

1 Developing and validating a pancreatic cancer 2 risk model for the general population using 3 multi-institutional electronic health records from 4 a federated network

5 Kai Jia¹, Steven Kundrot², Matvey Palchuk², Jeff Warnick², Kathryn
6 Haapala², Irving Kaplan³, Martin Rinard¹ *, and Limor Appelbaum³ *

7 ¹ Department of Electrical Engineering and Computer Science,
8 Massachusetts Institute of Technology, Cambridge MA 02139 USA

9 jiakai@mit.edu rinard@csail.mit.edu

10 ² TriNetX, LLC, Cambridge MA 02140 USA

11 steve.kundrot@trinetx.com matvey.palchuk@trinetx.com

12 jeff.warnick@trinetx.com kathryn.haapala@trinetx.com

13 ³ Beth Israel Deaconess Medical Center, Boston MA 02215 USA

14 ikaplan@bidmc.harvard.edu lappelb1@bidmc.harvard.edu

15 Abstract

16 **Purpose:** Pancreatic Duct Adenocarcinoma (PDAC) screening can enable de-
17 tection of early-stage disease and long-term survival. Current guidelines are
18 based on inherited predisposition; only about 10% of PDAC cases meet screening
19 eligibility criteria. Electronic Health Record (EHR) risk models for the general
20 population hold out the promise of identifying a high-risk cohort to expand the
21 currently screened population. Using EHR data from a multi-institutional fed-
22 erated network, we developed and validated a PDAC risk prediction model for
23 the general US population.

24 **Methods:** We developed Neural Network (NN) and Logistic Regression (LR)
25 models on structured, routinely collected EHR data from 55 US Health Care Or-
26 ganizations (HCOs). Our models used sex, age, frequency of clinical encounters,
27 diagnoses, lab tests, and medications, to predict PDAC risk 6-18 months before
28 diagnosis. Model performance was assessed using Receiver Operating Character-
29 istic (ROC) curves and calibration plots. Models were externally validated using
30 location, race, and temporal validation, with performance assessed using Area
31 Under the Curve (AUC). We further simulated model deployment, evaluating
32 sensitivity, specificity, Positive Predictive Value (PPV) and Standardized Inci-
33 dence Ratio (SIR). We calculated SIR based on the SEER data of the general
34 population with matched demographics.

35 **Results:** The final dataset included 63,884 PDAC cases and 3,604,863 controls
36 between the ages 40 and 97.4 years. Our best performing NN model obtained an
37 AUC of 0.829 (95% CI: 0.821 to 0.837) on the test set. Calibration plots showed

* Co-senior authors.

38 good agreement between predicted and observed risks. Race-based external val-
39 idation (trained on four races, tested on the fifth) AUCs of NN were 0.836 (95%
40 CI: 0.797 to 0.874), 0.838 (95% CI: 0.821 to 0.855), 0.824 (95% CI: 0.819 to
41 0.830), 0.842 (95% CI: 0.750 to 0.934), and 0.774 (95% CI: 0.771 to 0.777) for
42 AIAN, Asian, Black, NHPI, and White, respectively. Location-based external
43 validation (trained on three locations, tested on the fourth) AUCs of NN were
44 0.751 (95% CI: 0.746 to 0.757), 0.749 (95% CI: 0.745 to 0.753), 0.752 (95% CI:
45 0.748 to 0.756), and 0.722 (95% CI: 0.713 to 0.732) for Midwest, Northeast,
46 South, and West, respectively. Average temporal external validation (trained on
47 data prior to certain dates, tested on data after a date) AUC of NN was 0.784
48 (95% CI: 0.763 to 0.805). Simulated deployment on the test set, with a mean
49 follow up of 2.00 (SD 0.39) years, demonstrated an SIR range between 2.42-83.5
50 for NN, depending on the chosen risk threshold. At an SIR of 5.44, which ex-
51 ceeds the current threshold for inclusion into PDAC screening programs, NN
52 sensitivity was 35.5% (specificity 95.6%), which is 3.5 times the sensitivity of
53 those currently being screened with an inherited predisposition to PDAC. At
54 a chosen high-risk threshold with a lower SIR, specificity was about 85%, and
55 both models exhibited sensitivities above 50%.

56 **Conclusions:** Our models demonstrate good accuracy and generalizability across
57 populations from diverse geographic locations, races, and over time. At compa-
58 rable risk levels these models can predict up to three times as many PDAC cases
59 as current screening guidelines. These models can therefore be used to identify
60 high-risk individuals, overlooked by current guidelines, who may benefit from
61 PDAC screening or inclusion in an enriched group for further testing such as
62 biomarker testing. Our integration with the federated network provided access
63 to data from a large, geographically and racially diverse patient population as
64 well as a pathway to future clinical deployment.

65 1 Introduction

66 Most cases of Pancreatic Duct Adenocarcinoma (PDAC) are diagnosed as advanced-
67 stage disease, leading to a five-year relative survival rate of only 11% [26]. Ex-
68 panding the population currently being screened for this lethal disease is crucial
69 for increasing early detection and improving survival. Current screening guide-
70 lines [4, 10, 12] targeting stage I cancers and high-grade PDAC precursors have
71 been shown to significantly improve long-term survival [6, 18]. Current guide-
72 lines target patients with a family history or genetic predisposition to PDAC
73 [13, 21], with screening eligibility based on estimated absolute and relative risk
74 compared to the general population (5% or 5 times the relative risk, respectively)
75 [6]. These patients comprise only about 10% of all PDAC cases. No consensus
76 or guidelines exist for PDAC screening in the *general population* [20], where the
77 *majority* of PDAC cases are found.

78 Several groups have developed PDAC risk models for the general population
79 using various data sources [5, 15, 16]. A goal of most such models is eventual
80 integration with Electronic Health Record (EHR) systems and ultimately clinical

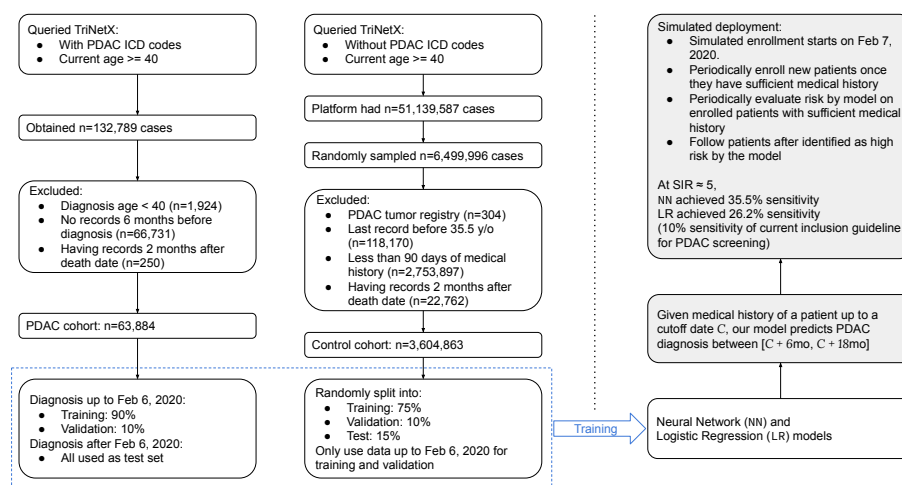


Fig. 1: Flowchart of our study with simulated deployment as an example

81 implementation. EHR integration has proven to be a significant barrier to the
 82 clinical adoption of models [28]. One effort developed a model using EHR data
 83 from an aggregated multi-institutional database [7]. The evaluation focused on
 84 identification of high risk patients up to one month before diagnosis and did not
 85 attempt to evaluate model generalization across locations or races. Several other
 86 efforts worked with real-world EHR data [3, 8, 22], but with limited validation
 87 across diverse locations and races. Other efforts worked with small sample sizes
 88 [5, 19] and internal validation only [16, 19].

