Title

Machine learning versus regression for prediction of sporadic pancreatic cancer.

Journal

Pancreatology, 23(4)

Authors

Chen, Wansu
Zhou, Botao
Xie, Fagen
et al.

Publication Date

2023-06-01

DOI

10.1016/j.pan.2023.04.009

Peer reviewed

HHS Public Access

Author manuscript

Pancreatology. Author manuscript; available in PMC 2024 June 01.

Published in final edited form as:

Pancreatology. 2023 June ; 23(4): 396–402. doi:10.1016/j.pan.2023.04.009.

Machine Learning versus Regression for Prediction of Sporadic Pancreatic Cancer

Wansu Chen, PhD[1], **Botao Zhou, MS**[1], **Christie Y. Jeon, ScD**[2], **Fagen Xie, PhD**[1], **Yu-Chen Lin, MPH**[2], **Rebecca K. Butler, ScM**[1], **Yichen Zhou, MS**[1], **Tiffany Q. Luong, MPH**[1], **Eva Lustigova, MPH**[1], **Joseph R. Pisegna, MD**[3], **Bechien U. Wu, MD, MPH**[4]

[1]Kaiser Permanente Southern California Research and Evaluation, Pasadena, CA

[2]Cedars-Sinai Medical Center, Los Angeles, CA

[3]Division of Gastroenterology and Hepatology, VA Greater Los Angeles Healthcare System, Los Angeles, CA and Departments of Medicine and Human Genetics David Geffen School of Medicine at UCLA

[4]Center for Pancreatic Care, Department of Gastroenterology, Los Angeles Medical Center, Southern California Permanente Medical Group, Los Angeles, CA

Abstract

BACKGROUND/OBJECTIVES: There is currently no widely accepted approach to identify patients at increased risk for **sporadic** pancreatic cancer (PC). We aimed to compare the performance of two machine-learning models with a regression-based model in predicting pancreatic ductal adenocarcinoma (PDAC), the most common form of PC.

METHODS: This retrospective cohort study consisted of patients 50–84 years of age enrolled in either Kaiser Permanente Southern California (KPSC, model training, internal validation) or the Veterans Affairs (VA, external testing) between 2008–2017. The performance of random survival forests (RSF) and eXtreme gradient boosting (XGB) models were compared to that of COX proportional hazards regression (COX). Heterogeneity of the three models were assessed.

Correspondence: Wansu Chen, Ph.D., Department of Research and Evaluation, Kaiser Permanente Southern California, 100 S Los Robles, 2nd Floor, Pasadena, CA 91101, Wansu.Chen@KP.org.

Author contributions: Wansu Chen: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision; Botao Zhou: Methodology, Software, Validation, Formal analysis, Investigation, Writing - Review & Editing, Visualization; Christie Jeon: Conceptualization, Validation, Investigation, Writing - Review & Editing, Supervision; Fagen Xie: Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Review & Editing; Yu-Chen Lin: Software, Validation, Investigation, Data Curation, Writing - Review & Editing; Rebecca Butler: Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Review & Editing; Yichen Zhou: Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Review & Editing, Visualization; Tiffany Luong: Writing - Review & Editing, Project administration; Eva Lustigova: Writing - Review & Editing, Supervision, Project administration; Joseph Pisegna: Writing - Review & Editing; Bechien Wu: Conceptualization, Methodology, Validation, Resources, Writing - Review & Editing, Supervision, Funding acquisition. All authors have approved the final submitted draft.

Disclosure Statement: The authors declare they have no conflict of interest for this study.

Ethics approval and consent to participate: The study was approved by KPSC IRB. IRB approved the request to waive the documentation of informed consent.

RESULTS: The KPSC and the VA cohorts consisted of 1.8 and 2.7 million patients with 1,792 and 4,582 incident PDAC cases within 18 months, respectively. Predictors selected into all three models included age, abdominal pain, weight change, and glycated hemoglobin (A1c). Additionally, RSF selected change in alanine transaminase (ALT), whereas the XGB and COX selected the rate of change in ALT. The COX model appeared to have lower AUC (KPSC: 0.737, 95% CI 0.710–0.764; VA: 0.706, 0.699–0.714), compared to those of RSF (KPSC: 0.767, 0.744–0.791; VA: 0.731, 0.724–0.739) and XGB (KPSC: 0.779, 0.755–0.802; VA: 0.742, 0.735–0.750). Among patients with top 5% predicted risk from all three models (N=29,663), 117 developed PDAC, of which RSF, XGB and COX captured 84 (9 unique), 87 (4 unique), 87 (19 unique) cases, respectively.

CONCLUSIONS: The three models complement each other, but each has unique contributions.

Keywords

risk prediction; pancreatic cancer; machine learning versus regression; random survival forest; eXtreme gradient boosting

BACKGROUND

Pancreatic cancer is a relatively uncommon but lethal cancer type, with an incidence rate of 13.3 per 100,000 people per year.[1] Survival in pancreatic cancer remains poor, with only a 11.5% 5-year survival after diagnosis.[1] Pancreatic ductal adenocarcinoma (PDAC) is the most prevalent and lethal neoplastic disease of the pancreas, accounting for over 90% of all pancreatic cancer cases.[2] Because of its low incidence rate, the United States Preventative Services Task Force (USPTF) does not recommend population-based screening for pancreatic cancer.[3] Therefore, additional tools are needed to identify high-risk patients to facilitate early detection to impact the natural history of this disease. While progress has been made related to approaches for screening individuals based on genetic or family history,[4] risk prediction models based on electronic health records have the potential to provide important adjunctive risk stratification tools applicable to a broader range of the general population.

