HHS Public Access

Author manuscript

*Am J Gastroenterol*. Author manuscript; available in PMC 2025 August 01.

Published in final edited form as:

*Am J Gastroenterol*. 2024 August 01; 119(8): 1466–1482. doi:10.14309/ajg.0000000000002870.

# Machine Learning Models for Pancreatic Cancer Risk Prediction Using Electronic Health Record Data - A Systematic Review and Assessment

**Anup Kumar Mishra, PhD, MS**[1], **Bradford Chong, MD**[1], **Shivaram P. Arunachalam, PhD, DBA**[1], **Ann L. Oberg, PhD**[2], **Shounak Majumder, MD**[1]

[1]Department of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN, USA

[2]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA

## Abstract

**Introduction:** Accurate risk prediction can facilitate screening and early detection of pancreatic cancer (PC). We conducted a systematic review to critically evaluate effectiveness of machine learning (ML) and artificial intelligence (AI) techniques applied to Electronic Health Records (EHR) for PC risk prediction.

**Methods:** Ovid MEDLINE(R), Ovid EMBASE, Ovid Cochrane Central Register of Controlled Trials, Ovid Cochrane Database of Systematic Reviews, Scopus, and Web of Science were searched for articles that utilized ML/AI techniques to predict PC, published between January 1st, 2012 to February 1st, 2024. Study selection and data extraction were conducted by two independent reviewers. Critical appraisal and data extraction was performed using CHARMS checklist. Risk of bias and applicability was examined using PROBAST.

**Results:** Thirty studies including 169,149 PC cases were identified. Logistic regression was the most frequent modeling method. Twenty studies utilized a curated set of known PC risk predictors or those identified by clinical experts. ML model discrimination performance (C-index) ranged from 0.57 to 1.0. Missing data was underreported, and most studies did not implement explainable-AI techniques or report exclusion time intervals.

**Discussion:** AI/ML models for PC risk prediction using known risk factors perform reasonably well and may have near-term applications in identifying cohorts for targeted PC screening if validated in real-world data sets. The combined use of structured and unstructured EHR data using emerging language models while incorporating explainable-AI techniques has the potential to identify novel PC risk factors and this approach merits further study.

Correspondence: Shounak Majumder, M.D., Associate Professor of Medicine, Division of Gastroenterology and Hepatology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Majumder.Shounak@mayo.edu.

Specific author contributions:

A.K.M., B.C., A.L.O., S.M.: conception, design, acquisition, analysis, and drafting manuscript. All authors: interpretation of data for the work and reviewing manuscript. All authors: final approval of manuscript.

Conflict of interests:

S.M., Mayo Clinic and Exact Sciences have an intellectual property development agreement. Dr. Majumder is listed as inventor under this agreement and could share potential future royalties as an employee of Mayo Clinic. The other authors of this manuscript have no conflict of interest to declare.

### Keywords

Artificial Intelligence; Electronic Health Records; Pancreatic Cancer; Early Detection of Cancer; Risk Factors

## INTRODUCTION:

While pancreatic cancer (PC) is ranked as the 11th most common cancer in the world with 458,918 new cases in 2018 [1], it is projected to be the second leading cause of cancer-related mortality in the United States by 2030 [2]. Most of the mortality is attributed to advanced stage at diagnosis, and hence, only a minority of patients (15–20%) are eligible for surgical resection [3, 4]. Earlier diagnosis of PC with localized disease correlates with improved survival [5]. The low incidence of PC and lack of accurate biomarkers for early-stage disease have made effective screening challenging and hindered efforts to improve overall survival. As PC screening in the general population is not recommended, efforts have been made to identify high-risk individuals who may benefit from PC screening [6]. In current practice PC screening is limited to individuals with pathogenic/likely pathogenic germline mutations in PC susceptibility genes and those with multiple affected family members [7, 8]. However, less than 20% of PC patients have known familial and genetic risk factors thereby limiting the ability to enrich and screen the population at risk. Therefore, identifying novel risk factors for PC is critical.

Electronic Health Records (EHR) data contain a variety of structured and unstructured data, which have shown promising results in disease and risk prediction. With EHR being more pervasively used across health systems and with the recent developments in the field of machine learning (ML) and deep learning (DL), EHR data could potentially be explored for effective prediction of PC risk [9]. Identified high risk individuals could then benefit from PC screening. Also, with emerging Explainable-Artificial Intelligence (X-AI) techniques, interpretable risk factors of PC could be identified from the EHR data [10].

We therefore sought to systematically review the existing ML/AI literature that utilizes EHR data to predict PC risk, and to summarize model development, evaluation strategies, and model effectiveness in predicting PC.

## METHODS:

### Data sources and searches

A comprehensive search of several databases from January 1st, 2012 to February 1st, 2024 in the English language, was conducted. The databases included Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, and Daily, Ovid EMBASE, Ovid Cochrane Central Register of Controlled Trials, Ovid Cochrane Database of Systematic Reviews, Scopus, and Web of Science. The search strategy was designed and conducted by an experienced librarian with input from the study's principal investigator. Controlled vocabulary supplemented with keywords was used to search for ML and natural language processing (NLP) models pertaining to prediction of PC and PC risk factors using EHR data.

The actual strategy listing all search terms used and how they are combined is available in the ‘Supplementary Digital Content 1 - Article Search Strategies’ document.

### Study inclusion criteria

We included articles that developed a multivariable ML model to predict PC using EHR data.

### Outcome

The outcome was PC.

### Compilation and screening of articles

Two independent reviewers (A.K.M., B.C.) screened articles for eligibility, based on title and abstract, followed by a second round of full-text review to identify eligible articles. This was followed by screening their respective reference lists and citation matching for additional articles. A third independent review (S.M.) adjudicated any disagreement in eligible articles. The articles were archived into Endnote software [11].

### Extraction and quality assessment

We used the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) checklist [12] to extract data for the appraisal of the articles. We extracted study details including study type and time-period, data sources, participants, reporting and handling of missing data, machine learning modeling methods, model calibration, validation, and performance. In addition to the CHARMS checklist, we extracted data including choice of candidate predictors in the study: curated PC predictors derived from literature or identified by experts versus non-curated predictors in the EHR; study population type: high-risk subgroups versus general population; prediction time window; and novel risk factor identification through model explainability. We also used PROBAST to evaluate risk-of-bias and applicability of the models developed and validated in the included articles [13, 14]. For quality assessment, we applied the PRISMA checklist to guide our systematic review [15].

We used the C-index metric as the metric for model performance. Studies included in our systematic review were very heterogeneous in data exclusion windows (excluding data immediately prior to diagnosis), prediction time windows (duration of future disease risk period from date of clinical assessment), number of independent datasets and subset groups used and modeling techniques. Thus, for studies that explored multiple data exclusion windows, results corresponding to the smallest exclusion window were used. For studies that experimented with multiple prediction time windows, results corresponding to the shortest prediction window were considered. If studies utilized multiple independent datasets, all datasets were included as individual data points. However, if studies performed both full cohort and subset analyses, e.g. subset to patients with new onset diabetes, full cohort results are reported but subset results were excluded. For studies that explored multiple modeling techniques, results from each modeling technique were included as individual data points if the corresponding results were reported consistently across datasets. If results from two or more similar modeling techniques (e.g. Light Gradient Boosting machine, Gradient

Boosting Machine) were reported in an article, only results from best performing model were reported. Also, modeling techniques were categorized into three groups: Group A included linear ML models, Group B included non-linear models excluding deep learning models, and Group C included deep learning models only.

## RESULTS:

### Study characteristics

With our PICO search, we identified 183 articles after removing duplicates. These articles were screened to identify 21 articles which have implemented ML algorithms to predict PC. We added nine additional articles from references that met our inclusion criteria. Figure 1 shows the process of study identification and inclusion for data extraction and analysis. Tables 1 and 2 describe the study characteristics for risk prediction models including study type, data sources, modeling development techniques, validation results using the CHARMS checklist framework. Supplementary Tables 1 and 2 describe novel risk factors identified by studies and additional modeling characteristics such as missing data handling, respectively. We excluded four articles due to unclear data sources [16], no multivariate model development [17], unclear predictor utilization in modeling [18], and significant overlap of data and modeling methods with another included study [19].

Most studies considered a composite PC outcome and did not differentiate between pancreatic ductal adenocarcinoma, neuroendocrine tumors, or other specific types of PC.

Most of the studies utilized curated high-risk predictors based on PC literature or clinical expertise (n=20) [20–39]. Figure 2a shows the percentage of studies that utilized curated vs non-curated predictors. Moreover, we observed that a greater proportion of models in Group A (linear models, 8/14) and Group B (non-linear models excluding deep learning, 9/14) utilized curated sets of candidate predictors as compared to Group C (deep learning, 1/9) (Figure 3a). Models that limited their analysis to curated risk-factors reported a similar discrimination performance (mean C-index=0.81, min=0.61, max=1.0, n=18) when compared to models that did not (mean C-index=0.80, min=0.72, max=0.93, n=19) (Figure 3b).

### ML model development and evaluation

Logistic regression was most frequently (n=18) utilized for model development (Table 2). In addition, a diverse range of modeling techniques were employed to build PC prediction models. These include tree-based models such as XGBoost, random forests, survival models such as random survival forests, cox regression, multistate models. Furthermore, neural network-based models such as artificial neural networks, as well as more advanced deep learning-based approaches including gated recurrent units, and transformers were utilized to build the models.

Sixteen studies provided information about missing data and how missing data was handled (Supplementary Table 2, Figure 2b). The most common approaches of missing data handling included exclusion of patients [28, 40], exclusion of predictors with large percentage of missingness [22–25, 33], and imputation of predictors [22, 24, 25, 39]. In three studies, missingness

had been replaced by categorical values such as ‘Not known’ [26] and ‘missing’ [28] or created a binary variable with value −1 for missing data [31]. We also observed that in one study, missing laboratory result values were considered the same as those with normal results [36].