89 We used EHR data from 55 US Health Care Organizations (HCOs) from
 90 a federated data network to develop and validate two PDAC risk prediction
 91 models for the general population, a Neural Network (NN) model and a Logistic
 92 regression (LR) model. The models can be used as a tool to identify individuals
 93 at high risk for PDAC from the general population, so they can be offered early
 94 screening or referred for lower overhead testing such as biomarker testing.

95 The network provides access to harmonized, de-identified EHR data of over
 96 89 million patients for model development and testing. It also provides a means
 97 to simulate deployment of the resultant models to identify high risk patients
 98 for screening within a research setting. Because the network is connected to the
 99 EHR systems of the participating HCOs, it provides a pathway to deploy the
 100 models to a clinical setting, a critical step in the progression towards successful
 101 clinical adoption [28].

102 We developed a methodology to train PDAC prediction models on federated
 103 network EHR data. Our evaluation reports AUC and PPV numbers for the re-
 104 sulting trained models, with the evaluation focusing on the ability of the models
 105 to identify high risk patients 6 to 18 months before an initial PDAC diagno-
 106 sis. We conducted three types of external validation: location-based, race-based,

107 and temporal. We simulated deployment of the model on real-world HCO data
108 to evaluate its performance in a more realistic setting. We compared the rela-
109 tive incidence of PDAC in our model-assigned high-risk group with that of a
110 demographically matched general US population based on SEER data [1].

111 2 Methods

112 2.1 Data source and setting

113 This is an observational retrospective study, with both a case-control and cohort
114 design, using data from the federated EHR database platform of TriNetX [27].
115 TriNetX is a federated global health research network that specializes in data
116 collection and distribution. HCOs contributing to the database include academic
117 medical centers, community hospitals, and outpatient clinics.

118 We used retrospective de-identified EHR data from 55 HCOs across the
119 United States. The majority of these HCOs are tertiary care centers and the data
120 used includes inpatient, outpatient, and Emergency Room encounters. Different
121 HCOs have different historical coverage; on average, each HCO provides approx-
122 imately 13 years of historical data. Data include values from structured EHR
123 fields (e.g. demographics, date-indexed encounters, diagnoses, procedures, labs,
124 and medications) as well as facts and narratives from free text (e.g. medications
125 identified through Natural Language Processing (NLP)). TriNetX harmonizes all
126 data from each HCO's EHR to the TriNetX common data model and common
127 set of controlled terminologies. TriNetX also has tools to identify anomalies and
128 outliers for quality assurance.

129 We used data from the TriNetX database under a no-cost collaboration agree-
130 ment between BIDMC, MIT, and TriNetX. Under this agreement, we accessed
131 de-identified data under the agreements and institutional approvals already in
132 place between TriNetX and their partner institutions.

133 2.2 Study population

134 We worked with two cohorts: a PDAC cohort and a control cohort. We obtained
135 all data from TriNetX during November and December, 2022. We obtained the
136 PDAC cohort by querying the TriNetX database to obtain EHR data for all
137 patients, 40 years of age or older, from 55 HCOs across the United States, with
138 one of the following ICD-10/ICD-9 codes:

- 139 – C25.0 Malignant neoplasm of head of pancreas
- 140 – C25.1 Malignant neoplasm of body of pancreas
- 141 – C25.2 Malignant neoplasm of tail of pancreas
- 142 – C25.3 Malignant neoplasm of pancreatic duct
- 143 – C25.7 Malignant neoplasm of other parts of pancreas
- 144 – C25.8 Malignant neoplasm of overlapping sites of pancreas
- 145 – C25.9 Malignant neoplasm of pancreas, unspecified

146 – 157 Malignant neoplasm of pancreas (ICD-9 without a corresponding ICD-10
147 code)

148 We obtained $n=132,789$ PDAC cases. We excluded patients who were diagnosed
149 before 40 years of age ($n=1,924$), patients with no medical history 6 months
150 prior to diagnosis ($n=66,731$), and patients with records 2 months after their
151 death record ($n=250$), to obtain a PDAC cohort with $n=63,884$ cases.

152 To prepare the control cohort, we queried the TriNetX database for patients
153 at least 40 years of age without any of the above ICD-10 or ICD-9 codes. There
154 were $n=51,139,587$ patients that met this criteria. From these patients we ran-
155 domly selected $n=6,499,996$ patients. We excluded patients with a PDAC tumor
156 registry entry but no PDAC diagnosis entries ($n=304$), patients whose last entry
157 was before age 35.5 ($n=118,170$), patients with less than 90 days of medical his-
158 tory ($n=2,753,897$), and patients with records 2 months after their death record
159 ($n=22,762$), to obtain a control cohort with $n=3,604,863$ cases. Our subsequent
160 training and testing procedures implement additional exclusion criteria (see be-
161 low).

162 2.3 Model development

163 We used the TRIPOD guidelines for multivariable prediction models for report-
164 ing on model development and validation [9].

165 We trained and evaluated two model classes, Neural Network (NN) and Lo-
166 gistic Regression (LR). Data was randomly partitioned into training, validation,
167 and test sets (75%, 10%, and 15%, respectively). We evaluated model perfor-
168 mance by AUC scores and sensitivity, specificity, PPV, and SIR in simulated
169 deployment. To calculate SIR, we used the SEER database [1] to estimate the
170 PDAC risk for our model’s high-risk group compared to the general population.

171 Our training and testing procedures work with a cutoff date C for every
172 patient, with entries after the cutoff date excluded. For a patient P and a cutoff
173 date C , the model uses entries available before the cutoff date C to predict the
174 risk of first diagnosis of PDAC between $C + 6\text{mo}$ to $C + 18\text{mo}$. We defined the
175 date of PDAC diagnosis D to be the first time a PDAC ICD code (as above)
176 appeared in the patient record. During training, we sampled the cutoff dates for
177 PDAC cases uniformly between $[D - 18\text{mo}, D - 6\text{mo}]$. Since control patients were
178 not diagnosed with PDAC, we sampled random cutoff dates for them from the
179 distribution of the PDAC diagnosis dates. For a control patient with a known
180 death date, we limited the cutoff date to at most 18 months before death, to rule
181 out undiagnosed PDAC that caused death. To avoid undiagnosed PDAC cases,
182 we also limited all cutoff dates of patients in the control cohort to be at most 18
183 months before the dataset query date.

184 We empirically defined any patient with at least 16 diagnosis, medication, or
185 lab result entries within 2 years before their cutoff date and whose first entry
186 is at least 3 months earlier than their last entry before the cutoff date to have
187 *sufficient medical history*. We excluded patients that did not have sufficient med-
188 ical history. We trained the NN with the iterative Stochastic Gradient Descent

189 (SGD) algorithm [17], sampling a new cutoff date for each patient at each step
190 of the iteration. Our LR training sampled one cutoff date for each patient.

