE Full text] [doi: 10.1007/s00198-007-0543-5] [Medline: 18292978]
5. Lansbury LN, Roberts HC, Clift E, Herklots A, Robinson N, Sayer AA. Use of the electronic frailty index to identify vulnerable patients: a pilot study in primary care. Br J Gen Pract. 2017;67(664):e751-e756. [FREE Full text] [doi: 10.3399/bjgp17X693089] [Medline: 28947622]
6. Stevens JA, Sogolow ED. Gender differences for non-fatal unintentional fall related injuries among older adults. Inj Prev. 2005;11(2):115-119. [FREE Full text] [doi: 10.1136/ip.2004.005835] [Medline: 15805442]
7. Curtis EM, van der Velde R, Moon RJ, van den Bergh JPW, Geusens P, de Vries F, et al. Epidemiology of fractures in the United Kingdom 1988-2012: variation with age, sex, geography, ethnicity and socioeconomic status. Bone. 2016;87:19-26. [FREE Full text] [doi: 10.1016/j.bone.2016.03.006] [Medline: 26968752]
8. Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: a systematic review of methods. Comput Methods Programs Biomed. 2018;164:49-64. [doi: 10.1016/j.cmpb.2018.06.006] [Medline: 30195431]
9. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. J Thorac Dis. 2019;11(Suppl 4):S574-S584. [FREE Full text] [doi: 10.21037/jtd.2019.01.25] [Medline: 31032076]
10. Matsuo K, Purushotham S, Jiang B, Mandelbaum RS, Takiuchi T, Liu Y, et al. Survival outcome prediction in cervical cancer: cox models vs deep-learning model. Am J Obstet Gynecol. 2019;220(4):381.e1-381.e14. [FREE Full text] [doi: 10.1016/j.ajog.2018.12.030] [Medline: 30582927]
11. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
12. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236-1246. [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
13. Carabantes M. Black-box artificial intelligence: an epistemological and critical analysis. AI & Soc. 2019;35(2):309-317. [doi: 10.1007/s00146-019-00888-w]
14. van Royen FS, Weerts HJP, de Hond AAH, Geersing G, Rutten FH, Moons KGM, et al. In humble defense of unexplainable black box prediction models in healthcare. J Clin Epidemiol. 2026;189:112013. [FREE Full text] [doi: 10.1016/j.jclinepi.2025.112013] [Medline: 41077324]
15. Carriero A, de Hond A, Cappers B, Paulovich F, Abeln S, Moons KG, et al. Explainable AI in healthcare: to explain, to predict, or to describe? Diagn Progn Res. 2025;9(1):29. [doi: 10.1186/s41512-025-00213-8] [Medline: 41345739]
16. Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Hum Factors. 2024;11:e48633. [FREE Full text] [doi: 10.2196/48633] [Medline: 39207831]
17. Varghese J. Artificial intelligence in medicine: chances and challenges for wide clinical adoption. Visc Med. 2020;36(6):443-449. [FREE Full text] [doi: 10.1159/000511930] [Medline: 33442551]
18. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon. 2024;10(4):e26297. [FREE Full text] [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
19. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput. 2023;16(1):45-74. [doi: 10.1007/s12559-023-10179-8]
20. Winter E. Chapter 53 the shapley value. In: Handbook of Game Theory with Economic Application. Amsterdam. North-Holland; 2002:30.
21. Ortigossa ES, Gonçalves T, Nonato LG. EXplainable artificial intelligence (XAI)—from theory to methods and applications. IEEE Access. 2024;12:80799-80846. [doi: 10.1109/access.2024.3409843]
22. Mridha K, Ghimire S, Shin J, Aran A, Uddin MM, Mridha MF. Automated stroke prediction using machine learning: an explainable and exploratory study with a web application for early intervention. IEEE Access. 2023;11:52288-52308. [doi: 10.1109/access.2023.3278273]
23. Ahmed ZU, Sun K, Shelly M, Mu L. Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA. Sci Rep. 2021;11(1):24090. [FREE Full text] [doi: 10.1038/s41598-021-03198-8] [Medline: 34916529]
24. Klauschen F, Dippel J, Keyl P, Jurmeister P, Bockmayr M, Mock A, et al. Toward explainable artificial intelligence for precision pathology. Annu Rev Pathol. 2024;19:541-570. [FREE Full text] [doi: 10.1146/annurev-pathmechdis-051222-113147] [Medline: 37871132]
25. Temkov S. Explainable AI for discovering disease biomarkers: a survey. Springer Nature Switzerland; 2024. Presented at: 7th EAI International Conference on Robotic Sensor Networks; December 15-16, 2024:185-192; Istanbul, Türkiye. [doi: 10.1007/978-3-031-64495-5_14]
26. Yagin FH, Aygun U, Algarni A, Colak C, Al-Hashem F, Ardigò LP. Platelet metabolites as candidate biomarkers in sepsis diagnosis and management using the proposed explainable artificial intelligence approach. J Clin Med. 2024;13(17):5002. [FREE Full text] [doi: 10.3390/jcm13175002] [Medline: 39274215]