Our research team developed a risk prediction model based on the EHR of patients 50–84 years of age who had at least one clinic-based visit in the past 12 months.[5] The risk prediction model demonstrated good performance via internal validation and external testing. In the previous study, the approach we adopted was Random Survival Forest (RSF), one of the mature machine learning approaches for analyzing time-to-event outcomes.[5] Compared to the Cox proportional hazards regression model (COX), the most popular regression-based model for predicting time-to-event outcomes, RSF has perceived advantages due to its ability to handle non-linear effects and interactions among predictors. Nevertheless, its superiority over the COX model in performance was shown when it was applied to predict cardiovascular diseases[6] and suicidal behaviors,[7] but not for the prediction of gastrointestinal bleeding,[8] breast cancer,[9] oral and pharyngeal cancer survival,[10] head and neck cancer survival[11] and overall survival.[12] Furthermore, the implementation of RSF-based risk prediction models in clinical operation is more time-consuming, resource intensive and challenging in various ways compared to a COX model. Therefore, comparing

RSF and COX models in the contexts of risk prediction of pancreatic cancer, a very rare disease, could be insightful. In this analysis, we also choose another well-developed machine learning-based approach eXtreme gradient boosting (XGB) to compare with RSF and COX modes due to its good performance in previously published clinical models.[8, 13] The purpose of this study is not only to compare the performance of RSF, XGB and COX for pancreatic cancer prediction, but also to examine the extent of overlap with respect to individually predicted risks based on the three modeling approaches. The results will provide insights into the advantages and disadvantages of the three methods for the purpose of predicting rare time-to-event outcomes including pancreatic cancer.

METHODS

Study design and setting

This is a retrospective cohort study involving two large health care organizations. Kaiser Permanente Southern California (KPSC) is an integrated healthcare system that provides comprehensive healthcare services for 4.8 million enrollees across 15 medical centers and 235 medical offices in Southern California. The Veterans Affairs (VA) is America's largest integrated health care system, providing care to 9 million enrollees every year at its 1,298 health care facilities including 171 medical centers and 1,113 medical offices nationwide. The study protocol was approved by the KPSC and VA Institutional Review Boards.

Sources of data

The data were extracted from KPSC Research Data Warehouse and Clarity, the repository of HealthConnect, the EHR system of KPSC, and VA's Corporate Data Warehouse (CDW), a repository derived from the Veterans Health Administration (VHA)'s electronic medical records system called Computerized Patient record System (CPRS)/VistA system.[14]

Participants

The study participants were patients between 50–84 years of age with 1+ clinic-based visit (index visit) within a KPSC or VA facility in 2008–2017. In addition, KPSC patients were required to have continuous enrollment in the KPSC health plan in the past 12 months (gaps 45 days or less were allowed) prior to the index visit, and the VA participants were required to have another clinic-based visit within the 12 months prior to the index date given the open-ended nature of the VA health coverage system. Patients who had history of pancreatic cancer prior to the index date were excluded. For patients with multiple qualifying index visits, we selected one randomly as the index visit (to serve as the date of risk assessment). The date corresponding to the index visit was referred to as the index date (t_0).

Follow-up

Follow-up started on t_0 and ended with the earliest of the following events: (1) disenrollment from the health plan (applicable to KPSC patients only), (2) end of the study (December 31, 2018), (3) reached the maximum length of follow-up (18 months), (4) non-PDAC related death, or (5) PDAC diagnosis or death due to PDAC (outcome). A minimum of 30 days of follow-up was required.

Outcome

The study outcome was PDAC diagnosis or death with pancreatic cancer in the 18 months after the index date. For the KPSC cohort, PDAC was identified from the Cancer Registry by using the Tenth Revision of International Classification of Diseases, Clinical Modification (ICD-10-CM) code C25.x and histology codes (eTable 1). Pancreatic cancer deaths were derived from the linkage with the California State Death Master files and identified using ICD-10-CM codes C25.x.[15] For the VA cohort, cases of PDAC were similarly identified through an internal VA Central Cancer Registry, and PDAC deaths identified through the VA Mortality Data Repository, which integrates vital status data from the National Death Index (NDI), VA, and DoD administrative records.

Predictors

A list of 500+ candidate features (eTable 2) were extracted in the original study. The comprehensive list includes patient demographics, diagnosis, medical procedures, dispensed medications, lab results, potential symptoms of pancreatic cancer and medical utilizations. From the candidate feature pool, twenty-nine (eTable 3) were pre-selected by using Random Survival Forest (RSF). Missing values were imputed[16] if the frequency of missing was <60%. We used predictive mean matching method[17] with k=5. Laboratory measures with ≥60% missingness or change/change rate measures with ≥80% missingness were not included in the model development process. Nine imputed datasets were generated at KPSC and 10 were created at VA (eFigure 1).

Statistical Analysis

Description of modeling approaches—RSF is an ensemble learning method for analyzing right censored time-to-event data.[18, 19] RSF approach uses bootstrapping and random node splitting to grow a group of decision trees. The results are averaged across all the trees to determine the relative importance of each variable based on the average minimal depth (average distance from the root node to the node where the variable first splits). The lower the average minimal depth values, the more important the variable. RSF can be computed using the R-package randomForestSRC.[20]

COX is a semi-parametric model commonly used to model the time it takes for an outcome to occur.[21, 22] It is often used to study the relationships between predictors and a time-to-event outcome,[21] and is more and more frequently utilized to build risk prediction models[23, 24] due to its easier model training and implementation process compared to machine-learning based models (e.g., RSF and XBG introduced below). As a regression-based model, Cox proportional hazards regression relies on the assumptions of log-linearity and proportional hazards across different covariates.

XGB is a scalable and flexible tree boosting system in which users can define various objective functions.[25] The parallel and distributed computing makes learning much faster compared to other tree-based models.[25] For the current study, the R-package XGBoost was applied with “survival: COX” (Cox proportional hazards regression) selected as the objective function and “COX-nloglik” (negative partial log-likelihood for Cox proportional hazards regression) specified as the evaluation metric.[26]

Training, validation, and testing datasets—For RSF and XGB models, the process to create the 45 training and internal validation datasets and the 10 externally testing datasets is displayed in eFigure 1. For Cox proportional hazards regression model, we randomly selected 80% of patients from the first imputation dataset to form the training dataset. The rest (20%) served as the corresponding internal validation dataset. The exact same patients (based on patient IDs) in the 2nd to 9th imputation datasets were identified for training and internal validation, respectively, forming a total of 9 training and 9 internal validation datasets. All 10 imputed datasets at VA were used for external testing.