191 Our feature extraction excluded entries after the cutoff date (and included
192 all entries up to the cutoff date). For each patient, we defined six basic features
193 including age, whether age is known, sex, whether sex is known, number of
194 diagnosis, medication, or lab entries in the medical record up to 18 months before
195 cutoff (the recent entries), and number of diagnosis, medication, or lab entries in
196 the medical record greater than 18 months before cutoff (the early entries). We
197 also included features that correspond to individual diagnosis, medication, or
198 lab codes, with the corresponding code empirically included in feature selection
199 if it appeared in the medical record of at least 1% of the patients in the cancer
200 cohort of the training set.

201 We manually grouped 827 commonly used diagnosis codes into 39 groups.
202 For ungrouped codes, we used the ICD-10 category plus the first digit of the
203 subcategory. We derived 3 features for each diagnosis code: whether or not it
204 exists $\{0, 1\}$, its first and last date (encoding for first and last date: greater
205 or equal to 4 years before cutoff=0; at cutoff=1). To use past ICD-9 data to
206 train the model for use on current and future ICD-10 data, we mapped all ICD-
207 9 codes to their ICD-10 equivalents. For ICD-9 codes that could be mapped
208 to more than one ICD-10 code, we included the features of all the mapped
209 ICD-10 codes in the feature vector. We also manually grouped 67 medication
210 codes into 8 different medication classes. Ungrouped codes were used as they
211 are. We derived 4 features for each medication code: whether or not it exists
212 $\{0, 1\}$, its frequency (i.e., number of times it appears in the medical record),
213 span (time between first and last appearance of a medication code), and last
214 date (same encoding as diagnosis first/last date). For lab features, we used a
215 grouping provided by TriNetX for similar lab tests, which had 98 groups for 462
216 codes. Ungrouped codes were used as they are. For each lab code, we derived 4
217 features: existence, frequency, first date, and last date. The frequency was the
218 number of lab results within three years before cutoff. We manually selected the
219 most relevant lab tests for PDAC prediction, based on clinical knowledge and
220 literature review. For these manually selected 44 quantitative labs, we derived
221 two additional features: lab test value and slope. Lab values were normalized
222 according to the median absolute deviation and the population median (range
223 -1 to 1). Slope was measured by calculating the yearly change in lab test values
224 up to three years before cutoff.

225 To account for the additional effect of the healthcare process on EHR data
226 [2], we did the following: For each feature type described above (except the
227 number of early and recent entries in basic features) there is a corresponding
228 existence feature $\{0, 1\}$; if the feature is missing in the data set, the value of the
229 corresponding existence feature is 1 and the value of the feature itself is 0. This
230 encoding enables the model to compute risk scores based on whether a feature
231 is present or missing. Because our NN models can use sophisticated nonlinear
232 reasoning to extract information from the chosen features, data imputation pro-

233vides little to no useful additional information for these models. Therefore, we
234did not use any imputation.

235Our NN models have three fully connected layers; each layer has 48, 16,
236and 1 output neurons. Hidden layers use the tanh nonlinearity. To ameliorate
237overfitting, we used sparse weights computed by the recently developed BinMask
238sparsification technique [14]. We used balanced numbers of PDAC and control
239patients in each mini-batch. For LR training, we used the SAGA solver [11]
240with balanced class weights. For each model type, we trained four models with
241different regularization parameters and selected the best one on the validation
242set.

243We calibrated the models on the validation set with a modified Platt cali-
244bration algorithm [23], where we fitted a two-segment piecewise-linear mapping
245with the turning point set as the median of model predictions. We accounted for
246the unbalanced sampling of control cohort and estimated the risk on the whole
247population in calibration. We evaluated our calibration by creating calibration
248plots on the test set. We chose 16 risk groups for calibration evaluation as a
249geometric sequence between the 85% percentile of predicted risk on the test set
250and the maximum predicted risk. To quantitatively compare calibration between
251models, we used the Geometric Mean of Over Estimation (GMOE), calculated
252as the geometric mean of the ratios of predicted risk to the true risk over all
253tested risk groups. Perfectly calibrated models have $GMOE=1$. A GMOE value
254greater than one means over estimation of risk and a value less than one means
255under estimation of risk.

256We also evaluated the stability of our algorithm by calculating the mean AUC
257and GMOE with confidence interval on nine independent runs with different
258random seeds for dataset split and weight initialization.

259For both the LR and NN models, we analyzed the impact of different num-
260bers of features on model performance. We reduced the number of input features
261by applying BinMask to the input of a small and densely connected neural net-
262work to automatically select important features. We varied the BinMask weight
263decay coefficient to obtain different numbers of input features and evaluated the
264performance of our models with those feature sets.

265We analyzed the feature importance for NN by calculating the partial AUC
266(up to 6% FPR) obtained with only each type of medical record entries. A larger
267score for a type of record means the NN makes better predictions based on the
268record entries alone.

269 2.4 External validation

270Our model validation considered three attributes: geographical location of the
271HCO, patient race, and time of diagnosis/last used entry in the medical record.
272For each attribute, we split the dataset according to that attribute, trained
273models on one split, and tested on the other split.

274Our location based validation used the TriNetX geographical location for each
275HCO; locations include Northeast, South, Midwest, and West. Our race based
276validation used the TriNetX racial classification of each patient; races include

277 American Indian or Alaska Native (AIAN), Asian, Black or African American
278 (Black), Native Hawaiian or Other Pacific Islander (NHPI), and White.

279 A primary assessment of model generalizability is the AUC gap between
280 test set and validation set. However, since different attribute splitting produces
281 training/validation/test sets with different sizes, the test/validation AUC gap
282 does not necessarily depict model generalizability. Therefore, we trained control
283 models that used the same training and test set size for each attribute-based
284 split, but used random splitting that ignores attribute values. We also assessed
285 model generalizability by checking the AUC gap between the external validation
286 models and corresponding control models.

287 For temporal validation, we selected the 50%, 60%, . . . , 90% percentile from
288 the distribution of diagnosis dates as the dataset split dates. The 90% percentile
289 was Sep 23, 2021. We trained the models only on data available prior to those
290 split dates. We also limited the cutoff date of control patients to earlier than 18
291 months before the split dates, to simulate model training with datasets queried
292 on the split dates. We evaluated the performance of the models on the same
293 subset of data only available after Sep 23, 2021. We also calculated the aver-
294 age performance of different models for the temporal validation. Since different
295 dataset split dates result in different training set sizes, we also trained control
296 models. For each split date, we randomly sampled the same number of PDAC
297 cases (equal to the 50% of the total number of PDAC cases) from cases up to
298 that split date. The control models allowed us to separate the contribution of
299 larger training set from the impact of smaller time gap between training and
300 test sets.

301 2.5 Simulated deployment

302 We estimate the performance of our model when deployed in a clinical setting by
303 simulating model deployment in a prospective study on the TriNetX database.
304 We trained the model only on data available prior to Feb 7, 2020, in the same
305 way as the above temporal validation, with the dataset split date chosen as
306 the 70% percentile of the distribution of the diagnosis dates. For each date D
307 separated by 90 days between Feb 7, 2020 and May 2, 2021 (18 months before
308 dataset query), we

- 309 1. Enrolled a new patient into the simulated deployment if the patient had a
310 known age, was at least 40 years old on date D , and had sufficient medical
311 history on D for the first time. We call the date D the *enrollment date* for
312 such a patient.
- 313 2. For each enrolled patient, we checked if that patient still had sufficient med-
314 ical history on D . If so, we evaluated the model risk by our model, with the
315 cutoff date set at D . We call the date D a *check date* for such a patient.