The studies predicted PC occurrence within a prediction time window of up to 8 years after the date of risk assessment (Table 2). We observed that six articles did not provide any information about the prediction time window or data exclusion time intervals [20, 26, 28, 31, 36, 41]. Only twelve studies experimented with 1 month to 5 years of data exclusion time intervals [21, 29, 30, 33, 35, 40, 42–47]. C-index for the models without a curated set of predictors and 1 year lead time or exclusion time interval ranged from 0.71 to 0.83 for internal validations and 0.60 to 0.78 for external validations. Figure 4 shows performance of the same models with no data exclusion (or smallest time interval data exclusion) settings versus data exclusion (or maximum time interval data exclusion) settings in different model groups. The figure represents results from internal validations of nine models presented in five different manuscripts (Group A: linear models, n=3 [21, 40]; B: non-linear models excluding deep learning models, n=3 [40, 43]; and C: deep learning models only, n=3 [33, 42]). Four studies that experimented with data exclusion intervals were excluded from this analysis due to no minimum and maximum data exclusion experiment results reported [29, 30], no c - index reported [35], or no internal validation results reported [44].

We observed that 24 studies performed either an internal or external or both internal and external validations (Table 2). Some internal validations were conducted by evaluating the model on a holdout test set, typically 20% of the dataset. Several studies used bootstrapping for internal validation. External validations were conducted by evaluating the model performance on an external dataset from a different health system or geographic region [42, 44, 46]. A distribution of model validation methods utilized in the different studies included in our review is presented in Figure 2c. Figure 5a and 5b present the performance of different model groups in internal and external validation settings, respectively. Models from the six studies that did not perform any form of validation were excluded from this illustration [26–28, 32, 34, 36]. For internal validation, the average C-index for models in Group A, B, and C were 0.77, 0.83, and 0.83, respectively. For external validation, the average C-index for models in Group A, B, and C were 0.77, 0.79, and 0.88, respectively. Group C for external validation included results from a single study only. Model performances on all exclusion/lead time intervals, prediction time windows, and datasets are presented in Table 2.

Ten studies performed a calibration analysis (Figure 2d, Supplementary Table 2) [20–25, 37, 43, 44, 46]. The model calibration analyses were conducted using Hosmer–Lemeshow chi-square goodness-of-fit tests, Greenwood-Nam-D’Agostino (GND) calibration tests, Platt calibration, and calibration graphs.

### Identifying novel risk factors of PC

Six studies that did not rely on a curated set of predictors [42–47] identified novel risk factors utilizing X-AI techniques (Supplementary Table 1). Chen et al. utilized XGBoost gains to identify that pancreatic disorders (noncancerous and not relating to diabetes mellitus) was the most important model predictor [43]. Placido et al. explored integrated gradients in neural

networks, finding jaundice, abdominal pain, and weight loss as key features 0–6 months before PC diagnosis [42]. With a longer interval before cancer diagnosis, key contributors included diabetes mellitus, anemia, functional bowel disease, and other pancreatic, bile duct diseases, and cancers. [42]. Salvatore et al. grouped relevant ICD codes into clinically relevant phenotypically related aggregates, 'phecodes'. Using co-occurrence analysis, they identified that digestive and neoplasm phecodes were strong predictors of PC [44]. Park et al. utilized SHapley Additive exPlanations (SHAP) values to identify that kidney group, liver function group, diabetes group, red blood cell, and white blood cell group of labs contributed the most in predicting PC risk from labs [47]. Jia et al. ranked features by univariate AUC to identify the independent contributors to PC risk prediction [46]. The top 5 predictors from their analysis were age, number of recent records, creatinine in serum, plasma, or blood, number of early records, diabetes mellitus without complications diagnosis group, and essential hypertension diagnosis group. Zhu et al. reported that disease of pancreas, unspecified (ICD10 K86.9), malignant neoplasm of transverse colon (ICD10 C18.4), pseudocyst of pancreas (ICD10 K86.3), hypertrophy of breast (ICD10 N62), neoplasm of unspecified behavior of digestive system (ICD 10 D49.0) were the key factors based on model odds ratios [45].

### Risk of bias assessment

We used PROBAST to assess risk of bias of the models included in our study[13]. If two or more models were developed in a study, risk of bias for the best performing model (highest C-index) was assessed using PROBAST. Models from only four studies had low risk of bias [33, 42, 46, 47]. Supplementary Table 3 presents a summary of the PROBAST risk of bias and applicability assessment.

## DISCUSSION:

We extracted and reviewed data from thirty studies to discern state-of-the-art ML methods for predicting PC risk and identifying novel risk factors from EHR data. Most studies could develop models with a discriminative performance ranging from 0.57–1.0. However, there were many potential sources for risk of bias including outcome definition, predictor selection, short prediction time window, and reporting and handling of missing data.

Most of the studies defined PC as a composite outcome, by using a range of ICD codes. Two types of PC account for most cases: pancreatic adenocarcinoma, PDAC; (85% of cases), and pancreatic neuroendocrine tumor, PNET; (less than 5%) [48]. PDAC, PNET, and other PC types have different tumor biology, natural history, and risk factors. Predicting PC as a composite outcome is problematic, as key contributing predictors identified for all PCs may not apply to PDAC or PNET, specifically.

Most of the studies used logistic regression for model development but did not provide information about assessing modeling assumptions. Nor was sufficient information provided to determine whether controls were sampled appropriately, ensuring they are representative of the population from which cases develop at the case index date [49]. Non-linear and deep learning-based AI models had similar discrimination performance (C - index) when compared with traditional linear machine learning models (Figure 5). It is crucial to note that

more caution is warranted for computationally expensive models to prevent overfitting [50]. This is because the model complexity that enables identifying signal in the training data to make accurate predictions can also make the model more susceptible to capturing nuanced noise that does not generalize to other populations as patterns. Therefore, to mitigate these issues, increasing sample size, employing regularization and resampling/internal model validation techniques, and conducting external validation in data from other populations and institutions is crucial. External validations will test model robustness and generalizability beyond the initial development setting. Also, it is important to note that Group C has only one sample. Hence, our understanding of the performance of Group C in an external validation setting is currently limited.

It is critical to examine performance of the final models in different subgroups to ensure that the model is fair to the subgroups (similar discrimination ability) and not significantly advantaged/disadvantaged in certain groups. We found that only four studies performed/mentioned any subgroup analysis by age[30, 43] and race[32, 33]. Jia et al. performed model development using data from different race groups and geographic locations and tested model performance using data from excluded races and locations[46]. However, none of the studies reported any fairness matrices such as equalized odds and equalized opportunity[51].

Most of the studies used a curated set of high-risk predictors based on PC literature or clinical expertise (Supplementary Table 2). The EHR clinical data include structured data such as medications, and unstructured data such as free text clinical notes. Few studies used a combination of structured and unstructured data to develop the models. Figure 3b shows that not utilizing a curated set of high-risk predictors resulted in similar mean discriminatory performance, though this could potentially favor identifying novel risk factors. Chen et al. used various EHR-based candidate predictors to develop their XGBoost models, but many of the features have limited interpretability, such as “strain” and “runny” [43]. The XGBoost model viewed each word in clinical notes individually, while a transformer-based approach can retain the context of words and phrases in the clinical notes data [52].

Several studies did not provide any information about missing data and missing data handling (Supplementary Table 2) [20, 21, 27, 30, 32, 34, 38, 41–44, 53]. Missing data and how the missingness has been handled could impact prognostic model performance and applicability [54]. The estimated predictor outcome associations and predictive performance measures of the model are unbiased only if excluded participants are a completely random subset of the original study sample [55]. A comparison of the participants with and without missing values could provide better understanding of potential bias in the data. For models utilizing structured data, multiple imputation has shown to perform superior in terms of bias and precision [56, 57]. Also, deep learning-based approaches including recurrent neural networks can efficiently handle irregularities and missing patterns in time series clinical data [58, 59].

The PC occurrence prediction time window in the studies ranged up to 8 years of the date of risk assessment (Table 2). Most studies did not consider data exclusion intervals. Such modeling strategies are not appropriate for early detection and can introduce high risk of bias (Supplementary Table 3), as the predictor data close to the time of PC diagnosis will most likely be symptoms of the disease instead of true predictors of future risk. Among

studies that did consider data exclusion intervals, deep learning-based modeling techniques performed better on average with minimum or no data exclusion and performed comparable to non-linear models for maximum data exclusion periods (3 months to one year) for the same models in each group; linear models had least discrimination performance with data exclusion intervals as shown in Figure 4. There was also a decline in performance with data exclusion in Group C models when compared with Group A and B models. With a sample size of three across groups it is difficult to draw any strong conclusions. However, this could suggest that the Group C deep learning models developed in these studies depended more on data closer to the PDAC event than other groups. Studies show that predictor data considered with a lead time of 24 to 36 months prior to PC diagnosis may be most appropriate [35, 60, 61].

Identification of novel risk factors is important as about 80% of PC is considered sporadic in etiology. Explainability of an ML model pertains to the clarity of its internal logic and mechanics, enabling deeper comprehension of its training and decision-making processes [10]. Few articles explored such techniques (Supplementary Table 1) [42–44]. Pancreatic disorders, diseases of biliary tract, abdominal-pelvic pain, digestive neoplasms, and jaundice were identified as the most common risk factors.

Table 3 presents a list of best practice recommendations for AI/ML model development to predict PC early using EHR data.

A limitation of this review was potentially missing studies that could be relevant. We excluded studies if they were written in non-English. Another limitation of this study is the sample size for different groups of models in the figures and analysis. For instance, we only have one Model C sample that performed external validation as seen in Figure 5. Therefore, it is important to consider the sample size when interpreting the results. The strength of this study is that we critically appraised the studies utilizing guidelines provided in the CHARMS checklist. Another strength of our study is that we did not limit our analysis to specific ML/AI modeling techniques. Our comprehensive review and discussion of model development, evaluation, and explainability strategies could guide future research studies attempting to develop PC risk prediction models and efforts on novel risk factor identification utilizing EHR data.

Real-world utilization of the models developed in these studies was limited. Only two of the studies conducted a prospective validation after model development [25, 35]. Multiple studies have considered identifying individuals at high risk, provided a decision curve, or reported model performance by thresholding predicted risks by the models in the validation cohort [22–25, 27–29, 31, 32, 35, 37, 40, 42–46]. None of the studies reported an integration of their model into the EHR or to identify high risk individuals in a real-world setting; in the authors' opinion this is appropriate since all of the algorithms potentially require further external model validation before being ready for this.