Model training—For all three methods described above, age was forced into each model. Preselected features were added incrementally to identify the feature that yielded the maximum improvement of c-index. This process continued until the c-index increased is <0.005 when a new feature is added. For RSF and XGB, out of the 45 models derived from the 45 training/internal validation datasets (9 imputation datasets x 5-fold cross validation), the model that appeared most frequently was selected as the final model. To train the Cox proportional hazards regression model, we appended the 9 imputed datasets into one single stacked dataset.[27] To account for the multiple observations for each subject, weights were applied according to Wood et al. (weight=1/M, where M, the number of imputed datasets=9).[27] All the possible interaction terms and quadratic terms (for continuous variables only) of the 29 pre-selected features were included in the training process.

Model validation and testing

Discrimination: The discriminative power for each final model was evaluated by c-index. When the analyses were limited to patients with complete follow up or developed PDAC within 18 months, area under the receiver operator curve (AUC), sensitivity, specificity, positive predictive value (PPV), and relative increase in risk in comparison to that of the entire cohort at various levels of risk thresholds were also calculated for each final model. The results were averaged across the internal validation and external testing datasets. Area under the ROC curves (AUROC) were also plotted for each model.

Calibration: Calibration was assessed by calibration plots with five risk groups (<50th, 50–74th, 75–89th, 90–94th, 95–100th percentiles). Greenwood-Nam-D'Agostino (GND) calibration test was also performed to assess goodness-of-fit.[28]

Heterogeneity in high-risk patient identification: To understand how the three risk prediction models complement each other in patients who are likely to benefit from cancer screening, we plotted Venn diagrams of all patients and cancer patients in those with the highest risks (top 5 percentiles and top 1 percentile, respectively). To visualize and describe the comparative distributions of the risks predicted by the three models, overlapping histograms were plotted.

Predicted risks for high-risk patients: To understand the absolute values of the risks predicted by each model for high-risk patients, the predicted risks were plotted using histogram.

In the previous study, we showed that the performance of the RSF model could be improved when the model was recalibrated in the testing datasets for the VA population.[5] In the current study, only the recalibrated results are presented.

Linear vs. quadratic regression models: For the COX model, we developed a separate model without including the quadratic (high order) terms. The developed models were applied to the internal validation datasets.

Feature contribution: To understand the relative importance of the PDAC predictors being selected into the final models, we used SurvSHAP(t)[29] implemented in R-package 'survex'[30] to explain the impact of individual features on model's prediction.

RESULTS

The numbers of eligible patients at KPSC (training and internal validation) and VA (testing) were 1.8 million and 2.7 million, respectively (eFigure 2). Out of the 1.8 million records of KPSC, 1.441,546 and 360,386 patients were utilized for training and internal validation, respectively. The incidence rates of PDAC were 0.77 and 1.11/1,000 person-years of follow-up at KPSC and VA, respectively. Patient demographics and clinical characteristics including time since health plan enrollment for both KPSC and VA patients are presented in eTable 4a and eTable 4b.

Final model derived from each modeling method

Out of the 29 pre-selected potential predictors (eTable 3), the RSF method selected the model containing age, abdominal pain, weight change, alanine transaminase (ALT) change and HbA1c as the final model (Table 1). For the final model identified by XGB, all the predictors remained the same as that of RSF model except that ALT change in one year was replaced by ALT change rate in one year. The COX model selected the exact same predictors identified by XGB model; however, it included quadratic terms of the following continuous predictors: ALT rate change, HbA1c and weight change. The size of training, internal validation samples and external testing samples are represented in eTable 5. The hyperparameters used for feature selection for each of the three models can be found in eTable 6.

Performance

Discrimination: The c-indexes of the RSF model (mean 0.77, SD 0.02) and XGB model (mean 0.78, SD 0.01) based on the KPSC internal validation data were comparable (Table 1). The c-index (mean 0.74, SD 0.01) for the COX model seemed to be slightly lower compared to those of RSF and XGB models. The mean c-indexes based on the 10 testing datasets at VA were 0.71 (SD 0.002), 0.74 (SD 0.008), and 0.70 (SD 0.005), respectively, for RSF, XGB and COX models (Table 1). For KPSC internal validation datasets, the COX model appeared to have lower AUC (0.737, 95% CI 0.710–0.764), compared to those of RSF (0.767, 95% CI 0.744–0.791) and XGB (0.779, 95% CI 0.755–0.802), although the 95% confidence intervals overlap. For the VA external testing datasets, the AUC of the COX model was 0.706 (95% CI 0.699–0.714) compared to those of RSF (0.731, 95% CI

0.724–0.739) and XGB (0.742, 95% CI 0.735–0.750). The advantage of the discriminative power of the two machine learning models (RSF, XGB) over the COX model depends on risk threshold. The AUROC curves for the three models are plotted in Figure 1.

Calibration: Calibration test of the three models based on the KPSC validations datasets revealed that the XGB model was poorly calibrated (Table 1, mean p-value of calibration test <0.01), while the RSF model and the COX model had reasonable calibration (Table 1, mean p-value of calibration test =0.4 and 0.3, for RSF and COX respectively). For both KPSC internal validation dataset and the VA testing dataset, the XGB model tended to overestimate the risks of PDAC, especially in the highest risk group (Figure 2).

Sensitivity: For KPSC, the sensitivity of the COX model was compromised only when the risk threshold was at the top 20^{th} or top 15^{th} percentile compared to that of RSF or XGB model (Table 2a). However, For VA, the sensitivity of the COX model was lower compared to that of RSF or XGB model for all risk thresholds.

PPV and fold increase in risk: The PPV of the highest risk identified by each model (top 2.5 percentile) were similar (1.1% for RSF, 1.2% for XGB, and 1.2 for COX). PPVs were higher for the VA enrollees, despite lower AUC compared to those of KPSC, as baseline risk of PDAC were higher at the VA (Table 2b). The top 2.5 percentile of risk predicted by all three models identified patients whose predicted PDAC risk exceeded 1% over 18 months in both the internal validation and external testing datasets. Between RSF and XGB, the performance measures were either the same or slightly favored XGB (Table 2).

Heterogeneity in high-risk patient identification: Figure 3 demonstrates the heterogeneity across the three models when they are used to identify high risk patients at KPSC. If we are able to screen the top 5% of patients (n=29,663) with the highest risk of developing PDAC identified any of the three models based on the KPSC validation dataset, only 8,049 patients will be consistently captured by all three models (Figure 3). Patients uniquely identified by RSF, XGB and COX were 4,466, 2,459 and 6,369, respectively. Out of the 117 PDAC cases developed in the 29,663 screened patients, 56 were identified by all three models (Figure 3). The COX model uniquely captured 19 PDAC. In other words, if both RSF and XGB are used to identify high-risk patients without the COX model, 19 true cases would have been missed. Similarly, RSF and XGB would uniquely add 9 and 4 cases, respectively.