316 We excluded patients who were diagnosed with PDAC either before enroll-
317 ment or within 6 months after enrollment, patients who had no medical entries
318 between first and last check dates, and patients with a known death but no

319 PDAC diagnosis within 18 months after enrollment. We started following up a
320 patient 6 months after their enrollment date. We stopped following up a patient
321 18 months after the last check date. During the followup period, we defined the
322 following outcomes:

- 323 1. A patient was diagnosed with PDAC. We counted this patient as a true
324 positive if the model made a high-risk prediction on any check date 6 months
325 prior to diagnosis and a false negative otherwise.
- 326 2. A patient was not diagnosed with PDAC. They might either have a known
327 death date, reached our dataset query date, or never had sufficient medical
328 history again after a certain check date. For patients with a known death
329 date, we only considered check dates up to 18 months before death, due
330 to the possibility of undiagnosed PDAC at death. For other patients, we
331 considered all check dates. If the model ever made a high-risk prediction for
332 this patient on any considered check dates, we counted the patient as a false
333 positive. Otherwise, we counted the patient as a true negative.

334 We chose the risk thresholds according to the 89.00%, 93.00%, 96.50%, 98.00%,
335 99.70%, 99.92% specificity levels on the validation set. For each risk threshold,
336 we computed sensitivity, specificity, Positive Predictive Value (PPV), and Stan-
337 dardized Incidence Ratio (SIR), based on the above protocol. Since we used all
338 the PDAC cases in the TriNetX database, but sampled a subset of control pa-
339 tients, we accounted for this imbalance to estimate the PPV and SIR that would
340 be obtained if we had evaluated the model on the full TriNetX population.

341 We calculated SIR by dividing the observed PDAC cases in the high-risk
342 group by the expected number of PDAC cases of that group. To calculate the
343 expected number of cases, we used the SEER database [1], matched with age,
344 sex, race, and calendar year for each individual in the high-risk group, as done
345 by Porter et al. [24].

346 3 Results

347 3.1 Model evaluation

348 The final LR model and NN models used 63,884 cancer patients and 3,604,863
349 controls up to 97.4 years old (determined at the time of diagnosis or last record).
350 Detailed demographics, including sex, age, race, and HCO location, are given in
351 [Table 1](#). [Fig. 1](#) presents a flowchart demonstrating how this dataset was derived.

352 The NN outperformed the LR model on the test set, with an AUC of 0.827
353 (95% CI: 0.822 to 0.833) and 0.809 (95% CI: 0.804 to 0.815), respectively ([Fig. 2a](#)).
354 The mean AUCs of NN and LR on nine random runs are 0.829 (95% CI: 0.821
355 to 0.837) and 0.810 (95% CI: 0.803 to 0.817), respectively. Because our models
356 predict based in part on the presence or absence of features, each feature is a
357 predictor and we have no participants with missing predictors [2].

358 [Fig. 2b](#) shows the log-scale calibration plots on the test set. The evaluated
359 risk levels are selected according to a geometric sequence between the 85% risk

Table 1: Demographics of our dataset.

Attribute		Cancer group (n=63,884) % (No.)	Control group (n=3,604,863) % (No.)
Sex	Female	50.40 (32,196)	55.27 (1,992,432)
	Male	49.59 (31,681)	43.42 (1,565,131)
	Unknown	0.01 (7)	1.31 (47,300)
Age at first record	Mean (SD)	60.88 (12.02)	53.90 (14.03)
	< 40	4.88 (3,116)	17.37 (626,073)
	40 - 50	12.38 (7,908)	21.52 (775,841)
	50 - 60	24.35 (15,556)	23.30 (840,092)
	60 - 70	30.69 (19,605)	19.76 (712,411)
	70 - 80	18.93 (12,091)	11.01 (396,806)
	> 80	3.54 (2,259)	2.50 (90,128)
Age at diagnosis / last record	Mean (SD)	67.67 (10.59)	60.20 (13.10)
	< 40	0.00 (0)	4.78 (172,349)
	40 - 50	6.01 (3,841)	20.00 (720,800)
	50 - 60	16.42 (10,490)	22.90 (825,442)
	60 - 70	30.56 (19,522)	23.44 (844,818)
	70 - 80	29.68 (18,958)	17.05 (614,615)
	> 80	12.09 (7,724)	7.30 (263,327)
Age	Unknown	5.24 (3,349)	4.54 (163,512)
Race	AIAN	0.26 (164)	0.36 (13,023)
	Asian	1.53 (976)	2.27 (81,726)
	Black	13.95 (8,910)	14.16 (510,444)
	NHPI	0.05 (35)	0.13 (4,694)
	White	72.70 (46,441)	67.24 (2,423,771)
	Unknown	11.52 (7,358)	15.85 (571,205)
HCO location	Midwest	21.17 (13,527)	15.41 (555,417)
	Northeast	33.42 (21,352)	28.40 (1,023,916)
	South	36.37 (23,234)	44.18 (1,592,634)
	West	7.41 (4,733)	8.54 (308,013)
	Unknown	1.62 (1,038)	3.46 (124,883)
No. medical records	Mean (SD)	779.11 (1506.23)	441.79 (1091.31)

Race abbreviations:

- AIAN: American Indian or Alaska Native
- Black: Black or African American
- NHPI: Native Hawaiian or Other Pacific Islander

360 percentile and the maximal risk given by the model on the test set. Geometric
 361 Mean of Over Estimation (GMOE), the geometric mean of ratios of predicted
 362 risks to observed risks, was calculated for both models. The GMOE for the NN
 363 was 1.037 and 0.861 for the LR. The GMOE on nine random runs was 1.148
 364 (95% CI: 1.092 to 1.203) and 0.992 (95% CI: 0.944 to 1.041) for NN and LR,
 365 respectively.

366 The impact of different feature numbers on model performance, for both
 367 the NN and LR models, is shown in Fig. 3a. Both models showed improved
 368 performance with an increasing number of features, reaching a plateau at an
 369 AUC of 0.83 (NN) and 0.81 (LR) for a combination of 1574 diagnoses features,
 370 862 medication features, and 719 lab features. Additional features produced no
 371 significant improvement in model performance.

372 Fig. 3b shows the top features selected by the LR model and ranked by feature
 373 importance. The top features include codes related to glucose metabolism and

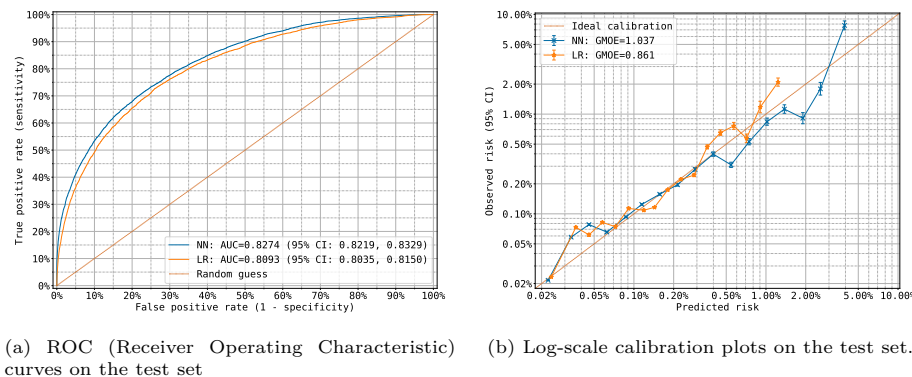


Fig. 2: Model performance on the test set.