In conclusion, through this systematic review, we found that several studies have attempted to develop ML models using EHR data to predict PC risk with some success. However, it was observed that most studies utilized a curated set of predictors instead of utilizing unbiased approaches within the EHR. Logistic regression was the most common modeling

technique. Lack of reporting on missing data was also common and a significant limitation. Novel risk factor identification was conducted in only six studies. We believe that utilization of longitudinal structured and unstructured data together in a population-based cohort coupled with utilization of X-AI techniques may identify novel PC risk factors and should be important considerations in future studies. We also recommend using the TRIPOD statement to report prediction model development and validation method details [62]. Finally, for PC risk modeling strategy, it is crucial to evaluate the modeling assumptions and ensure collaboration across a spectrum of content expertise, including physicians,epidemiologists, biostatisticians, data scientists, and AI/ML experts. Such multidisciplinary collaborative efforts will help develop the most effective model for early prediction of PC risk by judiciously utilizing the available EHR data while minimizing biased estimates, inefficient models, and incorrect conclusions.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGEMENTS:

We thank Larry J. Prokop for his support in article search and Karen A. Doering and Kathleen J. Johnson for their administrative support.

**Financial support:**

This study was supported by research funding from the Centene Foundation to Dr. Majumder. Dr. Majumder is also supported by U01 CA210138, National Cancer Institute. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

## Abbreviations:

| | |
|---|---|
| **EHR** | Electronic Health Record |
| **NLP** | Natural Language Processing |
| **AI** | Artificial Intelligence |
| **DL** | Deep Learning |
| **ML** | Machine Learning |
| **X-AI** | Explainable-Artificial Intelligence |
| **PC** | Pancreatic Cancer |
| **PDAC** | Pancreatic ductal adenocarcinoma |
| **PNET** | Pancreatic neuroendocrine tumors |
| **CHARMS** | CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies |

| | |
|---|---|
| **PICO** | Population, Intervention, Comparison and Outcomes Article Search Framework |
| **ROC** | Receiver operating characteristic |

## REFERENCES:


1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424. [PubMed: 30207593]
2. Rahib L, Smith BD, Aizenberg R, et al. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Volume 74(11):2913–21. Cancer Res, 2014.
3. Ryan DP, Hong TS, Bardeesy N. Pancreatic adenocarcinoma. N Engl J Med 2014;371:1039–49. [PubMed: 25207767]
4. Kleeff J, Korc M, Apte M, et al. Pancreatic cancer. Nat Rev Dis Primers 2016;2:16022. [PubMed: 27158978]
5. Blackford AL, Canto MI, Klein AP, et al. Recent Trends in the Incidence and Survival of Stage 1A Pancreatic Cancer: A Surveillance, Epidemiology, and End Results Analysis. J Natl Cancer Inst 2020;112:1162–1169. [PubMed: 31958122]
6. Owens DK, Davidson KW, Krist AH, et al. Screening for Pancreatic Cancer: US Preventive Services Task Force Reaffirmation Recommendation Statement. Jama 2019;322:438–444. [PubMed: 31386141]
7. Sawhney MS, Calderwood AH, Thosani NC, et al. ASGE guideline on screening for pancreatic cancer in individuals with genetic susceptibility: summary and recommendations. Gastrointest Endosc 2022;95:817–826. [PubMed: 35183358]
8. Aslanian HR, Lee JH, Canto MI. AGA Clinical Practice Update on Pancreas Cancer Screening in High-Risk Individuals: Expert Review. Gastroenterology 2020;159:358–362. [PubMed: 32416142]
9. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association 2018;25:1419–1428. [PubMed: 29893864]
10. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel) 2020;23.
11. The EndNote Team. EndNote. EndNote 20 ed. Philadelphia, PA: Clarivate, 2013.
12. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. 2014;11(10):e1001744.
13. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019;170:51–58. [PubMed: 30596875]
14. Fernandez-Felix BM, López-Alcalde J, Roqué M, et al. CHARMS and PROBAST at your fingertips: a template for data extraction and risk of bias assessment in systematic reviews of predictive models. BMC Medical Research Methodology 2023;23:44. [PubMed: 36800933]
15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Systematic Reviews 2021;10:89. [PubMed: 33781348]
16. Chauhan R, Kaur H, Sharma S. A feature based approach for medical databases, In 2016 International Conference on Advances in Information Communication Technology and Computing, AICTC 2016, Association for Computing Machinery, 2016.
17. Manias G, Op Den Akker H, Azqueta A, et al. IHELP: Personalised Health Monitoring and Decision Support Based on Artificial Intelligence and Holistic Health Records, In 26th IEEE Symposium on Computers and Communications, ISCC 2021, Institute of Electrical and Electronics Engineers Inc., 2021.

18. Matchaba S, Fellague-Chebra R, Purushottam P, et al. Early Diagnosis of Pancreatic Cancer via Machine Learning Analysis of a National Electronic Medical Record Database. JCO Clinical Cancer Informatics 2023;7:e2300076. [PubMed: 37816199]
19. Chen W, Zhou B, Jeon CY, et al. Machine learning versus regression for prediction of sporadic pancreatic cancer. Pancreatology 2023;23(4):396–402. [PubMed: 37130760]
20. Ahmed AE, Alzahrani FS, Gharawi AM, et al. Improving risk prediction for pancreatic cancer in symptomatic patients: a Saudi Arabian study. Cancer management and research 2018;10:4981–4986. [PubMed: 30464598]
21. Baecker A, Kim S, Risch HA, et al. Do changes in health reveal the possibility of undiagnosed pancreatic cancer? Development of a risk-prediction model based on healthcare claims data. PLoS ONE [Electronic Resource] 2019;14:e0218580. [PubMed: 31237889]
22. Boursi B, Finkelman B, Giantonio BJ, et al. A Clinical Prediction Model to Assess Risk for Pancreatic Cancer Among Patients With New-Onset Diabetes. Gastroenterology 2017;152:840–850 e3. [PubMed: 27923728]
23. Chen W, Butler RK, Zhou Y, et al. Prediction of Pancreatic Cancer Based on Imaging Features in Patients With Duct Abnormalities. Pancreas 2020;49:413–419. [PubMed: 32132511]
24. Chen W, Butler RK, Lustigova E, et al. Risk Prediction of Pancreatic Cancer in Patients With Recent-onset Hyperglycemia: A Machine-learning Approach. Journal of Clinical Gastroenterology 2023;57:103–110. [PubMed: 35470312]
25. Chen W, Zhou Y, Xie F, et al. Derivation and External Validation of Machine Learning-Based Model for Detection of Pancreatic Cancer. The American Journal of Gastroenterology 2023;118:157–167. [PubMed: 36227806]
26. Dayem Ullah AZM, Stasinos K, Chelala C, et al. Temporality of clinical factors associated with pancreatic cancer: a case-control study using linked electronic health records. BMC Cancer 2021;21:1279. [PubMed: 34837975]
27. Jeon CY, Chen Q, Yu W, et al. Identification of Individuals at Increased Risk for Pancreatic Cancer in a Community-Based Cohort of Patients With Suspected Chronic Pancreatitis. Clinical and Translational Gastroenterology 2020;11:e00147. [PubMed: 32352677]
28. Klein AP, Lindstrom S, Mendelsohn JB, et al. An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. PLoS One 2013;8:e72311. [PubMed: 24058443]
29. Li X, Gao P, Huang C-J, et al. A Deep-Learning Based Prediction of Pancreatic Adenocarcinoma with Electronic Health Records from the State of Maine. International Journal of Medical and Health Sciences 2020;14:358–365.
30. Malhotra A, Rachet B, Bonaventure A, et al. Can we screen for pancreatic cancer? Identifying a sub-population of patients at high risk of subsequent diagnosis using machine learning techniques applied to primary care data. PLoS ONE [Electronic Resource] 2021;16:e0251876. [PubMed: 34077433]
31. Muhammad W, Hart GR, Nartowt B, et al. Pancreatic Cancer Prediction Through an Artificial Neural Network. Frontiers in Artificial Intelligence 2019;2.
32. Munigala S, Singh A, Gelrud A, et al. Predictors for Pancreatic Cancer Diagnosis Following New-Onset Diabetes Mellitus. Clin Transl Gastroenterol 2015;6:e118. [PubMed: 26492440]
33. Park J, Artin MG, Lee KE, et al. Deep learning on time series laboratory test results from electronic health records for early detection of pancreatic cancer. Journal of Biomedical Informatics 2022;131:104095. [PubMed: 35598881]
34. Risch HA, Yu H, Lu L, et al. Detectable Symptomatology Preceding the Diagnosis of Pancreatic Cancer and Absolute Risk of Pancreatic Cancer Diagnosis. American Journal of Epidemiology 2015;182:26–34. [PubMed: 26049860]
35. Sharma A, Kandlakunta H, Nagpal SJS, et al. Model to Determine Risk of Pancreatic Cancer in Patients With New-Onset Diabetes. Gastroenterology 2018;155:730–739.e3. [PubMed: 29775599]
36. Stapley S, Peters TJ, Neal RD, et al. The risk of pancreatic cancer in symptomatic patients in primary care: a large case-control study using electronic records. Br J Cancer 2012;106:1940–4. [PubMed: 22617126]