Predicted risks for high-risk patients: eFigure 3 revealed that XGB tended to have higher predicted risks for the top three risk groups, compared to RSF and COX. The average observed risks based on XGB seemed to be slightly higher compared to those of RSF and COX for these three risk groups.

Linear vs. quadratic regression models: When the quadratic terms were dropped in the COX model, the c-index on based on the KPSC validation datasets dropped from 0.74 to 0.72 (SD 0.002).

Feature contribution: The feature importance is shown in eFigure 4. The features are ranked based on their relative influences on model's prediction. For all three models, age was the most important feature. Abdominal pain was the least important feature in both RSF and XGB models; however, its contribution in the COX model was bigger compared to ALT change and weight change.

Discussion

In this study, we compared the performance of two machine learning models with Cox proportional hazards regression model for prediction of pancreatic cancer, an uncommon but frequently lethal cancer. The two machine learning models appeared to have better discrimination compared to the COX model when measured by c-index or AUC; however, the discriminative powers are comparable among all three models when focused on the very high risk (patients who may be targeted for screening). An interesting finding of the study is the complementary nature of the three prediction methods. Each of the three models can identify unique patients to screen and identify unique PDAC cases due to the differences in methodology. Using a combination of models can potentially increase the sensitivity of a screening strategy.

Many previous studies comparing machine-learning based models to parametric models in cancer incidence or outcomes have found similar performance between the models. For example, Omurlu et al. have compared RSF with Cox proportional hazards regression model with simulation and a real data application and suggested that all the methods were almost similar.[9] Hazewinkel et al. compared RSF, survival neural networks (SNN) and COX based on the data collect from a group of cancer patients, and concluded that all three methods were similar in terms of performance measured by C-indices, Brier and KL scores.[12] Meanwhile, there are other studies favoring machine learning-based models.[6, 7] For instance, Miao et al. found that RSF improved discrimination performance greatly than Cox with an out-of-bag C-statistics of 0.812 comparing to 0.736 for the Cox based model.[6] A data simulation study conducted by Baralou et al. revealed that under the presence of interactions, RSF performed better than Cox-PH as the number of events increased.[31] The current study found a difference of 0.03–0.04 in c-index or AUC between a machine learning-based model and the COX model, a very minor improvement. Although RSF algorithms are a promising alternative to conventional COX model, a significant limitation is the requirement of higher number of events for training.[31]

Machine learning models do have advantages over regression model because they can handle non-linear effects and interactions among predictors. However, regression models are much easier to implement in operations. In addition, the required efforts to update or recalibrate a model being transferred from the development site are much less for a regression model. This puts the regression-based model in favor even if their performance is slightly compromised compared to machine learning-based models.

High order and interaction terms are often considered when the interest of performing the regression model is to understand the effects of covariates on the outcome; however, they are often ignored when the model is developed for the purpose of risk prediction. In the current

study, the disadvantage in discrimination measured by c-index for the Cox proportional hazards regression model compared to machine learning based model is enlarged when the high order terms are ignored. This is because Cox proportional hazards regression models are log-linear models, and thus, lacks the flexibility of data fitting unless high order terms and/or interaction terms are added.

The XGB model over-estimated the risks of PDAC in the top three highest risk groups. However, this flaw does not impact its ability to identify high risk patients, because the observed risks in the highest three risk groups of patients were also a bit higher, compared to those of RSF and XGB. In other words, if XGB model is used to identify high risk patients in a relative way (e.g., top 20% of patients in terms of risk) rather than know the actual risk, it is still a good choice.

The present study has several other limitations. First, of the 1479 and 4,582 events identified in the KPSC and VA cohorts, respectively, 300 and 2,564 events were captured by data sources other than Cancer Registry. An evaluation based on the KPSC Cancer Registry of the same time window showed that about 90% of pancreatic cancer cases were PDAC. Second, to estimate sensitivity, specificity, PPV and fold of risk increase, we relied on a subset of patients (~70% and ~80% of the total patients in the KPSC and VA cohorts, respectively) with complete follow up unless they died of pancreatic cancer. This restriction over-estimated the risk of PDAC, because the patients who were excluded from this analysis were at-risk for some periods of time. Third, some important predictors had high percentages of unknown values (i.e., missing data). Although multiple imputation was performed, bias may occur if the missing at random (MAR) assumption is violated. Finally, our comparison was performed based on EHR-based data extracted from integrated large health care systems and the outcome of interest is extremely rare. Our conclusions may not be applicable to other scenarios (e.g., more frequent outcomes).

Conclusion

Three common approaches for predicting time-to-event outcomes were compared in a large, diverse integrated health system and subsequently validated in a separate health system. All three models were parsimonious and identified key factors in determining risk of pancreatic cancer. Findings from the present study provide insights into model selection for targeted screening of pancreatic cancer based on automated analysis of data in EHR. Future studies may achieve better performance by developing deep learning models and/or fuse individual parametric, semi-parametric, and non-parametric risk prediction models.[32]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors thank the patients of Kaiser Permanente for helping to improve care through the use of information collected through our electronic health record systems. The authors also thank Sole Cardoso for the assistance with formatting the manuscript.

Source of Funding:

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA230442. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Abbreviations

ALT	Alanine transaminase
AUC	area under the curve
AUROC	area under the ROC curve
CI	confidence interval
CDW	Corporate Data Warehouse
COX	Cox proportional hazards regression
CPRS	Computerized Patient Record System
EHR	electronic health record
GND	Greenwood-Nam-D'Agostino
HbA1C	glycated hemoglobin
KPSC	Kaiser Permanente Southern California
MAR	missing at random
PC	pancreatic cancer
PDAC	pancreatic ductal adenocarcinoma
PPV	positive predictive value
RSF	Random Survival Forest
SD	standard deviation
SNN	survival neural networks
USPTF	United States Preventative Services Task Force
VA	Veterans Affairs Greater Los Angeles Healthcare System
VHA	Veterans Health Administration
XGB	eXtreme gradient boosting

References

1. National Cancer Institute Surveillance, Epidemiology and End Results Program: Cancer stat facts: Pancreatic cancer. https://seer.cancer.gov/statfacts/html/pancreas.html. Last accessed: March 31, 2023.