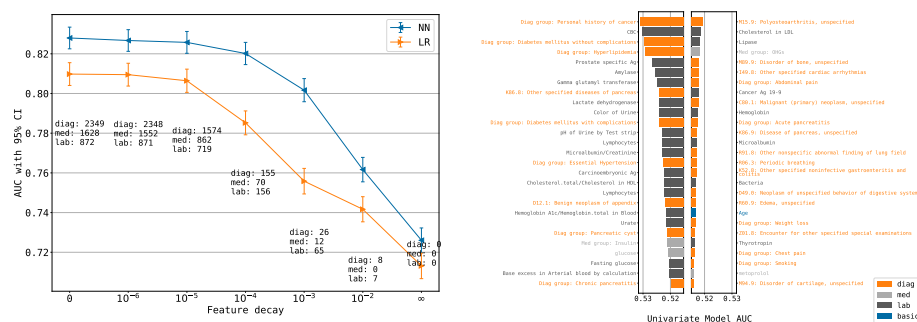
374 diabetes, medications such as Insulin and oral hypoglycemics, as well as blood
375 tests for glucose and fasting glucose and HbA1c. Top features also include known
376 PDAC risk factors such as age, pancreatitis, pancreatic cysts, personal history
377 of cancer, weight loss, and smoking.

378 3.2 External validation results

379 Fig. 4 shows the results for race-based, location-based, and temporal external
380 validations. The model performed similarly across racial groups without signif-
381 icant performance drop, as shown in Fig. 4a. AUCs on the test set were 0.836
382 (95% CI: 0.797 to 0.874), 0.838 (95% CI: 0.821 to 0.855), 0.824 (95% CI: 0.819
383 to 0.830), 0.842 (95% CI: 0.750 to 0.934), and 0.774 (95% CI: 0.771 to 0.777) for
384 AIAN, Asian, Black, NHPI, and White racial groups, respectively. The AUCs
385 of the LR models were 0.801 (95% CI: 0.755 to 0.846), 0.822 (95% CI: 0.804
386 to 0.840), 0.806 (95% CI: 0.800 to 0.811), 0.836 (95% CI: 0.742 to 0.929), and
387 0.773 (95% CI: 0.770 to 0.775). Test AUCs of NN models were -0.035 to 0.015
388 lower than the corresponding control models, and -0.024 to 0.008 lower for LR
389 models. The number of patients of each racial groups can be seen in Table 1. We
390 excluded patients with unknown race from this experiment.

391 Model performance was similar across the different geographic locations as
392 shown in Fig. 4b. NN AUCs on the test set were 0.751 (95% CI: 0.746 to 0.757),
393 0.749 (95% CI: 0.745 to 0.753), 0.752 (95% CI: 0.748 to 0.756), and 0.722 (95%
394 CI: 0.713 to 0.732) for the Midwest, Northeast, South, and West, respectively.
395 LR AUCs were 0.742 (95% CI: 0.737 to 0.748), 0.735 (95% CI: 0.730 to 0.739),
396 0.726 (95% CI: 0.722 to 0.730), and 0.623 (95% CI: 0.610 to 0.636). Test AUCs
397 of NN models were 0.074 to 0.112 lower than the corresponding control models,
398 and 0.060 to 0.191 lower for LR models. The number of patients in each location
399 can be seen in Table 1. We excluded patients with unknown HCO location from
400 this experiment.

12



(a) Model performance with different numbers of features induced by different feature decay penalties. All models also use the six-dimensional basic feature besides the indicated number of features in diag, med, and lab categories.

(b) Most important features

Fig. 3: Feature analysis. In the plots, diag refers to diagnosis features, med refers to medication features, and lab refers to lab features.

401 For temporal validation, model test performance varied over time, although
 402 they had relatively stable validation AUCs. Both NN and LR showed improved
 403 performance by adding more recent training data. The control models had worse
 404 performance and showed less stable improvement over time, which suggests that
 405 training set size is an important factor. The average test AUCs were 0.784 (95%
 406 CI: 0.763 to 0.805) and 0.768 (95% CI: 0.747 to 0.788) for the NN and LR models,
 407 respectively.

408 3.3 Simulated deployment results

Table 2: Simulated deployment results. Numbers in brackets are 95% CI.

Model	Risk level	Sensitivity	Specificity	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)
NN	1	54.5% (53.4 to 55.5)	85.6% (85.5 to 85.8)	0.30% (0.30 to 0.31)	2.42 (2.39 to 2.46)
	2	46.0% (44.9 to 47.1)	90.8% (90.7 to 90.9)	0.40% (0.39 to 0.41)	3.25 (3.20 to 3.29)
	3	35.5% (34.4 to 36.5)	95.6% (95.5 to 95.7)	0.64% (0.62 to 0.66)	5.44 (5.36 to 5.51)
	4	29.8% (28.8 to 30.8)	97.4% (97.3 to 97.4)	0.90% (0.86 to 0.94)	8.10 (7.98 to 8.21)
	5	17.4% (16.6 to 18.3)	99.5% (99.5 to 99.5)	2.66% (2.46 to 2.87)	26.0 (25.7 to 26.4)
	6	11.3% (10.7 to 12.0)	99.9% (99.9 to 99.9)	7.81% (6.80 to 8.99)	83.5 (82.1 to 84.7)
LR	1	52.9% (51.8 to 54.0)	84.1% (84.0 to 84.3)	0.27% (0.26 to 0.27)	2.02 (1.99 to 2.05)
	2	44.2% (43.1 to 45.3)	89.5% (89.4 to 89.7)	0.34% (0.33 to 0.35)	2.54 (2.49 to 2.57)
	3	33.4% (32.4 to 34.4)	94.6% (94.5 to 94.7)	0.49% (0.47 to 0.51)	3.71 (3.65 to 3.76)
	4	26.2% (25.3 to 27.2)	96.8% (96.7 to 96.9)	0.65% (0.62 to 0.68)	5.01 (4.93 to 5.08)
	5	10.3% (9.66 to 11.0)	99.5% (99.5 to 99.5)	1.57% (1.44 to 1.72)	12.8 (12.6 to 13.0)
	6	5.39% (4.91 to 5.90)	99.8% (99.8 to 99.9)	2.66% (2.31 to 3.07)	22.6 (22.2 to 22.9)

PPV: Positive Predictive Value

SIR: Standardized Incidence Ratio

TrxPop. Est.: Estimation on the whole TriNetX population that accounts for unbalanced sampling