37. Yu A, Woo SM, Joo J, et al. Development and Validation of a Prediction Model to Estimate Individual Risk of Pancreatic Cancer. PLOS ONE 2016;11:e0146473. [PubMed: 26752291]
38. Zhao X, Lang R, Zhang Z, et al. Exploring and validating the clinical risk factors for pancreatic cancer in chronic pancreatitis patients using electronic medical records datasets: three cohorts comprising 2,960 patients. Translational Cancer Research 2020;9:629–638. [PubMed: 35117408]
39. Chen S-M, Phuc PT, Nguyen P-A, et al. A novel prediction model of the risk of pancreatic cancer among diabetes patients using multiple clinical data and machine learning. Cancer Medicine 2023;12:19987–19999. [PubMed: 37737056]
40. Appelbaum L, Cambronero JP, Stevens JP, et al. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study. European Journal of Cancer 2021;143:19–30. [PubMed: 33278770]
41. Rasmy L, Xiang Y, Xie Z, et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine 2021;4(1) (no pagination).
42. Placido D, Yuan B, Hjaltelin JX, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nature Medicine 2023;29:1113–1122.
43. Chen Q, Cherry DR, Nalawade V, et al. Clinical Data Prediction Model to Identify Patients With Early-Stage Pancreatic Cancer. JCO Clinical Cancer Informatics 2021;5:279–287. [PubMed: 33739856]
44. Salvatore M, Beesley LJ, Fritsche LG, et al. Phenotype risk scores (PheRS) for pancreatic cancer using time-stamped electronic health record data: Discovery and validation in two large biobanks. Journal of Biomedical Informatics 2021;113:103652. [PubMed: 33279681]
45. Zhu W, Aphinyanaphongs Y, Kastrinos F, et al. Identification of patients at risk for pancreatic cancer in a 3-year timeframe based on machine learning algorithms. medRxiv. 2023;06.
46. Jia K, Kundrot S, Palchuk MB, et al. A pancreatic cancer risk prediction model (Prism) developed and validated on large-scale US clinical data. EBioMedicine 2023;98:104888. [PubMed: 38007948]
47. Park J, Artin MG, Lee KE, et al. Structured deep embedding model to generate composite clinical indices from electronic health records for early detection of pancreatic cancer. Patterns 2023;4:100636. [PubMed: 36699740]
48. Hidalgo M, Cascinu S, Kleeff J, et al. Addressing the challenges of pancreatic cancer: future directions for improving outcomes. Pancreatology 2015;15:8–18. [PubMed: 25547205]
49. Pottegård A Core concepts in pharmacoepidemiology: Fundamentals of the cohort and case-control study designs. Pharmacoepidemiol Drug Saf 2022;31:817–826. [PubMed: 35621007]
50. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nature Reviews Cancer 2004;4:309–314. [PubMed: 15057290]
51. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Advances in neural information processing systems 2016;29.
52. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems 2017;30:6000–6010.
53. Rasmy L, Tiryaki F, Zhou Y, et al. Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. Journal of the American Medical Informatics Association 2020;27:1593–1599. [PubMed: 32930711]
54. Royston P, Moons KG, Altman DG, et al. Prognosis and prognostic research: Developing a prognostic model. Bmj 2009;338:b604. [PubMed: 19336487]
55. Donders ART, van der Heijden GJMG, Stijnen T, et al. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology 2006;59:1087–1091. [PubMed: 16980149]
56. Janssen KJM, Donders ART, Harrell FE, et al. Missing covariate data in medical research: To impute is better than to ignore. Journal of Clinical Epidemiology 2010;63:721–727. [PubMed: 20338724]
57. Vergouwe Y, Royston P, Moons KGM, et al. Development and validation of a prediction model with missing predictor data: a practical approach. Journal of Clinical Epidemiology 2010;63:205–214. [PubMed: 19596181]

58. Che Z, Purushotham S, Cho K, et al. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific Reports 2018;8:6085. [PubMed: 29666385]

59. Chen RT, Rubanova Y, Bettencourt J, et al. Neural ordinary differential equations. Advances in neural information processing systems 2018;31:6572–6583.

60. Pannala R, Leibson CL, Rabe KG, et al. Temporal association of changes in fasting blood glucose and body mass index with diagnosis of pancreatic cancer. Am J Gastroenterol 2009;104:2318–25. [PubMed: 19513024]

61. Sah RP, Sharma A, Nagpal S, et al. Phases of Metabolic and Soft Tissue Changes in Months Preceding a Diagnosis of Pancreatic Ductal Adenocarcinoma. Gastroenterology 2019;156:1742–1752. [PubMed: 30677401]

62. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55–63. [PubMed: 25560714]

63. Yang Z, Mitra A, Liu W, et al. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nature communications 2023;14:7857.

## STUDY HIGHLIGHTS:

**WHAT IS KNOWN:**

- Pancreatic cancer (PC) is often diagnosed at an advanced stage when treatment options are limited.
- PC detection at an early stage can improve survival.
- Artificial Intelligence (AI) -based models have been developed to predict pancreatic cancer utilizing Electronic Health Records (EHR).
- There is limited guidance on the optimal selection of modeling techniques, study design, and utilization of EHR data for PC prediction.

**WHAT IS NEW HERE:**

- The review provides recommendations for optimal ML/AI modeling approaches to utilizing EHR data for PC prediction.
- Underutilization of EHR data, sparse use of advanced AI methods, and limited experimentation with data exclusion intervals, were some of the major limitations.
- Efforts on identifying novel risk factors to predict PC from EHR are currently limited.
- Non-linear and deep learning-based AI models were found to perform similar to traditional linear statistical and ML models in predicting PC.
- Deep learning models generally utilized a wide range of candidate predictors, instead of a set of curated known risk factors for PC.

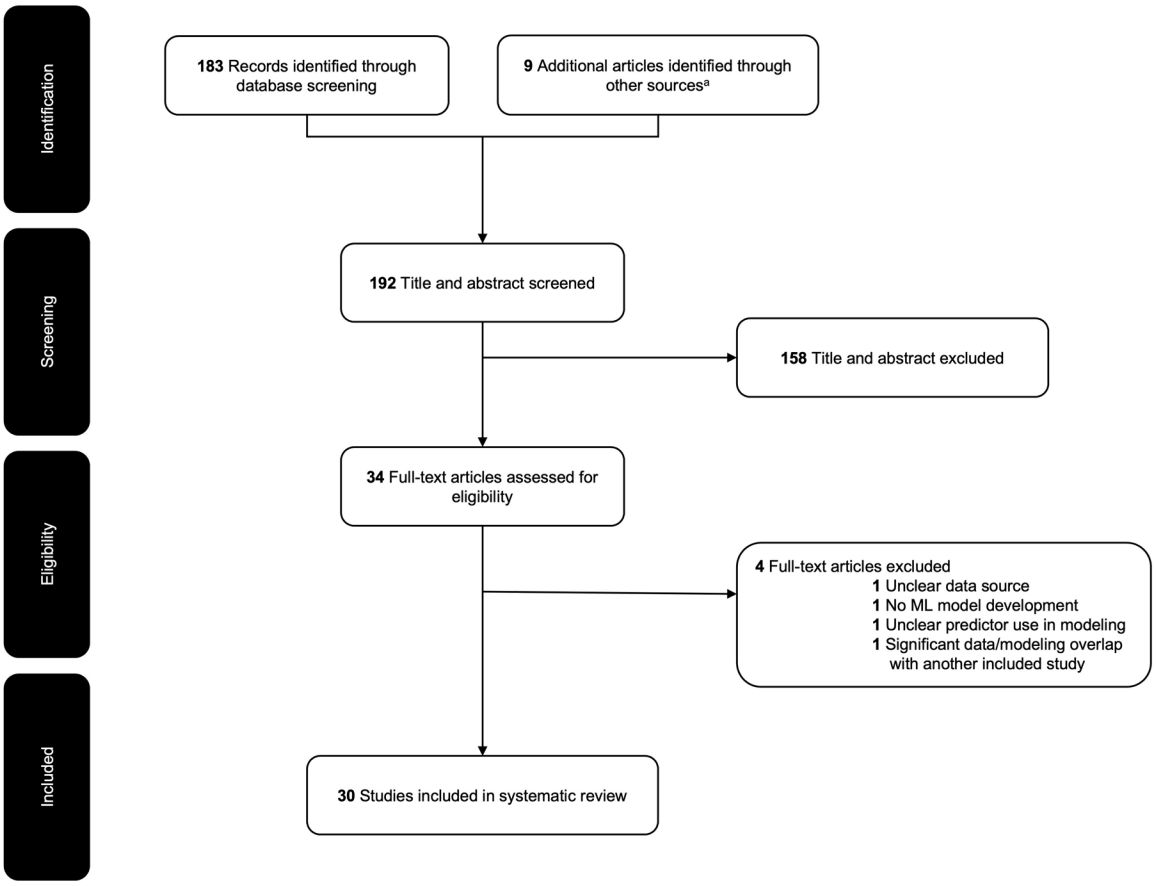


**Figure 1.**
Systematic review flow diagram — selection of articles

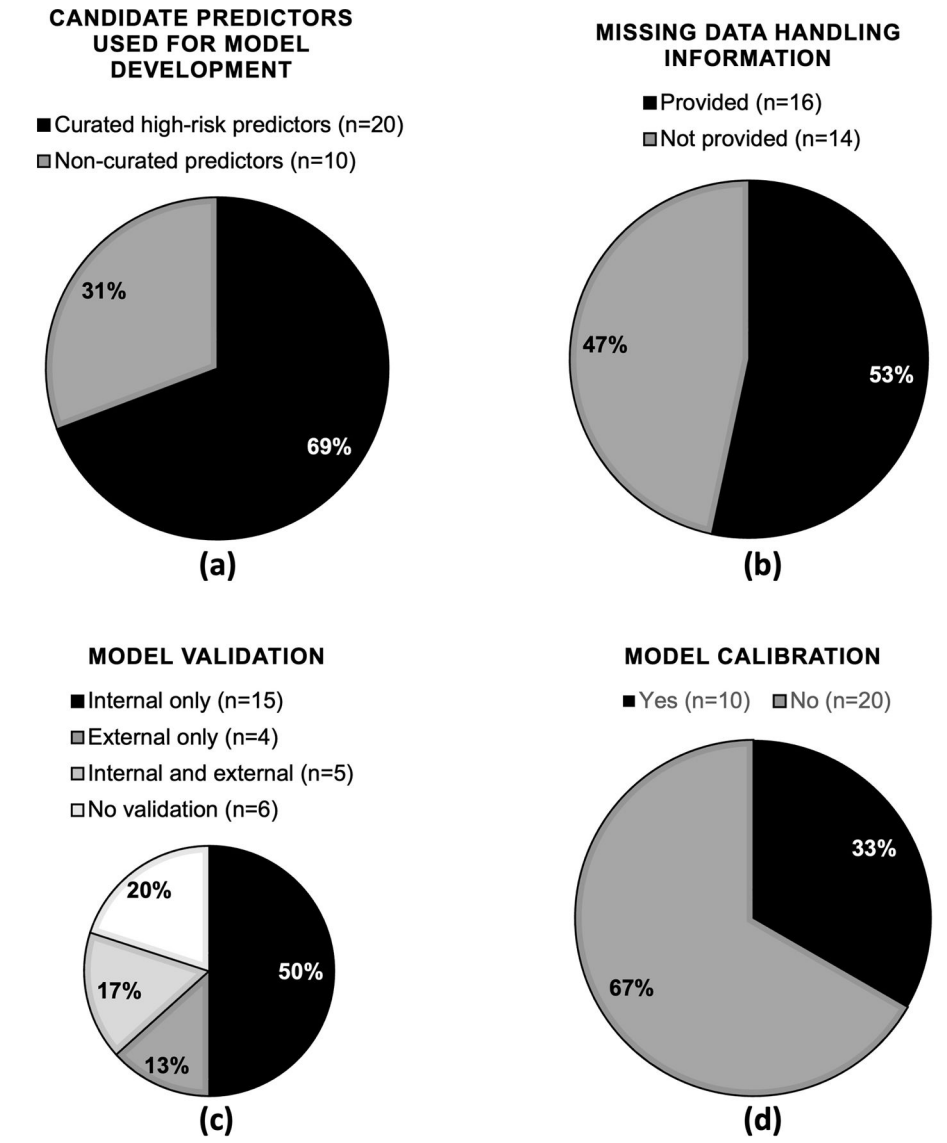


**Figure 2.**

Study and ML/AI modeling characteristics - (a): EHR candidate predictors used for model development by the studies. (b): Missing data reported by the studies. (c): Model validation conducted by the studies. (d): Model calibration conducted by the studies.