2. Kleeff J, Korc M, Apte M, La Vecchia C, Johnson CD, Biankin AV et al. : Pancreatic cancer. Nat Rev Dis Primers 2016; 2: 16022. [PubMed: 27158978]

3. Force USPST, Owens DK, Davidson KW, Krist AH, Barry MJ, Cabana M et al. : Screening for pancreatic cancer: Us preventive services task force reaffirmation recommendation statement. Jama 2019; 322: 438–444. [PubMed: 31386141]

4. Dbouk M, Katona BW, Brand RE, Chak A, Syngal S, Farrell JJ et al. : The multicenter cancer of pancreas screening study: Impact on stage and survival. J Clin Oncol 2022; 40: 3257–3266. [PubMed: 35704792]

5. Chen W, Zhou Y, Xie F, Butler RK, Jeon CY, Luong TQ et al. : Derivation and external validation of machine learning-based model for detection of pancreatic cancer. Am J Gastroenterol 2023; 118: 157–167. [PubMed: 36227806]

6. Miao F, Cai Y-P, Zhang Y-T, Li C-Y: Is random survival forest an alternative to cox proportional model on predicting cardiovascular disease? Cham, Springer International Publishing, 2015; pp. 740–743.

7. Grendas LN, Chiapella L, Rodante DE, Daray FM: Comparison of traditional model-based statistical methods with machine learning for the prediction of suicide behaviour. J Psychiatr Res 2021; 145: 85–91. [PubMed: 34883411]

8. Herrin J, Abraham NS, Yao X, Noseworthy PA, Inselman J, Shah ND et al. : Comparative effectiveness of machine learning approaches for predicting gastrointestinal bleeds in patients receiving antithrombotic treatment. JAMA Netw Open 2021; 4: e2110703.

9. Kurt Omurlu I, Ture M, Tokatli F: The comparisons of random survival forests and cox regression analysis with simulation and an application related to breast cancer. Expert Systems with Applications 2009; 36: 8582–8588.

10. Du M, Haag DG, Lynch JW, Mittinty MN: Comparison of the tree-based machine learning algorithms to cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on seer database. Cancers (Basel) 2020; 12: 2802. [PubMed: 33003533]

11. Datema FR, Moya A, Krause P, Back T, Willmes L, Langeveld T et al. : Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression. Head Neck 2012; 34: 50–58. [PubMed: 21322080]

12. Hazewinkel A-D, Gelderblom H, Fiocco M: Prediction models with survival data: A comparison between machine learning and the cox proportional hazards model. medRxiv 2022: 2022.2003.2029.22273112.

13. Khene ZE, Bigot P, Doumerc N, Ouzaid I, Boissier R, Nouhaud FX et al. : Application of machine learning models to predict recurrence after surgical resection of nonmetastatic renal cell carcinoma. Eur Urol Oncol 2022.

14. Fihn SD, Francis J, Clancy C, Nielson C, Nelson K, Rumsfeld J et al. : Insights from advanced analytics at the veterans health administration. Health affairs (Project Hope) 2014; 33: 1203–1211. [PubMed: 25006147]

15. Chen W, Yao J, Liang Z, Xie F, McCarthy D, Mingsum L et al. : Temporal trends in mortality rates among kaiser permanente southern california health plan enrollees, 2001–2016. The Permanente journal 2019; 23.

16. Wright MN, Ziegler A: Ranger: A fast implementation of random forests for high dimensional data in c++ and r. Journal of Statistical Software 2017; 77.

17. Little RJA: Missing-data adjustments in large surveys. Journal of Business & Economic Statistics 1988; 6: 287–296.

18. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS: Random survival forests. Ann Appl Stat 2008: 841–860.

19. Dietrich S, Floegel A, Troll M, Kuhn T, Rathmann W, Peters A et al. : Random survival forest in practice: A method for modelling complex metabolomics data in time to event analysis. International journal of epidemiology 2016; 45: 1406–1420. [PubMed: 27591264]

20. Ishwaran H KU: Randomforestsrc: Fast unified random forests for survival, regression, and classification (rf-src);

21. Allison PD: Survival analysis using the sas system. Cary, nc: Sas institute. Inc Google Scholar 1995.

22. Cox DR: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 1972; 34: 187–202.
23. Jia X, Baig MM, Mirza F, GholamHosseini H: A cox-based risk prediction model for early detection of cardiovascular disease: Identification of key risk factors for the development of a 10-year cvd risk prediction. Adv Prev Med 2019; 2019: 8392348.
24. Lin WP, Xing KL, Fu JC, Ling YH, Li SH, Yu WS et al. : Development and validation of a model including distinct vascular patterns to estimate survival in hepatocellular carcinoma. JAMA Netw Open 2021; 4: e2125055.
25. Chen T, Guestrin C: Xgboost; In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016; pp. 785–794.
26. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H et al.: Xgboost: Extreme gradient boosting;
27. Wood AM, Royston P, White IR: The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. Biom J 2015; 57: 614–632. [PubMed: 25630926]
28. Demler OV, Paynter NP, Cook NR: Tests of calibration and goodness-of-fit in the survival setting. Statistics in medicine 2015; 34: 1659–1680. [PubMed: 25684707]
29. Krzyziński M, Spytek M, Baniecki H, Biecek P: Survshap(t): Time-dependent explanations of machine learning survival models. Knowledge-Based Systems 2023; 262: 110234.
30. Spytek M, Krzyziński M, Baniecki H, Biecek P: Survex: Explainable machine learning in survival analysis. R package version 100 2023.
31. Baralou V, Kalpourtzi N, Touloumi G: Individual risk prediction: Comparing random forests with cox proportional-hazards model by a simulation study. Biom J 2022.
32. Wey A, Connett J, Rudser K: Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. Biostatistics 2015; 16: 537–549. [PubMed: 25662068]