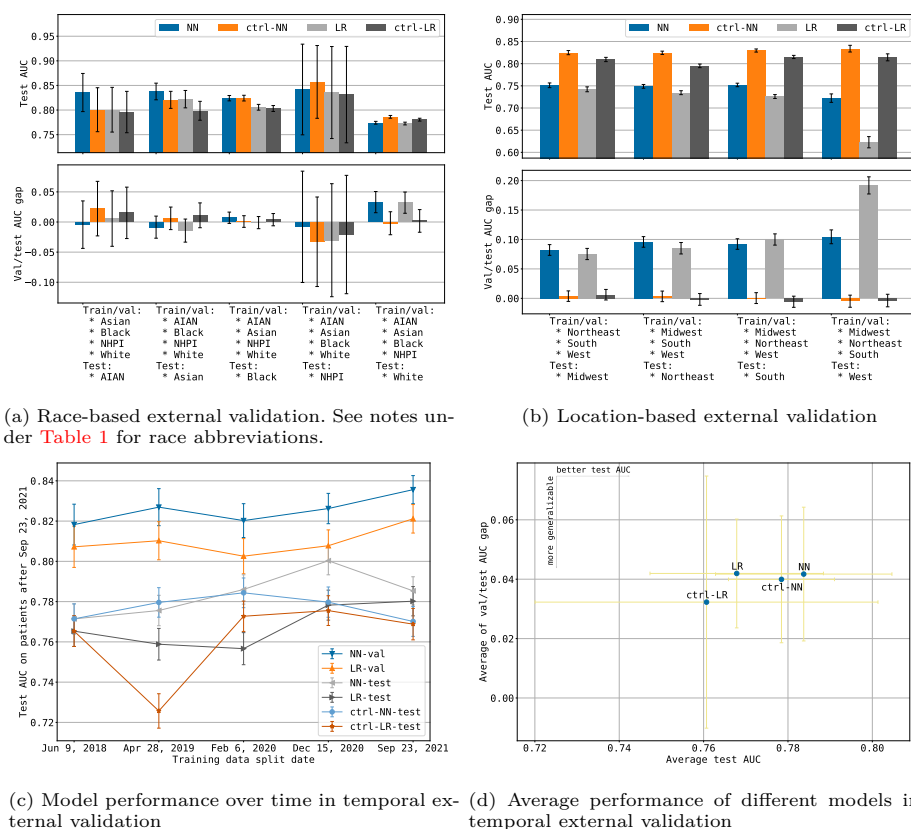


Fig. 4: Results for location-based, race-based, and temporal external validations. Error bars indicate 95% CI.

409 The simulated deployment of the NN and LR models was on 201,703 patients
 410 (with 8,113 PDAC cases) in the test set, with enrollment from Feb 7, 2020 to
 411 May 2, 2021. Mean age at enrollment was 61.45 (SD 11.97). Mean age at PDAC
 412 diagnosis was 69.65 (SD 10.40). Each patient was followed up for 2.00 (SD 0.39)
 413 years ([Table 2](#)).

414 Having accounted for unbalanced sampling of PDAC and control cohorts,
 415 we estimated that the model PPV range on the whole TriNetX population was
 416 0.30%-7.81% for the NN and 0.27%-2.66% for the LR. NN and LR SIR ranges
 417 were 2.42-83.5 and 2.02-22.6, respectively. The SIR of all the enrolled patients
 418 during the follow-up period was 0.95 (95% CI: 0.94 to 0.96). An SIR close to 1
 419 indicates that our TriNetX test population with patient exclusion has similar
 420 PDAC incidence as the general US population.

421 We determined the high-risk group to be any individuals that have an SIR of
 422 5.44 or above, based on the NN model. This threshold is correlated with a 35.5%
 423 sensitivity and 95.6% specificity. We use this SIR threshold because it is similar

14

424 to the currently used eligibility cutoff for inclusion of individuals into screening
425 programs [12].

426 **4 Discussion**

427 Our study leveraged routinely collected EHR data from a federated network
428 including 55 HCOs across the United States to develop and validate two ML
429 models (NN and LR) that can accurately identify patients in the general popu-
430 lation at high risk for PDAC, 6 to 18 months before first PDAC diagnosis. Both
431 models were trained on 63,884 PDAC cases and 3,604,863 controls; both models
432 worked with features derived from medical record entries including diagnosis,
433 medication, and lab results, as well as basic features including sex, age, and
434 number of clinical encounters. Our NN model obtained an AUC of 0.829 (95%
435 CI: 0.821 to 0.837) on the test set; the LR model obtained an AUC of 0.810
436 (95% CI: 0.803 to 0.817).

437 **4.1 Potential use cases**

438 We anticipate two potential clinical use cases for our models. The first is to
439 expand the eligibility for current screening programs, which are based on imaging
440 modalities such Endoscopic UltraSound (EUS) and MRI/MRCP [6]. Current
441 eligibility criteria are based on familial PDAC or a known germline mutation
442 syndrome (e.g., Lynch, Peutz-Jeghers) [6]. The identified population is known
443 to have an SIR of minimum 5 times the SIR of the general population and
444 includes only 10% of PDAC cases [13, 21]. Depending on the chosen high-risk
445 threshold, our NN model exhibited an SIR of 2.42 to 83.5. At an SIR of 5.44,
446 our NN model identifies 35.5% of the PDAC cases as high risk 6 to 18 months
447 before diagnosis, a significant improvement over current screening criteria.

448 The second use case is to identify an enriched group for lower overhead testing
449 (such as biomarker testing) followed by screening based on the lower overhead
450 test. In this use case we anticipate that it will be feasible to deploy the model at
451 a higher sensitivity than in our first use case. For example, at 85.6% specificity,
452 NN exhibited 54.5% sensitivity.

453 **4.2 Race-based, location-based, and temporal validation**

454 Our race based validation worked with the five racial groups recorded within
455 the TriNetX EHR data: AIAN, Asian, Black, NHPI, and White. We trained
456 models on four of these five racial groups, then tested on the fifth. The results
457 showed similar performance across all training/test pairs, highlighting the gen-
458 eralizability across diverse racial populations. There was a small AUC drop for
459 models when trained on all groups except White and tested on White, which we
460 attribute to the fact that the White group included over 70% of the PDAC cases
461 in the data set.

462 Our location based validation divided the HCOs into four regions: Midwest,
463 Northeast, South, and West. We trained models on three of the regions, then
464 tested on the fourth. In comparison with models trained on all regions with
465 randomly sampled size-matched training data, these models showed modest AUC
466 drops (0.074 to 0.112 for NN and 0.060 to 0.191 for LR).

467 Our temporal validation trained models on data before different dataset split
468 dates, then tested the models on future dates not used for training. We found
469 that NN models outperformed LR models, exhibiting average AUCs 0.784 (95%
470 CI: 0.763 to 0.805) and 0.768 (95% CI: 0.747 to 0.788), respectively.

471 **4.3 Simulated deployment**

472 We envision the eventual deployment of our models into clinical practice to im-
473 prove patient outcomes by promoting the detection of early stage disease. We
474 evaluated the effectiveness of our models for this purpose by simulating the de-
475 ployment of our models. A key aspect of this simulated deployment was using
476 models trained only on data available before a simulated enrollment date to iden-
477 tify high-risk individuals after the simulated enrollment date. We then followed
478 the identified high-risk individuals over time to evaluate the performance of our
479 models.

480 This simulated deployment methodology stands in contrast to methodologies
481 used in previous studies that do not temporally separate the training and test
482 data [5, 7]. By more closely tracking the envisioned deployment scenario, we
483 eliminated a potential source of inaccuracy and hope to obtain a more accurate
484 prediction of model performance in clinical use.

485 **4.4 Federated network**

486 A significant strength of our work is the development and validation of our
487 models using a federated EHR network. This network ingests EHR data from
488 multiple HCOs and EHR sources, with the data remaining stored behind each
489 institution's firewall. The ingested data is de-identified, harmonized, and con-
490 verted into a single format, supporting ease of integration and deployment of
491 models within the same platform. This federated network enabled us to train
492 and externally validate our models on racially, geographically, and temporally
493 diverse data from 55 HCOs within the United States. The results show that
494 our models perform well on all geographic and racial groups and generalize well
495 across time. The network also enabled us to simulate deployment of the model
496 over time to identify high-risk individuals across the entire network.