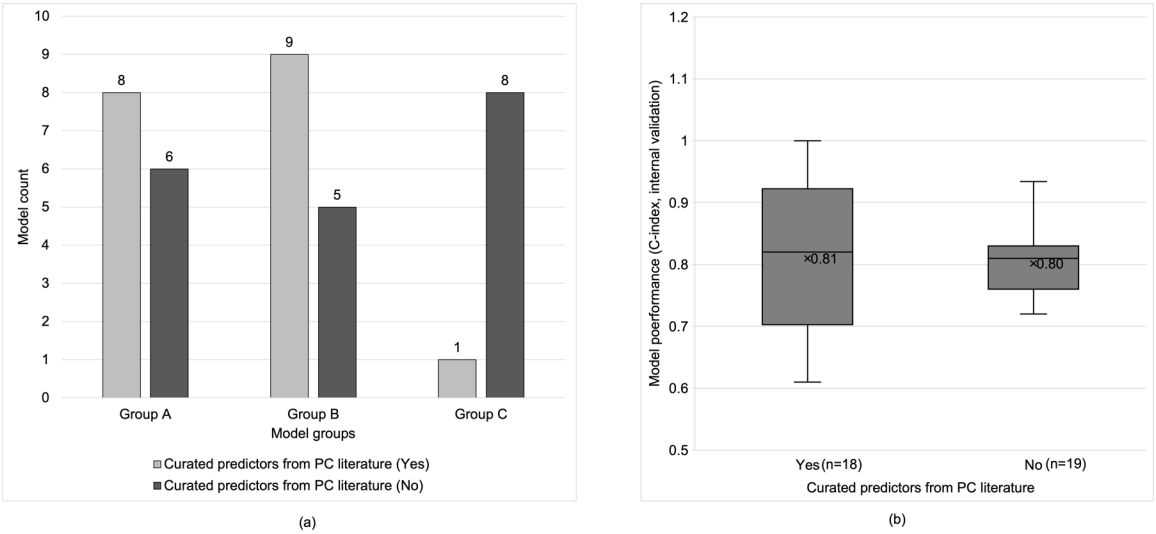


**Figure 3.**

Use of curated risk factors by models. (a) Number of models with and without using curated risk factors per model groups (A: linear, B: non-linear excluding deep learning models, and C: deep learning models only). (b) Performance of models in internal validations with and without using curated risk factors of PC from literature.

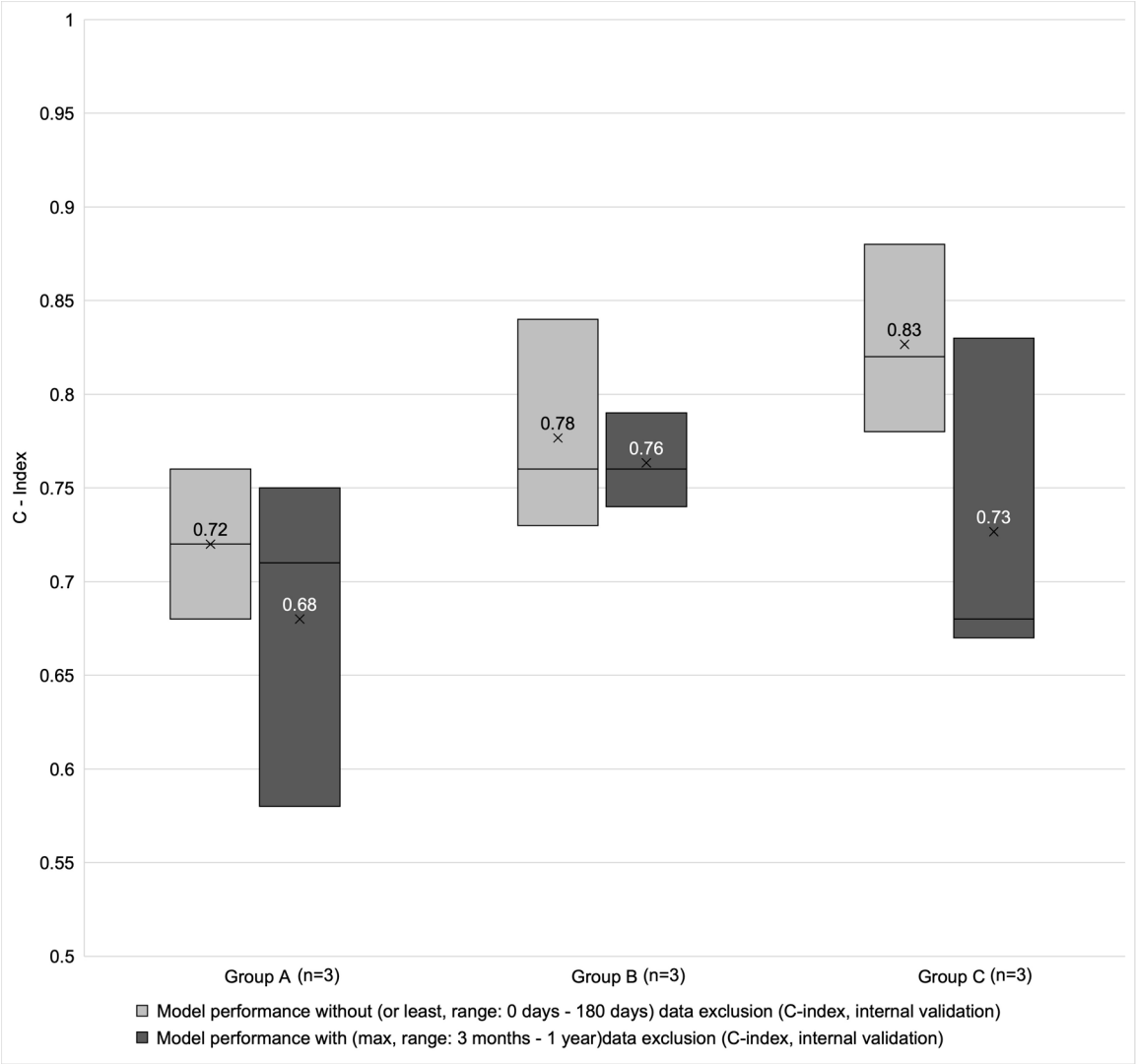


**Figure 4.**

Model performance by groups (A: linear, B: non-linear excluding deep learning models, and C: deep learning models only) with and without data exclusion intervals before diagnosis.

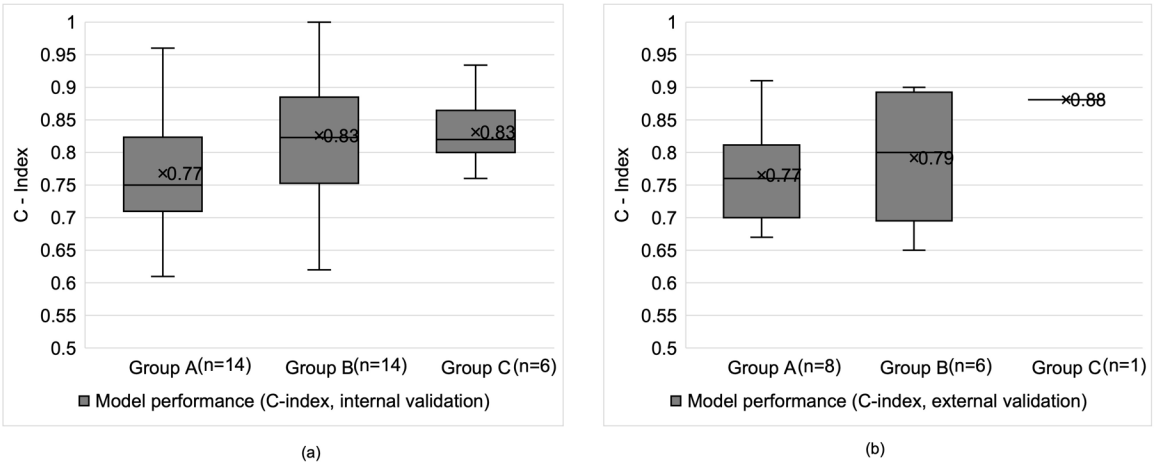


**Figure 5.**

Internal and external validation model performances by groups (A: linear, B: non-linear excluding deep learning models, and C: deep learning models only).