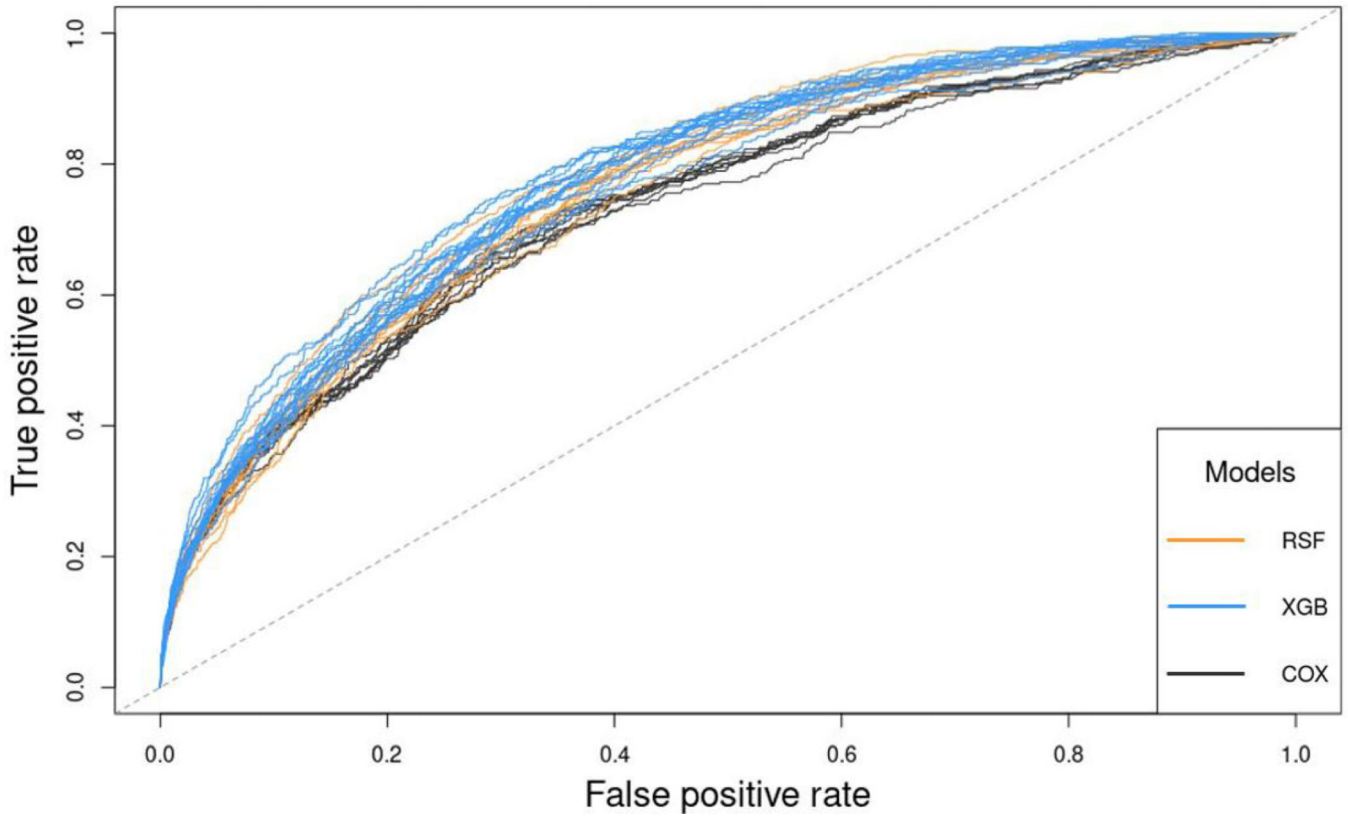

Figure 1. Area under the ROC curve (AUROC) for final model by model type based on KPSC validation data.

x-axis: False positive rate or (1-specificity); y-axis: True positive rate or sensitivity; RSF: orange, XGB: blue, COX: black.