497 The eventual clinical deployment of PDAC risk prediction models depends
498 not only on model accuracy and generalizability, but also on productive inte-
499 gration into EHR systems for inclusion into the clinical workflow. Lack of sys-
500 tem integration and model automation comprises a significant barrier to clinical
501 adoption of such models [28]. Because of their close integration with existing

502 HCO EHR systems, federated networks can solve these integration and deploy-
503 ment challenges to provide a clear pathway for integrated model development,
504 validation, and clinical deployment all within a single federated system [25].

505 **4.5 Related work**

506 Other researchers have used EHR data to develop PDAC risk prediction models
507 for the general population [3, 5, 7, 8, 22]. Data set sizes range from 1,792 PDAC
508 cases/1.8M controls [8] to 24,000 PDAC cases/6.2M controls [22]. Some studies
509 lack an external validation [7], complete the external validation/evaluate model
510 generalizability only with data from a single geographic area [3, 22], or validate
511 only on one gender (male) [8] or race [15]. While some studies work with data
512 obtained from multiple organizations [7, 8, 22], none work with a federated data
513 network that harmonizes and standardizes the data, none provides a clear path
514 to clinical deployment, and none supports the seamless deployment of the model
515 to new HCOs as they join the federated network.

516 Some previous studies evaluate the ability of their models to identify high-risk
517 individuals either until or shortly before the date of PDAC diagnosis [7, 8, 22],
518 when clinical benefit is improbable. To focus on time frames in which detection
519 of early stage disease and potential cure are most likely, we evaluate the ability
520 of our models to identify high-risk patients at least six months before diagnosis.

521 **4.6 Limitations**

522 Our study has limitations. Notably, model development and validation were
523 retrospective. Prospective studies are needed to evaluate efficacy of clinical de-
524 tection of early stage disease in high-risk individuals.

525 Our results also show that our models performed well on data from the
526 TriNetX network, including multiple HCOs located in different geographic re-
527 gions across the United States. We do not know, however, if our models will
528 perform similarly on data from different sources or different countries. Future
529 work should evaluate the models on data from different EHR sources and pop-
530 ulations selected from different countries and global regions.

531 The use of neural networks and the fact that our model needed thousands of
532 features to reach its best performance make it harder to interpret the reasoning
533 process or extract knowledge for clinicians. Future work should try to gain a
534 deeper understanding of how the model makes predictions and to simplify the
535 model if possible.

536 **5 Conclusion**

537 In conclusion, we have built, validated, and simulated deployment of a PDAC
538 risk prediction model for the general population on multi-institutional EHR data
539 from a federated network. This model can be used to help primary care providers
540 across the country identify high-risk individuals for PDAC screening or used

541 as a first filter before subsequent biomarker testing. The model maintained its
542 accuracy across diverse racial groups and geographic regions in the US, as well
543 as over time, and outperformed widely-used clinical guideline criteria [10, 12] for
544 inclusion of individuals into PDAC screening programs.

545 Our approach enables potential expansion of the population targeted for
546 screening beyond the traditionally screened minority with an inherited predispo-
547 sition. To our knowledge, this is the first PDAC risk prediction model developed,
548 externally validated, with simulated deployment, using a federated network. The
549 developed models set the stage for deployment of the model within the network
550 to identify high risk patients at multiple institutions within the network. A
551 prospective study to validate the models before full clinical deployment is the
552 next step.

553 Acknowledgment

554 We are grateful to Gadi Lachman and TriNetX for providing support and re-
555 sources for this work. We thank Lydia González for her help on identifying and
556 mitigating data quality issues. We also thank the Prevent Cancer Foundation
557 for supporting this work (LA).

558 **Funding:** LA acknowledges support from the Prevent Cancer Foundation for
559 this work. MR, LA, KJ acknowledge the contribution of resources by TriNetX,
560 including secured laptop computers, access to the TriNetX EHR database, and
561 clinical, technical, legal, and administrative assistance from the TriNetX team of
562 clinical informaticists, engineers, and technical staff. MR and KJ received fund-
563 ing from DARPA and Boeing. MR also received funding from the NSF, Aarno
564 Labs, and Boeing. During the time the research was performed MR consulted
565 for Comcast, Google, Motorola, and Qualcomm.

566 **Author contributions:** LA, MR, KJ, SK conceptualization. Data acquisi-
567 tion KH, JW, KJ. Data curation KJ, MR, LA. Data interpretation KJ, MR, LA,
568 MP, IDK. Project administration LA, MR, KH. Supervision MR, LA, SK, MP.
569 ALL writing review and editing. ALL approved published version and agreed to
570 be accountable for all aspects of the work.

571 **Competing interests:** JK and MR are not aware of any payments or ser-
572 vices, paid to themselves or MIT, that could be perceived to influence the sub-
573 mitted work. LA is not aware of any payments or services, paid to her or BIDMC,
574 that could be perceived to influence the submitted work.

575 **Data availability:** The de-identified data in TriNetX federated network
576 database can only be accessed by researchers that are either part of the network
577 or have a collaboration agreement with TriNetX. As stated in the manuscript,
578 we accessed data as part of a no-cost collaboration agreement between BIDMC,
579 MIT, and TriNetX.

580 References

- 581 1. Surveillance, epidemiology, and end results (SEER) program SEER*Stat
582 database: Incidence - SEER research limited-field data, 22 registries, nov
583 2021 sub (2000-2019) - linked to county attributes time dependent (1990-
584 2019) income/rurality, 1969-2020 counties (2022), <https://www.seer.cancer.gov>,
585 Released April 2022, based on the November 2021 submission
- 586 2. Agniel, D., Kohane, I.S., Weber, G.M.: Biases in electronic health record data
587 due to processes within the healthcare system: retrospective observational
588 study. *BMJ* **361** (2018)
- 589 3. Appelbaum, L., Cambronero, J.P., Stevens, J.P., Horng, S., Pollick, K., Silva,
590 G., Haneuse, S., Piatkowski, G., Benhaga, N., Duey, S., et al.: Development
591 and validation of a pancreatic cancer risk model for the general population
592 using electronic health records: An observational study. *European Journal*
593 *of Cancer* **143**, 19–30 (2021)
- 594 4. Aslanian, H.R., Lee, J.H., Canto, M.I.: Aa clinical practice update on pan-
595 creas cancer screening in high-risk individuals: expert review. *Gastroenterol-*
596 *ogy* **159**(1), 358–362 (2020)
- 597 5. Baecker, A., Kim, S., Risch, H.A., Nuckols, T.K., Wu, B.U., Hendifar, A.E.,
598 Pandol, S.J., Pisegna, J.R., Jeon, C.Y.: Do changes in health reveal the pos-
599 sibility of undiagnosed pancreatic cancer? development of a risk-prediction
600 model based on healthcare claims data. *PloS one* **14**(6), e0218580 (2019)
- 601 6. Canto, M.I., Harinck, F., Hruban, R.H., Offerhaus, G.J., Poley, J.W., Kamel,
602 I., Nio, Y., Schulick, R.S., Bassi, C., Kluijdt, I., et al.: International cancer
603 of the pancreas screening (CAPS) consortium summit on the management
604 of patients with increased risk for familial pancreatic cancer. *Gut* **62**(3),
605 339–347 (2013)
- 606 7. Chen, Q., Cherry, D.R., Nalawade, V., Qiao, E.M., Kumar, A., Lowy, A.M.,
607 Simpson, D.R., Murphy, J.D.: Clinical data prediction model to identify
608 patients with early-stage pancreatic cancer. *JCO Clinical Cancer Informatics*
609 **5**, 279–287 (2021)
- 610 8. Chen, W., Zhou, Y., Xie, F., Butler, R.K., Jeon, C.Y., Luong, T.Q., Lin,
611 Y.C., Lustigova, E., Pisegna, J.R., Kim, S., et al.: Prediction model for
612 detection of sporadic pancreatic cancer (pro-TECT) in a population-based co-
613 hort using machine learning and further validation in a prospective study.
614 *medRxiv* (2022)
- 615 9. Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G.: Transparent re-
616 porting of a multivariable prediction model for individual prognosis or di-
617 agnosis (TRIPOD): the TRIPOD statement. *Journal of British Surgery* **102**(3),
618 148–158 (2015)
- 619 10. Daly, M.B., Pal, T., AlHilli, Z., Arun, B., Buys, S.S., Cheng, H., Churpek,
620 J., Domchek, S.M., Elkhanany, A., Friedman, S., Giri, V., Goggins, M.,
621 Hagemann, A., Hendrix, A., Hutton, M.L., Karlan, B.Y., Kassem, N.,
622 Khan, S., Klein, C., Kohlmann, W., Kurian, A.W., Laronga, C., Mak,
623 J.S., Mansour, J., Maxell, K., McDonnell, K., Menendez, C.S., Merajver,
624 S.D., Norquist, B.S., Offit, K., Reiser, G., Senter-Jamieson, L., Shannon,