**Table 1.**

Machine Learning based Pancreatic Cancer Prediction Study Characteristics

| Study ID | Study Type (Time Period) | Participants | Study Outcome | Candidate Predictors |
|---|---|---|---|---|
| AhmedAE-2018[20] | Retrospective cohort study (January 2013 to December 2016) | Patients suspected as having pancreatic cancer (PC) and underwent biopsy examination of pancreatic mass; 206 total patients, 87 benign and 119 malignant | Biopsy-confirmed PC | Age, gender, body mass index (BMI), symptoms (nausea/vomiting, jaundice, weight loss, dark urine, persistent fatigue, abdominal pain, back pain, bowel obstruction, blood clots, pancreatitis), comorbidities (diabetes, hypertension, depression, renal disease), abnormal imaging findings |
| AppelbaumL-2021[40] | Case control study (July 1997 to December 2017 (BIDMC) and 1979 to 2017 (PHC)) | Patients of all ages with pancreatic ductal adenocarcinoma (PDAC); cross-checked against the BIDMC tumour registry; Training set: 594 cases, 100,787 controls, Validation set: 408 cases, 160,185 controls | International Classification of Diseases (ICD) 9 and 10 codes for PDAC | 4,150 diagnoses based on ICD 9 and 10 codes that are seen at least 100 times in training data |
| BaeckerA-2019[21] | Case control study (2004 to 2011) | Newly diagnosed patients with PDAC over age 68; Study only included people with PDAC that was confirmed by microscopy, laboratory test, direct visualization, or imaging, and excluded cases with unknown months of diagnoses or those diagnosed at autopsy; 29,646 cases, 88,938 age and sex-matched controls | PDAC confirmed by SEER topographic C25.x and ICD-O-3 histology codes for adenocarcinoma of the pancreas (8000, 8010, 8020, 8021, 8022, 8050, 8140, 8141, 8211, 8230, 8260, 8441, 8450, 8453, 8470, 8471, 8472, 8473, 8480, 8481, 8500, 8503, 8521). | 16 risk factors extracted from Medicare claims data: acute pancreatitis, chronic pancreatitis, any abdominal pain, chest pain, diabetes mellitus, weight loss/anorexia/cachexia, nausea and/or vomiting, digestive problems, dyspepsia/gastritis/peptic ulcer disease, fatigue, itching/pruritis, depression, jaundice, gallbladder disease, acute cholecystitis, and esophageal reflux |
| BoursiB-2017[22] | Retrospective cohort study (1995 to 2013) | Individuals with incident diabetes after the age of 35 years and 3 or more years of follow-up after diagnosis of diabetes; 109,385 patients, 390 cases diagnosed with PDAC within 3 years of diabetes diagnosis | Incident diagnosis of PDAC (defined according to diagnostic Read codes) within 3 years after the diagnosis of diabetes mellitus | PDAC risk factors as well as variables related to glucose metabolism (54 candidate predictors in total) including anthropometric variables, lifestyle factors, medical comorbidities, medications, and laboratory studies |
| ChenQ-2021[43] | Case control study (2009 to 2017) | Patients with PC diagnosed at age 40 and over; 3,322 cases with early-stage PC, and 25,908 cases with late-stage PC, and 7,039,056 controls | PC diagnosis with two or more ICD 9 or 10 in single calendar year | 18,220 variables generated from EHR data; final model consisted of 582 variables. |
| ChenSM-2023[39] | Retrospective cohort study (January 1, 2009, to December 31, 2019) | Individuals diagnosed with diabetes (International Classification of Diseases (ICD)-9:250; ICD-10:E11). excluded individuals under the age of 40 years, patients with Type I DM (T1DM), and those who had previously been diagnosed with pancreatic cancer (ICD-O-3:C25) prior to a Type 2 DM (T2DM) diagnosis. 66,384 total patients, 89 PDAC cases. | ICD-O-3 code C25 | Demographic characteristics (i.e., gender, age, and body mass index (BMI); comorbidities before the prescription date of antidiabetic drugs (i.e., cardiovascular, chronic obstructive pulmonary, and rheumatic diseases) and the Charlson comorbidity index (CCI) score; long-term medications (i.e., antacids, gastroesophageal reflux disease (GORD), and gastrointestinal disorder agents) are prescribed during the 6 months before a prescription for an antidiabetic drug; laboratory test results (i.e., glycated hemoglobin (HbA1c), glucose AC, and albumin) within 12 months before prescription of an antidiabetic drug. |

| Study ID | Study Type (Time Period) | Participants | Study Outcome | Candidate Predictors |
|---|---|---|---|---|
| ChenW-2020[23] | Retrospective cohort study (Jan 2006 to June 2016) | Patients who (1) were <18 years of age, (2) had a clearly defined mass (>2 cm) in the pancreas or a history of PDAC (cancer site code C25.0-25.9 in the Cancer Registry (CR) of the organization) on or prior to index date,or (3) were not continuously enrolled in the health plan in the 12 months prior to index date, were excluded; adult patients with CT or MRI with dilated main pancreatic duct, identified based on radiology reports using natural language processing (NLP); 7,819 total patients, 781 developed PDAC within 3 years of which 712 (91%) and 756 (97%) were diagnosed within 1 and 2 years, respectively | PDAC defined by Site Codes C25.x in the Cancer Registry of the organization, or ICD 10-CM codes C25.x as the cause of death in State Death Master Files | Pancreas morphological features including atrophy, calcification, pancreatic cyst, pancreatic ductal irregularity, focal pancreatic duct stricture with distal (upstream) dilatation, focal pancreatic side branch dilation, granular pancreatic duct filling defects, as well as intra-ductal calculi (duct stone); clinical/demographic features including age, sex, race/ethnicity, tobacco and alcohol use, medical insurance type, years since health plan enrollment, neighborhood educational level (% of population with high school completion), family history of pancreatic cancer, diabetes, acute and chronic pancreatitis, dyspepsia, gallstone disorders, depression, insulin resistance (ICD-9 code 790.29 or ICD-10 code R73.03) and weight change in the 12 months prior to the index date were captured; laboratory measures including fasting glucose, hemoglobin A1c, creatinine, cholesterol, alanine transaminase, aspartate aminotransferase, alkaline phosphatase, bilirubin, total protein, conjugated bilirubin, and albumin |
| ChenW-2023 (AJG)[25] | Retrospective cohort study (2008 to 2017) | Patients age 50–84 with at least one clinic-based visit; Kaiser: 1,801,931 patients, 1,792 patients developed PDAC; VA: 2,633,112 patients, 4,582 patients developed PDAC | PDAC diagnosis based on ICD-10-CM or histology codes identified from Kaiser and VA Cancer Registries; pancreatic cancer deaths identified through California State Death Master Files and VA Mortality Data Repository respectively | More than 500 clinical predictors including demographics and lifestyle variables (e.g., smoking status), medical conditions (coded by Ninth Revision of International Classification of Diseases or ICD-10 codes), laboratory test values, medication dispensing, medical procedures (coded by CPT, Ninth Revision of International Classification of Diseases/ICD-10, or KPSC internal procedure codes), symptoms (e.g., abdominal pain), healthcare utilization, and other features (e.g., year of index visit) |
| ChenW-2023 (JCG)[24] | Retrospective cohort study (January 2010 to September 2018) | 50 to 84 years of age who had an elevated (6.5% +) glycated hemoglobin (HbA1c) with recent-onset hyperglycemia; 109,266 patients, 319 PDAC cases | PDAC or death with pancreatic cancer in the 3 years after the index date captured by ICD 10 and histology codes | 102 candidate predictors including demographics, clinical characteristics, and symptoms before index date |
| DayemUllahAZM-2021[26] | Case control study (April 2008 to March 2020) | Patients with at least one hepato-pancreaticobiliary disease or control group disease recorded in secondary care EHR; 965 cases of PC, 3,963 cases with non-malignant pancreatic disease, 4,355 controls | Participants divided into incidence outcome groups – PC, non-malignant pancreas disease, and controls per ICD10, SNOMED CT, Read V2 or CTV3 codes or GP records during the observation period | 19 clinicodemographic factors: gender, ethnicity, age group, diabetes, hypertension, hyperlipidemia, cardiovascular disease, chronic respiratory disease, chronic renal disease, acute pancreatic disease, chronic pancreatic disease, chronic biliary disease, chronic liver disease, upper GI disease, lower GI disease, smoker, alcohol drinker, substance user, obesity |
| JeonCY-2020[27] | Retrospective cohort study (2006–2015) | Patients with suspected chronic pancreatitis identified by diagnostic code and at least one abnormal pancreatic finding on radiographic imaging and who had survived at least one year without PC; 1,766 patients, 46 cases | PC cases with at least 2 outpatient or inpatient visits with ICD-9 code, patients registered in the internal cancer registry as having a malignant neoplasm in the pancreas, and those with PC as cause of death in the Death | Imaging features (parenchymal calcification, ductal stones, glandular atrophy, pseudocyst, main duct dilatation, duct irregularity, abnormal side branch, or stricture), age, gender, race, alcohol, smoking, BMI, history of acute pancreatitis, diabetes |