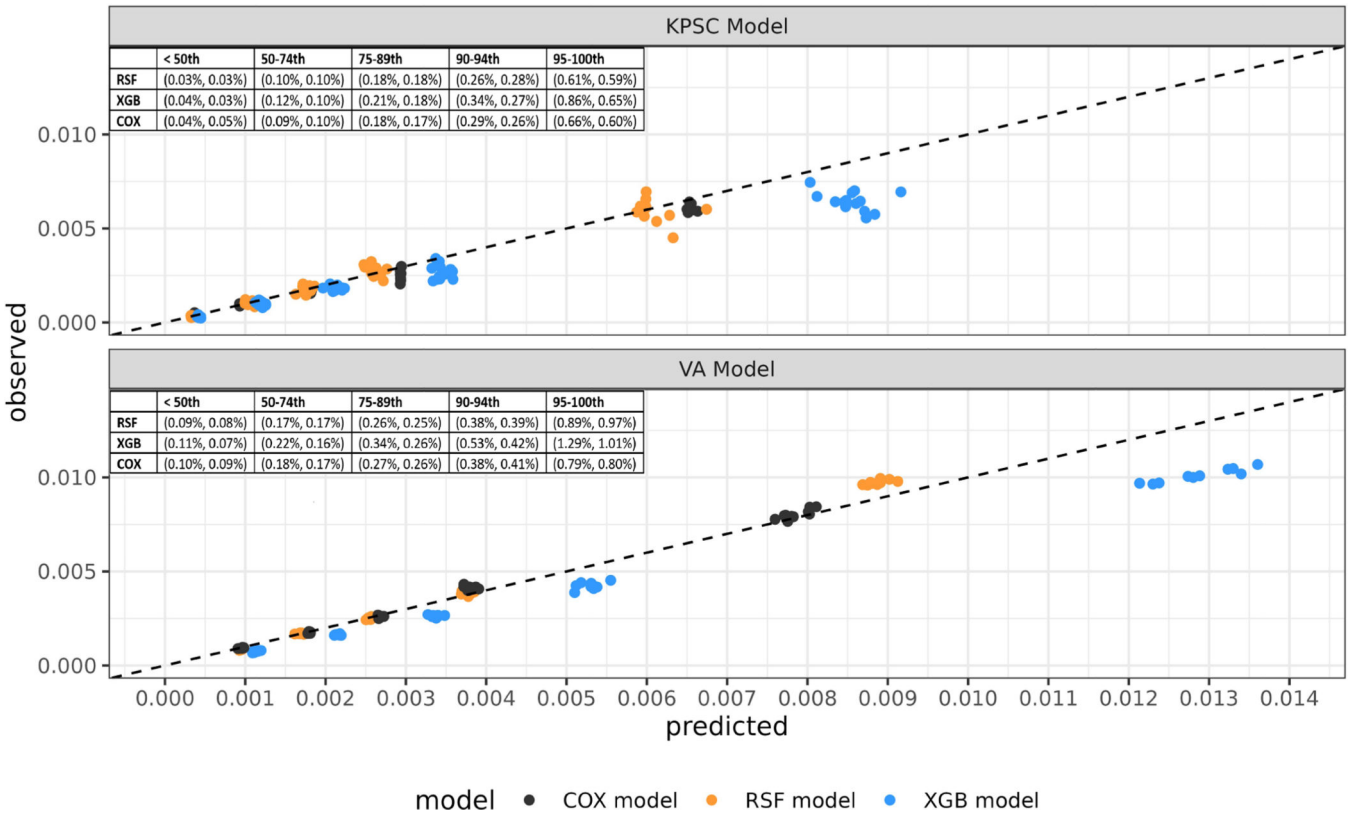

Figure 2. Calibration plots by model type (RSF, XGB, COX) based on KPSC validation data (left) and VA testing data (right).

x-axis: predicted; y-axis: observed. The five clusters represent the five risk groups defined by the ranges of predicted risks: <50th, 50–74th, 75–89th, 90–94th, 95–100th percentiles. Within each cluster, there are multiple dots representing the pairs of predicted and observed risks, calculated based on the corresponding validation or testing datasets.

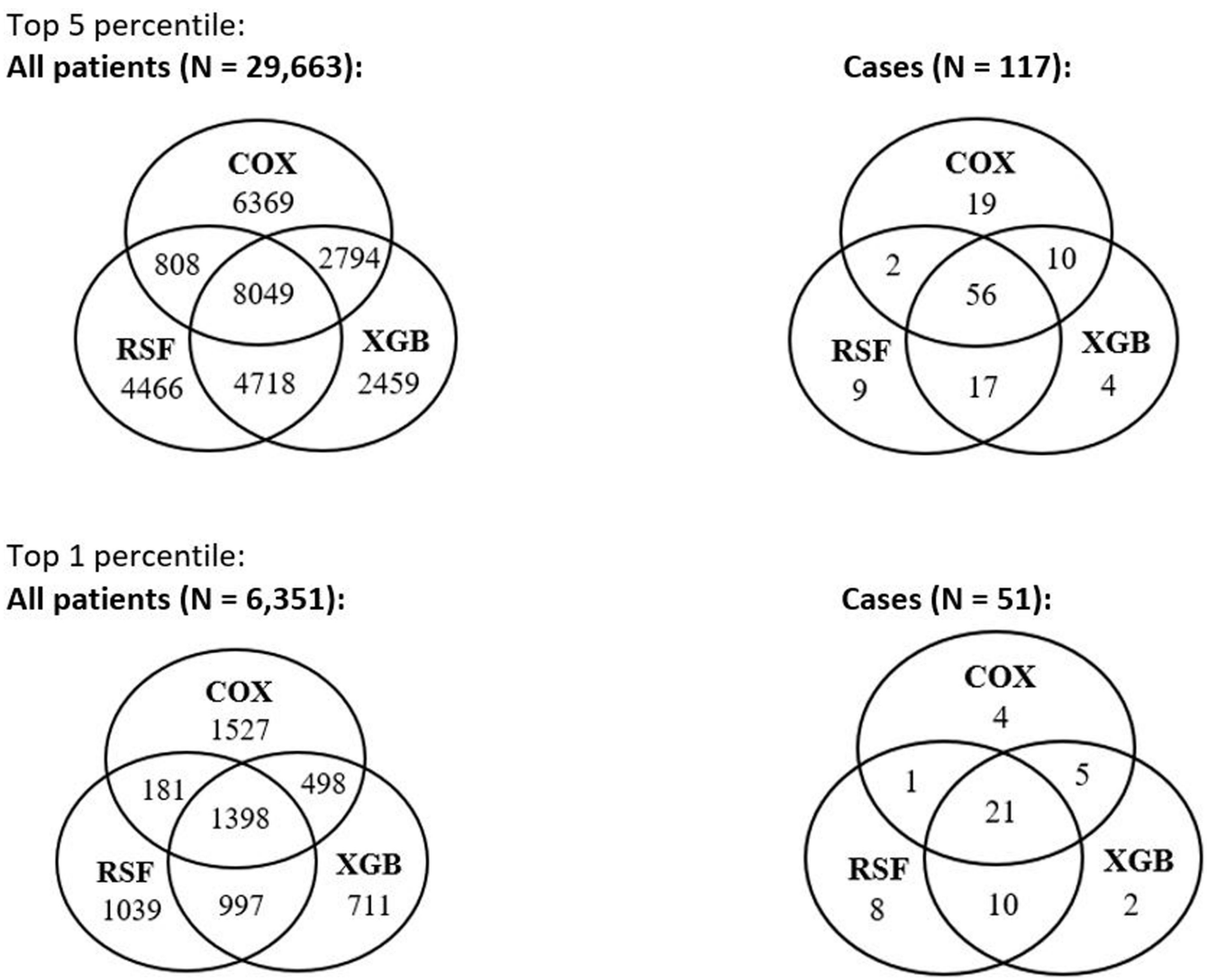

Figure 3. Venn diagram demonstrating heterogeneity across model type (KPSC validation data only).

Top: patients with predicted risk within at the top 5 percentiles (N=29,663 identified by any of the three models) ; Bottom: patients with predicted risk at the top 1st percentile (N=6,351 identified by any of the three models).

Table 1.

Predictors selected and model performance based on internal validation and external testing data.

Model Type	Predictors							Validation KPSC			Testing[a] VA
	Age	ALT		HgA1c		Abdominal Pain	Weight Change in 1 yr	c-index Mean (SD)	χ^2 Mean (SD)	p-value Mean (SD)	c-index Mean (SD)
		Change in 1 yr	Rate of Change in 1 yr	HbA1c	Rate of Change in 1 yr						
RSF	X	X		X		X	X	0.77 (0.02)	5.9 (4.1)	0.4 (0.3)	0.71 (0.002)
XGB	X		X	X		X	X	0.78 (0.01)	34.3 (14.5)	< 0.01 (< 0.01)	0.74 (0.008)
COX[b]	X		X	X		X	X	0.74 (0.01)	6.0 (3.0)	0.3 (0.1)	0.70 (0.005)

Model type: RSF: Random Survival Forest; XGB: eXtreme gradient boosting; COX: COX proportional hazards regression model.

Other abbreviations: ALT, alanine transaminase; C-index, concordance index; HgA1c, hemoglobin A1c; KPSC, Kaiser Permanente Southern California; NA, not applicable; SD, standard deviation; VA, Veterans Affairs

[a]Models were recalibrated using the exact features selected by RSF, XGB, and COX, respectively.

[b]The formula of the resulting prediction model is:

$P(t) = P_0(t)\ exp[\beta X]$, where $\beta X = 0.073 \times$ age $+ 1.012 \times$ Abdominal Pain[yes] $+ 0.331 \times$ Abdominal Pain[unknow] $+ 0.089 \times$ HGBA1C $- 0.055 \times$ HGBA1C$^2 + 0.074 \times$ [ALT Rate] $- 0.001 \times$ [ALT Rate]^2 $+ 0.814 \times$ [Weight Change] $+ 0.090 \times$ [Weight Change]^2

P(18 month) $= 0.000052 \times$ age $+ 0.000726 \times$ Abdominal Pain[yes] $+ 0.000237 \times$ Abdominal Pain[unknow] $+ 0.000064 \times$ HGBA1C $- 0.000039 \times$ HGBA1C$^2 + 0.000053 \times$ [ALT Change Rate] $- 0.000001$ $\times$ [ALT Change Rate]$^2 + 0.000584 \times$ [Weight Change] $+ 0.000065 \times$ [Weight Change]^2

Table 2a.

Percent of patients[a] whose risk was among the top 20%, 15%, 10%, 5%, and 2.5%, sensitivity, specificity, positive predictive value (PPV), and risk fold increase based on KPSC validation datasets, by model type (RSF, XGB, COX).

	RSF					XGB					COX				
	High-Risk Patients (percentile)					High-Risk Patients (percentile)					High-Risk Patients (percentile)				
	Top 20th	**Top 15th**	**Top 10th**	**Top 5th**	**Top 2.5th**	**Top 20th**	**Top 15th**	**Top 10th**	**Top 5th**	**Top 2.5th**	**Top 20th**	**Top 15th**	**Top 10th**	**Top 5th**	**Top 2.5th**
N[b]	50952	37454	24959	12498	6249	49941	37457	24970	12485	6243	49958	37468	24982	12488	6245
Sensitivity (%)	56.6	48.8	39.0	26.9	19.5	57.9	50.3	41.3	29.5	21.1	51.5	44.6	38.2	27.6	20.3
Specificity (%)	79.6	85.1	90.0	95.0	97.5	80.1	85.1	90.0	95.0	97.5	80.0	85.0	90.0	95.0	97.5
PPV (%)	0.4	0.5	0.6	0.8	1.1	0.4	0.5	0.6	0.8	1.2	0.4	0.4	0.5	0.8	1.2
Fold Increase in Risk[c]	2.8	3.3	3.9	5.4	7.9	2.9	3.4	4.1	5.9	8.5	2.6	3.0	3.8	5.5	8.2
AUC & 95% CI	0.767 (0.744 – 0.791)					0.779 (0.755 – 0.802)					0.737 (0.710 – 0.764)				

Abbreviations: RSF: Random Survival Forest; XGB: eXtreme gradient boosting; COX: COX proportional hazards regression model; KPSC, Kaiser Permanente Southern California; PPV, positive predictive value.

[a]Estimated in patients with complete 18 months follow up or those who developed PDAC in 18 months.

[b]Number of eligible patients whose risk was above each risk threshold.

[c]Compared with the incidence rate in the entire cohort.

Performance for model RSF: estimated based on 11 validation samples. Performance for model XGB: estimated based on 14 validation samples. Performance for model COX: estimated based on 9 validation samples.

Table 2b.

Percent of patients[a] whose risk was among the top 20%, 15%, 10%, 5%, and 2.5%, sensitivity, specificity, positive predictive value (PPV), and risk fold increase based on VA testing dataset, by model type (RSF, XGB, COX).

	RSF					XGB					COX				
	High-Risk Patients (percentile)					High-Risk Patients (percentile)					High-Risk Patients (percentile)				
	Top 20th	Top 15th	Top 10th	Top 5th	Top 2.5th	Top 20th	Top 15th	Top 10th	Top 5th	Top 2.5th	Top 20th	Top 15th	Top 10th	Top 5th	Top 2.5th
N[b]	424761	318408	212392	106118	53528	424540	318406	212274	115139	53068	424308	318255	212172	106098	53055
Sensitivity (%)	51.7	45.5	37.7	27.2	19.4	53.7	47.2	39.2	28.2	20.3	48.6	41.9	34.1	23	15.5
Specificity (%)	80.2	85.1	90.1	95.1	97.5	80.1	85.1	90.1	95.1	97.6	80.1	86.2	90.1	95.0	97.5
PPV (%)	0.6	0.7	0.8	1.2	1.7	0.6	0.7	0.8	1.2	1.7	0.5	0.6	0.7	1.0	1.3
Fold Increase in Risk[c]	2.6	3.0	3.7	5.4	7.7	2.6	3.1	3.8	5.5	7.9	2.4	2.8	3.3	4.5	6.1
AUC & 95% CI	0.731 (0.724 – 0.739)					0.742 (0.735 – 0.750)					0.706 (0.699 – 0.714)				

Abbreviations: RSF: Random Survival Forest; XGB: eXtreme gradient boosting; COX: COX proportional hazards regression model; PPV, positive predictive value; VA, Veterans Affairs.

[a]Estimated in patients with complete 18 months follow up or those who developed PDAC in 18 months.

[b]Number of eligible patients whose risk was above each risk threshold.

[c]Compared with the incidence rate in the entire cohort.

Performance was estimated based on 10 imputed testing datasets for all three models.