XSL•FO
RenderX

27. Burugadda VR, Dutt V, Mamta, Vyas N. Personalized cardiovascular disease risk prediction using random forest: an optimized approach. 2023. Presented at: Proceedings of the 2023 IEEE World Conference on Applied Intelligence and Computing, AIC 2023; July 29-30, 2023; Varanasi, India. [doi: 10.1109/AIC57670.2023.10263915]
28. Khani M, Luo J, Assadi Shalmani M, Taleban A, Adams J, Friedland DR. Advancing personalized healthcare: leveraging explainable AI for BPPV risk assessment. Health Inf Sci Syst. 2025;13(1):1. [doi: 10.1007/s13755-024-00317-3] [Medline: 39606094]
29. Gosiewska A, Kozak A, Biecek P. Simpler is better: lifting interpretability-performance trade-off via automated feature engineering. Decis Support Syst North-Holland. 2021;150:113556. [doi: 10.1016/j.dss.2021.113556]
30. Liu X, Morelli D, Littlejohns TJ, Clifton DA, Clifton L. Combining machine learning with cox models to identify predictors for incident post-menopausal breast cancer in the UK Biobank. Sci Rep. 2023;13(1):9221. [FREE Full text] [doi: 10.1038/s41598-023-36214-0] [Medline: 37286615]
31. Blayney MC, Reed MJ, Masterson JA, Anand A, Bouamrane MM, Fleuriot J, et al. Multimorbidity and adverse outcomes following emergency department attendance: population based cohort study. BMJ Med. 2024;3(1):e000731. [FREE Full text] [doi: 10.1136/bmjmed-2023-000731] [Medline: 39184567]
32. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. [FREE Full text] [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
33. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. J Clin Epidemiol. 2020;125:183-187. [doi: 10.1016/j.jclinepi.2020.03.028] [Medline: 32540389]
34. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3):841-860. [doi: 10.1214/08-AOAS169]
35. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
36. Sauerbrei W, Bastert G, Bojar H, Beyerle C, Neumann RLA, Schmoor C, et al. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients: an update based on 10 years' follow-up. German breast cancer study group. J Clin Oncol. 2000;18(1):94-101. [doi: 10.1200/JCO.2000.18.1.94] [Medline: 10623698]
37. Hosmer DW, Lemeshow S, May S. Applied Survival Analysis: Regression Modeling of Time to Event Data. 2nd ed. New Jersey. Wiley; 2011. [doi: 10.1002/9780470258019]
38. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circ Cardiovasc Qual Outcomes. 2011;4(1):39-45. [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.110.939371] [Medline: 21098782]
39. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30(10):1105-1117. [FREE Full text] [doi: 10.1002/sim.4154] [Medline: 21484848]
40. DataLoch. DataLoch. URL: https://dataloch.org/ [accessed 2026-10-22]
41. Yang Z, Lin H, Jiang G, Chu Y, Gao J, Tong Z, et al. Frailty is a risk factor for falls in the older adults: a systematic review and meta-analysis. J Nutr Health Aging. 2023;27(6):487-495. [FREE Full text] [doi: 10.1007/s12603-023-1935-8] [Medline: 37357334]
42. El Haber N, Erbas B, Hill KD, Wark JD. Relationship between age and measures of balance, strength and gait: linear and non-linear analyses. Clin Sci (Lond). 2008;114(12):719-727. [doi: 10.1042/CS20070301] [Medline: 18092948]
43. Pouwels S, Lalmohamed A, Thio B, Leufkens H, Cooper C, van Staa T, et al. Risk of hip fracture after stroke: a population-based case-control study. Bone. 2009;44:S416-S417. [doi: 10.1016/j.bone.2009.03.342]
44. de Vries M, Seppala L, Daams J, van de Glind EMM, Masud T, van der Velde N, et al. EUGMS Task and Finish Group on Fall-Risk-Increasing Drugs. Fall-risk-increasing drugs: a systematic review and meta-analysis: I. cardiovascular drugs. J Am Med Dir Assoc. Apr 2018;19(4):371.e1-371.e9. [doi: 10.1016/j.jamda.2017.12.013] [Medline: 29396189]
45. Stroes ES, Thompson PD, Corsini A, Vladutiu GD, Raal FJ, Ray KK, et al. European Atherosclerosis Society Consensus Panel. Statin-associated muscle symptoms: impact on statin therapy-European atherosclerosis society consensus panel statement on assessment, aetiology and management. Eur Heart J. 2015;36(17):1012-1022. [FREE Full text] [doi: 10.1093/eurheartj/ehv043] [Medline: 25694464]
46. Gaist D, Jeppesen U, Andersen M, García Rodríguez LA, Hallas J, Sindrup SH. Statins and risk of polyneuropathy: a case-control study. Neurology. 2002;58(9):1333-1337. [doi: 10.1212/wnl.58.9.1333] [Medline: 12011277]
47. Reynolds E. Vitamin B12, folic acid, and the nervous system. Lancet Neurol. 2006;5(11):949-960. [doi: 10.1016/S1474-4422(06)70598-1] [Medline: 17052662]
48. Zhang Y, Muller S. Robust variable selection methods with Cox model-a selective practical benchmark study. Brief Bioinform. 2024;25(6):bbae508. [FREE Full text] [doi: 10.1093/bib/bbae508] [Medline: 39400113]
49. Montero-Odasso M, Pieruccini-Faria F, Son S, Carvalho de Abreu DC, Hunter S, Liu JQ, et al. Fall risk stratification in older adults: low and not-at-risk status still associated with falls and injuries. Age Ageing. 2025;54(3):afaf064. [FREE Full text] [doi: 10.1093/ageing/afaf064] [Medline: 40139220]