| Study ID | Study Type (Time Period) | Participants | Study Outcome | Candidate Predictors |
|---|---|---|---|---|
| | | | Index, confirmed by manual chart review | |
| JiaK-2023[46] | Case control study (All data available before Dec 2022 in the federated EHR database of TriNetX) | 35,387 PDAC cases, 1,500,081 controls | ICD-10/ICD-9 codes: C25.0, C25.1, C25.2, C25.3, C25.7, C25.8, C25.9, and 157 | 87 model predictors including demographic features, diagnosis, medications, and labs selected automatically from over 5,000 EHR features using L0 regularisation and iterative feature removal. |
| KleinAP-2013[28] | Case control study (1985 to 2002) | Non-Hispanic white patients of European ancestry with diabetes diagnosed earlier than 3 years of PC diagnosis; 3,349 cases, 3,654 controls | Outcome of PC not clearly defined | Age, sex, ethnicity, current smoking, diabetes, BMI, heavy alcohol consumption, family history of pancreatic cancer, and GWAS-identified risk markers including ABO blood group |
| LiX-2020[29] | Retrospective cohort study (January 2014 to March 2018) | Patients age 35 and older who visited Maine Healthcare Facilities; 265,225 patients, 4,361 cases | PC defined by ICD-10 codes C25.0-C25.9 | 233 predictors from demographics, admission information, vital signs, measurements, laboratory tests, medications, and diagnoses of chronic conditions data |
| MalhotraA-2021[30] | Case control study (January 2005 to June 2009) | Patients between 15–99 years of age with PC diagnosis with 2 years of data before diagnosis. Controls were patients diagnosed with unrelated primary cancer 18 months after index date. Excluded patients diagnosed with cancers of the lip, oral cavity and pharynx (ICD-10 codes C00-14), digestive organs (C15-26), respiratory and intrathoracic organs (C30-39), breast (C50) and female genital organs (C51-58); 1,139 cases, 4,556 controls | Primary pancreatic tumor with ICD-10 code C25 | 57 symptoms and health statuses associated with medical or product (drug) codes including cardiovascular, circulatory system, digestive, endocrine and metabolic, genitourinary, hematological, immunological, and oncological disorders, diseases of the musculoskeletal system and connective tissue, infections, nervous system, medications, general clinical symptoms such as weight loss and fever, and health behaviors including history of smoking and heavy drinking |
| MuhammadW-2019[31] | Retrospective cohort study (1993 to 2017) | 800,114 participants with 898 PC cases from National Health Interview Survey (NHIS) and Pancreatic, Lung, Colorectal, and Ovarian cancer (PLCO) datasets | Outcome of PC not clearly defined | age, diabetes age, smoking age, years quit, pack-years of smoking, vigorous exercise, moderate exercise, drinking frequency, drinking amount, binging frequency, family members with PC, family members over age 50 with PC, BMI, gender, emphysema, asthma, stroke, coronary heart disease, angina pectoris, heart attack, other heart disease, ulcer, drink, other cancer, hypertension, Hispanic, diabetes, smoking status, smoking frequency, and race |
| MunigalaS-2015[32] | Retrospective cohort study (Fiscal Year 1998) | Age over 40 and diagnosis of PC after diagnosis of diabetes was included; final cohort contained 452,804 patients, and new onset of diabetes cohort contained 73,811 patients with 234 PC patients (183 PC cases in less than or equal to 3 years); non diabetic patient cohort contained 378,993 patients with 858 PC patients (434 PC cases in less than or equal to 3 years) | PC - presence of two or more ICD-9 codes 157.0, 157.1, 157.2, 157.3, and 157.9 less than 1 year apart | CP (ICD-9 code 577.1), history of obesity (ICD-9 code 278.0), history of smoking (nicotine dependence, ICD-9 codes 305.1 or V15.82), presence of gallstones (ICD codes 574, 574.1, 574.3, 574.5, 574.7, 574.8, or 574.9) defined on the basis of 41 ICD codes before PC diagnosis or censorship, age at the time of entry into the study, race, and sex |
| ParkJ-2022[33] | Case control study (2004 to 2021) | Patients who met the following criteria: ICD code for smoking, obesity, diabetes, or chronic pancreatitis; underwent a CT, MRI or MRCP; had a pathology report containing both the terms “pancrea” and any one of the terms “malignan, carcinoma, cancer, | PDAC diagnosis based on ICD-9, ICD-10, or histology code | 418 of the most clinically relevant lab variables were identified by human experts, 33 selected based on data completeness |

| Study ID | Study Type (Time Period) | Participants | Study Outcome | Candidate Predictors |
|---|---|---|---|---|
| | | neoplas" 834 cases; 8,223 controls without PC ICD Codes | | |
| ParkJ-2023[47] | Case control study (2004 to 2021) | Patients who met the following criteria: ICD code for smoking, obesity, diabetes, or chronic pancreatitis; underwent a CT, MRI or MRCP; had a pathology report containing both the terms "pancrea" and any one of the terms "malignan, carcinoma, cancer, neoplas"; 834 cases, 8,223 controls controls without PC ICD | PDAC diagnosis based on ICD-9, ICD-10, or histology code | 206 final lab variables obtained from 6,392 unique variables through domain expertise and removing redundancies. |
| PlacidoD-2021[42] | Retrospective cohort study (January 1977 to April 2018) | All patients with at least five recorded diagnosis codes in the Danish Registry and United States Veterans Affairs (US-VA) Corporate Data Warehouse; 6.2 million patients with 23,985 cases from Danish Registry; 2.0 million patients with 3,418 cases from US-VA | PC diagnosis defined based on ICD-8 code 157 and ICD-10 code C25 | More than 2,000 ICD disease codes |
| RamsyL-2021[41] | Case control study (2000 to 2017 (Cerner Dataset) and 2011 to 2015 (Truven dataset)) | Patients with PC diagnosed at age 45 and over did not report any other cancer disease before their first PC diagnosis (Cerner) 11,486 cases, 17,919 controls | PC diagnosis based on ICD-9 codes that start with 157 and ICD-10 codes that start with C25 | Diagnosis codes based on ICD 9 and 10 (26,427 codes) |
| RasmyL-2020[53] | Case control study (Over fifteen years, Cerner HealthFacts dataset version 2017) | Patients with pancreatic cancer diagnosed at age 45 and over and did not report any other cancer diseases before their first PC diagnosis; 11,486 cases and matched 17,919 controls | Pancreatic cancer diagnosis based on ICD-9 and 10 codes | 17,629 ICD-9 codes, 94,044 ICD-10-CM codes, and 16,044 ICD-10-CA codes |
| RischHA-2015[34] | Case control study (January 2005 to June 2009) | Cases were 35–83 years old individuals with newly diagnosed PC; 362 case and 690 control subjects with blood samples were considered in final analysis | Outcome of PC was confirmed through examination of clinical or pathology records | Jewish ancestry, ABO blood group, diagnosis of diabetes mellitus, time since diabetes diagnosis, diagnosis of pancreatitis, time since pancreatitis diagnosis, current cigarette smoking, time since quitting smoking, current use of proton pump inhibitors (PPI), time since starting use of PPIs |
| SalvatoreM-2021[44] | Case control study (2006 to 2010 UK Biobank Study (UKB); dates of diagnoses not available for Michigan Genomics Initiative (MGI)) | Patients with PC of inferred, recent European ancestry; MGI dataset had 429 cases and 37,930 controls; UKB dataset contained 659 cases and 392,640 controls | PC diagnosis, based on the PC phecode, constructed using ICD9 codes 157, 157.1, 157.2, 157.3, 157.4, 157.8, 157.9 and ICD10 codes C25, C25.0, C25.1, C25.2, C25.3, C25.4, C25.7, C25.8, C25.9 | Age, sex, genotyping array, first four principal components of genotype data, BMI (continuous), alcohol (ever vs. never), and smoking status (ever vs. never), polygenic risk score, and phenotype risk score constructed using 1,683 unique phenotype codes (developed by grouping clinically relevant diagnosis codes in the EHR) |
| SharmaA-2018[35] | Retrospective cohort study (January 1, 2000 to December 31, 2015) | 4 independent, non-overlapping cohorts of patients greater than or equal to 50 years of age and new-onset diabetes (NOD) (based on hyperglycemia; data collected at date of diagnosis and 12 months before). Three retrospectively identified and annotated cohorts - 1) Discovery set of PC - NOD (n=64), 2) Discovery set of Type 2 Diabetes - NOD (n=192), and 3) A population based new-onset diabetes validation set (n=1,096, 9 patients had PC within 3 years of NOD). | PC diagnosis was manually verified to exclude mimickers including ampullary cancer, islet cell cancer | Change in weight, change in blood glucose, change in blood glucose category, age at new onset diabetes |

| Study ID | Study Type (Time Period) | Participants | Study Outcome | Candidate Predictors |
|---|---|---|---|---|
| | | 4) There was also a prosectively identified cohort of NOD subjects for pilot screening (n=100) | | |
| StapleyS-2012[36] | Case control study (January 2000 to December 2009) | Patients age 40 and older and cases with PC tumor with at least 1 year of data before the first diagnostic code; 3,635 cases, 16,459 controls | List of 25 PC tumor diagnostic codes collated from the General Practice Research Database master code library | Symptoms: abdominal pain, nausea, back pain, constipation, diarrhoea, weight loss, malaise; Signs: jaundice; Diseases: new-onset diabetes; Investigations: abnormal liver function, low hemoglobin, raised inflammatory markers |
| YangZ-2023[63] | Case control study (2016 to 2019) | 6,475,218 from more than 1200 health care facilities of the US VHA for pretraining transformer model. Pancreatic cancer disease prediction cohort: Cases: 4639 patients of 45 years or older with no report of any other cancer disease before their first pancreatic cancer diagnosis, Controls: 5,089 patients of 45 years or older without any cancer diagnosis. | New onset pancreatic cancer, ICD 10 code C25 | Demographic information and ICD-10CM codes as predictors. Demographic information includes gender, age, race, and marital status. |
| YuA-2016[37] | Retrospective cohort study (1996 to 1997 (KCCR), 1998 to 1999 (NHIC)) | Korean Central Cancer Registry (KCCR): 1,289,933 men and 557,701 women age 30 to 80 years who had no history of any cancer at baseline and during the first two years of follow-up and without any missing values for the primary risk factors (age, height, BMI, fasting glucose, urine glucose, cholesterol, smoking, age at smoking initiation, meal preference, frequency of meat consumption, eating habits), 1,634 men and 561 women cases; National Health Insurance Corporation (NHIC) validation cohort: 500,046 men and 627,629 women free of any cancer at baseline, 711 men and 576 women cases. | PC diagnosis based on ICD 10 codes | Previous disease history (hepatitis, diabetes, and any other cancer), eating habits (bland, moderate, spicy, or salty), meal preference (meat vs. vegetables), frequency of meat intake (≤1 time/week, 2–3 times/ week, or ≥4 times/week), drinking habit (≤2–3 times/ month or ≥1–2 times/week), amount of alcohol consumed at a time, duration of smoking, amount of smoking per day (never, ever, current and <0.5 pack/day, current and ≥0.5–1 pack/day, or current and ≥1 pack/day), year of smoking cessation, physical activity (none, light, moderate, or heavy), height (grouped by quartiles), BMI (<18.5, 18.5–22.9, 23.0–24.9, or ≥25), systolic and diastolic blood pressure, total cholesterol, and fasting blood and urine glucose levels |
| ZhaoX-2020[38] | Retrospective cohort study (January 2000 to October 2015 (WFBMC), June 2000 to August 2015 (MMH), and February 2010 to October 2015 (BJCYH)) | Patients diagnosed with chronic pancreatitis (ICD 9 and 10 codes 577.1, 577.8, K86.0 and K86.1) were enrolled in the study; Derivation cohort: 2,545 patients with chronic pancreatitis, 14 with PC; Validation cohort: 415 patients with chronic pancreatitis, 7 with PC | PC diagnosis based on ICD 9 and 10 codes PC (157.0, 157.1, 157.2, 157.3, 157.4, 157.8, 157.9, C25.0, C25.1, C25.2, C25.3, C25.4, C25.7, C25.8, C25.9) | Demographic data including age, sex, race etc.; history of alcohol consumption and smoking; family history of malignancy; accompanying disease including hypertension, type II diabetes mellitus (DM), coronary heart disease; symptoms such as abdominal pain, diarrhea, loss of weight (LW); laboratory findings including routine blood examinations; and serum biochemical indexes |
| ZhuW-2023[45] | Case control study (2000 to 2021) | 1,923 pancreatic cancer cases and 7,728 matched controls | ICD 10 codes C25.0, C25.1, C25.2, C25.3, C25.7, C25.8, C25.9 | Demographics; 73 diagnosis codes and 5 lab test obtained from 19,304 diagnosis records and 10 lab tests performing PheWAS analysis. |

**Table 2.**

Machine Learning Modeling Results of the Included Studies

| Study ID | Modeling Method | Model evaluation | Data exclusion time | Prediction time window | Model performance (C-index, internal validation) | Model performance (C-index, external validation) |
|---|---|---|---|---|---|---|
| AhmedAE-2018[20] | Logistic regression | Internal validation using bootstrapping | N/A | Unclear | enriched cohort=0.96 | N/A |
| AppelbaumL-2021[40] | Logistic regression, neural network models | Internal and external validation | 180, 270, and 365 days | N/A | BIDMC Data - Cut-off 365 days: Logistic regression (LR)= 0.71, Neural Net(NN)= 0.76; Cut-off 270 days: LR=0.72, NN=0.71; Cut-off 180 days: LR=0.72, NN=0.73. PHC Data- Cut-off 365 days: LR=0.75; NN=0.74; cut-off 270 days: LR=0.76, NN=0.75; cut-off 180 days: LR=0.76, NN=0.76 | Model Trained on BIDMC and validated on PHC Data - Cut-off 365 days: LR=0.68; NN=0.6; cut-off 270 days: LR=0.68, NN=0.69; cut-off 180 days: LR=0.70, NN=0.65 |
| BaeckerA-2019[21] | Logistic regression | Internal validation using bootstrapping | 3 months | N/A | All data: 0 month exclusion= 0.68 and 3 month exclusion= 0.58; Among new-onset diabetes patients: 0 month exclusion= 0.73 and 3 month exclusion= 0.63. | N/A |
| BoursiB-2017[22] | Logistic regression | Internal validation using bootstrapping | N/A | 3 years | 0.82 | N/A |
| ChenQ-2021[43] | XGBoost | Internal validation using holdout testset (30%) | 1, 2, and 3 months | N/A | 1 month exclusion, 0.84, 2 month exclusion, 0.80, and 3 month exclusion 0.79 | N/A |
| ChenSM-2023[39] | Logistic Regression, Linear Discriminant Analysis (LDA), Random Forest (RF), Light Gradient Boosting machine (LightGBM), Gradient Boosting Machine (GBM), EXtreme Gradient Boosting (XGB), Support Vector Classifier (SVC), and Voting Ensemble (Voting) | Internal and external validation | N/A | 4 year | LDA = 0.91; Voting =0.99; GBM = 0.91, RF = 0.99, XGB = 0.99; LGBM = 1.0; SVC = 0.78; Logistic Regression = 0.72 | LDA = 0.91; Voting =0.90; GBM = 0.90, RF = 0.89, XGB = 0.88; LGBM = 0.86; SVC = 0.77; Logistic Regression = 0.67 |
| ChenW-2020[23] | Multi-state model | Internal validation using bootstrapping | N/A | 1, 2, and 3 years | 1 year prediction window= 0.833, 2 year prediction window= 0.830, and 3 year prediction window= 0.825 | N/A |
| ChenW-2023 (AJG)[25] | Random survival forests | Internal and external validation | N/A | 3 years | main cohort= 0.77, early detection cohort= 0.77 | main cohort= 0.71, early detection cohort= 0.68 |

| Study ID | Modeling Method | Model evaluation | Data exclusion time | Prediction time window | Model performance (C-index, internal validation) | Model performance (C-index, external validation) |
|---|---|---|---|---|---|---|
| ChenW-2023 (JCG)[24] | Random survival forests | Internal validation using bootstrapping | N/A | 3 years | 0.81–0.82 | N/A |
| DayemUllahAZM-2021[26] | Logistic regression | No information provided | N/A | N/A | N/A | N/A |
| JeonCY-2020[27] | Cox regression | No information provided | N/A | >=1 year follow up | N/A | N/A |
| JiaK-2023[46] | Neural networks and logistic regression | Internal and external validation | 6–18 months | N/A | Neural network=0.826, Logistic regression=0.80 | Neural network (average AUCs for different locations)=0.74 |
| KleinAP-2013[28] | Logistic regression | No information provided | N/A | Unclear | N/A | N/A |
| LiX-2020[29] | XGBoost with Artificial Neural Networks (Unclear DNN) | Internal validation using stratified 25% hold out testset | 3 months | 2 years | 0.81 | N/A |
| MalhotraA-2021[30] | Logistic regression | Internal validation using 25% hold out testset | 1–20 months | 0–24 months | Age < 60, 20 months exclusion, AUC = 0.66; age > 60, 17 months exclusion, AUC = 0.61 | N/A |
| MuhammadW-2019[31] | Artificial Neural Network (ANN) | Internal validation using 30% holdout testset and 10 fold cross validation | N/A | Unclear | NHIS dataset=0.71, PLCO dataset=0.62, NHIS+PLCO dataset=0.85 | N/A |
| MunigalaS-2015[32] | Logistic regression | No information provided | N/A | 3 years following new-onset diabetes mellitus | N/A | N/A |
| ParkJ-2022[33] | Grouped Neural Networks with random masking | Internal validation using 20% hold out testset | 0 and 12 months | N/A | 0 month lead time= 0.82, 12 month lead time= 0.67 | N/A |
| ParkJ-2023[47] | Neural networks | Internal validation using 20% hold out testset, 10 repeatations with random splits | Experimented, but results not reported in a text or tables | N/A | Neural network=0.85 | N/A |
| PlacidoD-2021[42] | Bag of words, Multi Layer Perceptron, Gated Recurrent Units, and Transformers | Internal and external validation | 0, 3, 6, 12 months | 3, 6, 12, 36,60 months after risk assessment | 0–36 months prediction window- Danish Dataset: Transformer model with 0 month exclusion=0.88, 3 month exclusion= 0.84, 6 month exclusion=0.83, 12 month exclusion= 0.83; Boston Dataset: Transformer model with 0 month exclusion=0.87, 3 month exclusion, 0.80, 6 month exclusion=0.79, 12 month exclusion= 0.79 | Validation of model built with Danish data tested on VA data- 0–36 month prediction, Transformer model, 0 month exclusion= 0.78, 3 month exclusion 0.70, 6 month exclusion=0.72 |
| RamsyL-2021[41] | Transformers | Internal validation using 20% hold out test set | N/A | Unclear | cerner= 0.82, truven= 0.81 | N/A |

| Study ID | Modeling Method | Model evaluation | Data exclusion time | Prediction time window | Model performance (C-index, internal validation) | Model performance (C-index, external validation) |
|---|---|---|---|---|---|---|
| RasmyL-2020[53] | Logistic regression, bidirectional Recurrent Neural Networks | Internal validation using 20% hold out testset | N/A | Next appointment | Logistic regression= 0.81 and RNN= 0.83 | N/A |
| RischHA-2015[34] | Unconditional logistic regression | No information provided | N/A | 5 years | N/A | N/A |
| SalvatoreM-2021[44] | Logistic regression | External validation | 0, 1, 2, and 5 years | N/A | N/A (no independent test set) | PheRS only −0 year lead time=0.70, 1 year lead time=0.66, 2 year lead time= 0.61, 5 year lead time=0.60; PheRS, PRS, covariates + risk factors - 0 year lead time=0.812, 1 year lead time=0.78, 2 year lead time= 0.75, 5 year lead time=0.74 |
| SharmaA-2018[35] | Logistic regression, Weighted scoring model | External validation with independent cohort | 6 months, 6–12 months, 12–18 months, and longer than 18 months (sensitivity analysis) | 3 years | N/A | N/A |
| StapleyS-2012[36] | Logistic regression | No information provided | N/A | N/A | N/A | N/A |
| YangZ-2023[63] | Logistic regression, LSTM, BERT, Transformer neural networks | Internal validation using 20% hold out testset | N/A | 12 and 36 months | Logistic regression=0.73, LSTM=0.76, Transformer neural network=0.82, BERT=0.79 | N/A |
| YuA-2016[37] | Cox proportional hazards model | External validation | N/A | 8 year | N/A | men, 0.81; women, 0.80 |
| ZhaoX-2020[38] | Logistic regression | External validation | N/A | median followup=7 years (range=3–12 years) | N/A | 0.72 |
| ZhuW-2023[45] | Logistic regression | Internal validation with a test set | 2.5 years | N/A | Logistic regression=0.74 | N/A |

**Table 3.**

Best practices and recommendations for future ML/AI modeling studies in early detection of PC using EHR data

| AI/ML Modeling Topic | Best Practice Recommendations |
|---|---|
| PC outcome definition | Identify a homogeneous outcome definition and avoid combining different PC types such as PDAC and PNET |
| Modeling strategy | Consider and evaluate modeling assumptions to reduce biased estimates, inefficient models, and incorrect conclusions |
| Candidate predictors | Utilize a wide variety of candidate predictors, including structured and unstructured EHR data without limiting selection to known risk predictors of PC |
| Model Validation | At a minimum, perform internal validation via a resampling technique. Strive to perform external validation. |
| Predicting PC “early” | Consider a data exclusion time window of at least 12 months to avoid identifying features associated with clinically overt and advanced disease |
| Explaining AI models | Utilize explainable-AI techniques to provide additional context to the model predictions |
| Missing data handling | Employ methods that minimize bias when deciding how to handle missing data and explain how missing data is handled |