- 625 K.M., Shatsky, R., Visvanathan, K., Welborn, J., Wick, M.J., Yurgelun,
626 M.B., et al.: Genetic/familial high-risk assessment: Breast, ovarian, and
627 pancreatic (2023), [https://www.nccn.org/professionals/physician_gls/pdf/
628 genetics_bop.pdf](https://www.nccn.org/professionals/physician_gls/pdf/genetics_bop.pdf), Accessed: 1-21-2023
- 629 11. Defazio, A., Bach, F., Lacoste-Julien, S.: Saga: A fast incremental gradi-
630 ent method with support for non-strongly convex composite objectives. *Ad-
631 vances in neural information processing systems* **27** (2014)
- 632 12. Goggins, M., Overbeek, K.A., Brand, R., Syngal, S., Del Chiaro, M., Bartsch,
633 D.K., Bassi, C., Carrato, A., Farrell, J., Fishman, E.K., et al.: Management
634 of patients with increased risk for familial pancreatic cancer: updated recom-
635 mendations from the international cancer of the pancreas screening (caps)
636 consortium. *Gut* **69**(1), 7–17 (2020)
- 637 13. Humphris, J.L., Johns, A.L., Simpson, S.H., Cowley, M.J., Pajic, M., Chang,
638 D.K., Nagrial, A.M., Chin, V.T., Chanthrill, L.A., Pinese, M., et al.: Clinical
639 and pathologic features of familial pancreatic cancer. *Cancer* **120**(23), 3669–
640 3675 (2014)
- 641 14. Jia, K., Rinard, M.: Efficient exact verification of binarized neural networks.
642 In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.)
643 *Advances in Neural Information Processing Systems*, vol. 33, pp. 1782–1795,
644 Curran Associates, Inc. (2020)
- 645 15. Kim, J., Yuan, C., Babic, A., Bao, Y., Clish, C.B., Pollak, M.N., Amundadot-
646 tir, L.T., Klein, A.P., Stolzenberg-Solomon, R.Z., Pandharipande, P.V.,
647 et al.: Genetic and circulating biomarker data improve risk prediction for
648 pancreatic cancer in the general population. *Cancer Epidemiology, Biomark-
649 ers & Prevention* **29**(5), 999–1008 (2020)
- 650 16. Klein, A.P., Lindström, S., Mendelsohn, J.B., Steplowski, E., Arslan, A.A.,
651 Bueno-de Mesquita, H.B., Fuchs, C.S., Gallinger, S., Gross, M., Helzlsouer,
652 K., et al.: An absolute risk model to identify individuals at elevated risk for
653 pancreatic cancer in the general population. *PloS one* **8**(9), e72311 (2013)
- 654 17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436–
655 444 (2015)
- 656 18. Lu, C., Xu, C.F., Wan, X.Y., Zhu, H.T., Yu, C.H., Li, Y.M.: Screening
657 for pancreatic cancer in familial high-risk individuals: A systematic review.
658 *World journal of gastroenterology: WJG* **21**(28), 8678 (2015)
- 659 19. Muhammad, W., Hart, G.R., Nartowt, B., Farrell, J.J., Johung, K., Liang,
660 Y., Deng, J.: Pancreatic cancer prediction through an artificial neural net-
661 work. *Frontiers in Artificial Intelligence* **2**, 2 (2019)
- 662 20. Owens, D.K., Davidson, K.W., Krist, A.H., Barry, M.J., Cabana, M.,
663 Caughey, A.B., Curry, S.J., Doubeni, C.A., Epling, J.W., Kubik, M., et al.:
664 Screening for pancreatic cancer: Us preventive services task force reaffirma-
665 tion recommendation statement. *Jama* **322**(5), 438–444 (2019)
- 666 21. Petersen, G.M.: Familial pancreatic cancer. In: *Seminars in oncology*, vol. 43,
667 pp. 548–553, Elsevier (2016)
- 668 22. Placido, D., Yuan, B., Hjaltelin, J.X., Haue, A.D., Chmura, P.J., Yuan,
669 C., Kim, J., Umeton, R., Antell, G., Chowdhury, A., Franz, A., Brais, L.,
670 Andrews, E., Marks, D.S., Regev, A., Kraft, P., Wolpin, B.M., Rosenthal,

- 671 M., Brunak, S., Sander, C.: Pancreatic cancer risk predicted from disease
672 trajectories using deep learning. *BioRxiv* (2021), [https://doi.org/10.1101/
673 2021.06.27.449937](https://doi.org/10.1101/2021.06.27.449937)
- 674 23. Platt, J., et al.: Probabilistic outputs for support vector machines and com-
675 parisons to regularized likelihood methods. *Advances in large margin classi-
676 fiers* **10**(3), 61–74 (1999)
- 677 24. Porter, N., Laheru, D., Lau, B., He, J., Zheng, L., Narang, A., Roberts, N.J.,
678 Canto, M.I., Lennon, A.M., Goggins, M.G., et al.: Risk of pancreatic cancer
679 in the long-term prospective follow-up of familial pancreatic cancer kindreds.
680 *JNCI: Journal of the National Cancer Institute* **114**(12), 1681–1688 (2022)
- 681 25. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S.,
682 Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al.: The future
683 of digital health with federated learning. *NPJ digital medicine* **3**(1), 1–7
684 (2020)
- 685 26. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer statistics, 2022.
686 *CA: A Cancer Journal for Clinicians* **72**(1), 7–33 (2022), [https://doi.org/
687 https://doi.org/10.3322/caac.21708](https://doi.org/https://doi.org/10.3322/caac.21708)
- 688 27. Topaloglu, U., Palchuk, M.B.: Using a federated network of real-world data
689 to optimize clinical trials operations. *JCO clinical cancer informatics* **2**, 1–10
690 (2018)
- 691 28. Videha Sharma, I.A., van der Veer, S., Martin, G., Ainsworth, J., Augustine,
692 T.: Adoption of clinical risk prediction tools is limited by a lack of integration
693 with electronic health records. *BMJ Health & Care Informatics* **28**(1) (2021)