XSL•FO
RenderX

50. Huang Y, Li J, Li M, Aparasu RR. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. BMC Med Res Methodol. 2023;23(1):268. [FREE Full text] [doi: 10.1186/s12874-023-02078-1] [Medline: 37957593]
51. Bektaş AB, Gönen M. Machine learning for medicine must be interpretable, shareable, reproducible and accountable by design. arXiv. Aug 22, 2025:1-10. [FREE Full text]
52. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-215. [FREE Full text] [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
53. Kwak H, Lee M, Yoon S, Chang J, Park S, Jung K. Drug-disease graph: predicting adverse drug reaction signals via graph neural network with clinical data. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham. Springer; 2020:633-644. [doi: 10.1007/978-3-030-47436-2_48]
54. Al-Rabeah MH, Lakizadeh A. Prediction of drug-drug interaction events using graph neural networks based feature extraction. Sci Rep. 2022;12(1):15590. [FREE Full text] [doi: 10.1038/s41598-022-19999-4] [Medline: 36114278]
55. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18(1):24. [FREE Full text] [doi: 10.1186/s12874-018-0482-1] [Medline: 29482517]
56. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017. Presented at: ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70; August 6-11, 2017:5109-5118; Sydney NSW Australia. URL: https://proceedings.mlr.press/v70/sundararajan17a.html
57. Krzyziński M, Spytek M, Baniecki H, Biecek P. SurvSHAP(t): time-dependent explanations of machine learning survival models. Knowledge-Based Systems. 2023;262:110234. [doi: 10.1016/j.knosys.2022.110234]
58. Kuzma JW. Basic Statistics for the Health Sciences. California, United States. Mayfield Publishing Company; 1992.
59. Singh S, Bajorek B. Defining 'elderly' in clinical practice guidelines for pharmacotherapy. Pharm Pract (Granada). 2014;12(4):489. [FREE Full text] [doi: 10.4321/s1886-36552014000400007] [Medline: 25580172]
60. Butler K, Feng G, Djurić P. Higher-order feature attribution: bridging statistics, explainable AI, and topological signal processing. 2026. Presented at: Proceedings of the ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE; May 03-08, 2026:5261-5265; Barcelona, Spain. [doi: 10.1109/icassp55912.2026.11461829]
61. Datasets — scikit-survival 0.28.0. scikit-survival. URL: https://scikit-survival.readthedocs.io/en/stable/api/datasets.html [accessed 2026-10-22]
62. Systolic heart failure data. Fast Unified Random Forests with randomForestSRC. URL: https://www.randomforestsrc.org/reference/peakVO2.html [accessed 2026-10-22]
63. CHAI-UK/XAI-insight-discovery. GitHub. URL: https://github.com/CHAI-UK/XAI-insight-discovery/ [accessed 2026-10-22]
64. Explainable AI to guide the design of interpretable prediction models from high-dimensional data, applied to falls and related injuries in adults aged 50 years or older: algorithm development and validation. Zenodo. URL: https://doi.org/10.5281/zenodo.23216891 [accessed 2026-10-22]

Abbreviations

ACT: AIDS Clinical Trial
C-index: concordance index
CPH: Cox proportional hazards
eFI: electronic Frailty Index
FA: feature attribution
FDR: false discovery rate
FRI: falls or related injuries
GBSG2: German Breast Cancer Study Group 2
GP: general medical practice
HR: hazard ratio
IBS: integrated Brier score
***ICD-10*:** *International Statistical Classification of Diseases and Related Health Problems 10th Revision*
LASSO: least absolute shrinkage and selection operator
NSAID: nonsteroidal anti-inflammatory drug
RSF: random survival forest
SHAP: Shapley Additive Explanations
SIMD: Scottish Index of Multiple Deprivation
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

XSL•FO
RenderX

Edited by I Steenstra; submitted 23.Jun.2026; peer-reviewed by S Bhetwal, T Xiao; comments to author 30.Jul.2026; revised version received 07.Oct.2026; accepted 13.Oct.2026; published 12.Nov.2026

Please cite as:
Yan J, Machlanski D, Butler K, Dimitrakopoulos P, Harrison EM, Guthrie B, Tsaftaris SA
Explainable AI to Guide the Design of Interpretable Prediction Models From High-Dimensional Data, Applied to Falls and Related Injuries in Adults Aged 50 Years or Older: Algorithm Development and Validation
J Med Internet Res 2026;28:e105364
URL: https://www.jmir.org/2026/1/e105364
doi: 10.2196/105364
PMID:

©Junyu Yan, Damian Machlanski, Kurt Butler, Panagiotis Dimitrakopoulos, Ewen M Harrison, Bruce Guthrie, Sotirios A Tsaftaris. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX