A Pancreatic Cancer Risk Prediction Model (PRISM) Developed and Validated on Large-scale US Clinical Data (Supplemental Material)

Kai Jia[1], Steven Kundrot[2], Matvey B. Palchuk[2], Jeff Warnick[2], Kathryn Haapala[2], Irving D. Kaplan[3], Martin Rinard[1] *, and Limor Appelbaum[3] * †

[1] Department of Electrical Engineering and Computer Science,
Massachusetts Institute of Technology, Cambridge MA 02139 USA
jiakai@mit.edu
rinard@csail.mit.edu*
[2] TriNetX, LLC, Cambridge MA 02140 USA
steve.kundrot@trinetx.com
matvey.palchuk@trinetx.com
jeff.warnick@trinetx.com
kathryn.haapala@trinetx.com
[3] Beth Israel Deaconess Medical Center, Boston MA 02215 USA
ikaplan@bidmc.harvard.edu
lappelb1@bidmc.harvard.edu*†

A Supplemental material

A.1 More demographics distribution

Table A1 presents more detailed demographics of our dataset.

A.2 Feature extraction details

We used the following ICD codes to define the PDAC group

- C25.0 Malignant neoplasm of head of pancreas
- C25.1 Malignant neoplasm of body of pancreas
- C25.2 Malignant neoplasm of tail of pancreas
- C25.3 Malignant neoplasm of pancreatic duct
- C25.7 Malignant neoplasm of other parts of pancreas
- C25.8 Malignant neoplasm of overlapping sites of pancreas
- C25.9 Malignant neoplasm of pancreas, unspecified
- 157 Malignant neoplasm of pancreas (ICD-9 without a corresponding ICD-10 code)

As stated in the body text, our features were derived from demographics, diagnosis, medication, and lab entries in the EHR given a cutoff date C. The feature extraction excludes all entries after C. We defined the date of PDAC diagnosis D to be the first time a PDAC ICD code appeared in the patient's medical record. If the patient has tumour registry records, D is the earliest of PDAC diagnosis and PDAC tumour registry. During training or testing, we sampled the cutoff dates for PDAC cases uniformly between $[D - 18 \text{ months}, D - 6 \text{ months}]$. We sampled random cutoff dates for control patients matched with the

* Co-senior authors.
† Corresponding author

Table A1: More detailed demographics of our dataset

Location Race	N	Cancer group				Control group			
		N (%)	Age median(IQR)	Female n(%)	Male n(%)	N (%)	Age median(IQR)	Female n(%)	Male n(%)
Midwest	238,459	8,371 (3·51)	68·0 (59·6 to 75·9)	4,527 (1·90)	3,844 (1·61)	230,088 (96·49)	60·1 (49·4 to 70·1)	127,034 (53·27)	103,033 (43·21)
AIAN	648	21 (3·24)	65·4 (58·1 to 71·9)	12 (1·85)	9 (1·39)	627 (96·76)	55·0 (45·8 to 65·2)	374 (57·72)	252 (38·89)
Asian	3,299	66 (2·00)	64·2 (54·9 to 74·5)	38 (1·15)	28 (0·85)	3,233 (98·00)	53·3 (43·5 to 65·4)	1,798 (54·50)	1,433 (43·44)
Black	29,142	853 (2·93)	63·4 (56·5 to 71·3)	510 (1·75)	343 (1·18)	28,289 (97·07)	55·5 (45·8 to 64·6)	16,949 (58·16)	11,339 (38·91)
NHPI	165	2 (1·21)	69·9 (N/A)	0 (0·00)	2 (1·21)	163 (98·79)	56·9 (45·4 to 67·1)	92 (55·76)	71 (43·03)
White	176,579	6,181 (3·50)	68·0 (59·6 to 75·8)	3,330 (1·89)	2,851 (1·61)	170,398 (96·50)	60·9 (50·4 to 70·7)	93,419 (52·90)	76,963 (43·59)
Unknown	28,626	1,248 (4·36)	71·6 (63·4 to 78·7)	637 (2·23)	611 (2·13)	27,378 (95·64)	61·3 (49·3 to 72·0)	14,402 (50·31)	12,975 (45·33)
Northeast	438,300	11,831 (2·70)	69·1 (61·6 to 76·0)	6,015 (1·37)	5,815 (1·33)	426,469 (97·30)	59·9 (49·4 to 69·6)	228,598 (52·16)	176,606 (40·29)
AIAN	894	19 (2·13)	65·1 (60·1 to 72·7)	11 (1·23)	8 (0·89)	875 (97·87)	56·2 (46·4 to 66·0)	489 (54·70)	320 (35·79)
Asian	9,885	165 (1·67)	68·3 (60·2 to 75·7)	82 (0·83)	83 (0·84)	9,720 (98·33)	54·8 (44·6 to 66·5)	5,524 (55·88)	3,900 (39·45)
Black	58,787	1,687 (2·87)	66·6 (58·7 to 73·7)	1,029 (1·75)	658 (1·12)	57,100 (97·13)	56·2 (46·6 to 65·8)	34,060 (57·94)	22,863 (38·89)
NHPI	360	9 (2·50)	62·3 (56·7 to 70·9)	6 (1·67)	3 (0·83)	351 (97·50)	54·1 (44·4 to 65·8)	189 (52·50)	149 (41·39)
White	314,965	8,755 (2·78)	69·3 (62·0 to 76·3)	4,289 (1·36)	4,465 (1·42)	306,210 (97·22)	60·8 (50·6 to 70·2)	159,168 (50·54)	126,914 (40·29)
Unknown	53,409	1,196 (2·24)	70·4 (63·6 to 76·7)	598 (1·12)	598 (1·12)	52,213 (97·76)	59·8 (48·1 to 70·0)	29,168 (54·61)	22,460 (42·05)
South	694,663	12,246 (1·76)	67·9 (60·1 to 74·9)	6,331 (0·91)	5,915 (0·85)	682,417 (98·24)	59·1 (48·6 to 69·0)	395,968 (57·00)	286,375 (41·23)
AIAN	1,908	25 (1·31)	72·1 (68·3 to 75·1)	14 (0·73)	11 (0·58)	1,883 (98·69)	55·4 (46·4 to 65·3)	1,126 (59·01)	757 (39·68)
Asian	16,223	189 (1·17)	68·3 (61·5 to 74·8)	111 (0·68)	78 (0·48)	16,034 (98·83)	54·8 (44·8 to 67·0)	9,595 (59·14)	6,437 (39·68)
Black	135,793	2,631 (1·94)	65·3 (58·0 to 72·7)	1,529 (1·13)	1,102 (0·81)	133,162 (98·06)	56·4 (46·8 to 65·6)	82,189 (60·53)	50,965 (37·53)
NHPI	1,073	6 (0·56)	75·1 (69·8 to 80·5)	2 (0·19)	4 (0·37)	1,067 (99·44)	57·0 (46·0 to 67·4)	640 (59·65)	427 (39·79)
White	468,387	8,623 (1·84)	69·0 (61·1 to 75·7)	4,272 (0·91)	4,351 (0·93)	459,764 (98·16)	60·4 (49·7 to 70·2)	262,117 (55·96)	197,618 (42·19)
Unknown	71,279	772 (1·08)	65·0 (57·6 to 72·7)	403 (0·57)	369 (0·52)	70,507 (98·92)	57·3 (47·0 to 67·3)	40,301 (56·54)	30,171 (42·33)
West	123,556	2,595 (2·10)	67·7 (59·8 to 74·4)	1,295 (1·05)	1,300 (1·05)	120,961 (97·90)	58·5 (47·3 to 68·7)	66,106 (53·50)	54,850 (44·39)
AIAN	2,166	28 (1·29)	62·2 (54·0 to 70·6)	13 (0·60)	15 (0·69)	2,138 (98·71)	53·2 (43·9 to 63·7)	1,197 (55·26)	941 (43·44)
Asian	2,865	74 (2·58)	67·9 (59·3 to 74·9)	38 (1·33)	36 (1·26)	2,791 (97·42)	58·5 (47·1 to 69·6)	1,601 (55·88)	1,190 (41·54)
Black	5,802	115 (1·98)	64·2 (58·4 to 69·0)	61 (1·05)	54 (0·93)	5,687 (98·02)	54·7 (44·5 to 63·8)	3,002 (51·74)	2,685 (46·28)
NHPI	181	4 (2·21)	61·8 (56·0 to 68·4)	4 (2·21)	0 (0·00)	177 (97·79)	53·2 (43·0 to 62·0)	105 (58·01)	72 (39·78)
White	81,819	1,815 (2·22)	68·7 (60·8 to 75·3)	884 (1·08)	931 (1·14)	80,004 (97·78)	59·8 (48·3 to 69·7)	43,625 (53·32)	36,379 (44·46)
Unknown	30,723	559 (1·82)	65·4 (57·8 to 72·5)	295 (0·96)	264 (0·86)	30,164 (98·18)	56·3 (45·9 to 66·5)	16,576 (53·95)	13,583 (44·21)
Unknown	40,490	344 (0·85)	70·6 (62·3 to 76·1)	173 (0·43)	171 (0·42)	40,146 (99·15)	62·9 (52·0 to 73·0)	23,336 (57·63)	16,810 (41·52)
AIAN	4	0 (0·00)	N/A	0 (0·00)	0 (0·00)	4 (100·00)	59·6 (55·3 to 64·5)	3 (75·00)	1 (25·00)
Asian	1,230	10 (0·81)	73·0 (68·3 to 78·4)	5 (0·41)	5 (0·41)	1,220 (99·19)	63·1 (50·5 to 73·5)	758 (61·63)	462 (37·56)
Black	4,047	29 (0·72)	64·9 (60·1 to 72·4)	19 (0·47)	10 (0·25)	4,018 (99·28)	58·5 (48·7 to 68·0)	2,367 (58·49)	1,651 (40·80)
NHPI	125	0 (0·00)	N/A	0 (0·00)	0 (0·00)	125 (100·00)	55·2 (41·0 to 66·3)	75 (60·00)	50 (40·00)
White	30,124	260 (0·86)	70·8 (63·1 to 76·1)	123 (0·41)	137 (0·45)	29,864 (99·14)	64·1 (53·1 to 73·9)	17,242 (57·24)	12,622 (41·90)
Unknown	4,960	45 (0·91)	72·4 (62·6 to 76·4)	26 (0·52)	19 (0·38)	4,915 (99·09)	59·7 (49·4 to 70·6)	2,891 (58·29)	2,024 (40·81)

Race abbreviations:

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

Patients with unknown sex were excluded from the female/male breakdown.

distribution of the PDAC diagnosis dates. For a control patient with a known death date, we limited the cutoff date to at most 18 months before death, to rule out undiagnosed PDAC that caused death. To avoid undiagnosed PDAC cases, we limited all cutoff dates of patients in the control group to be at most 18 months before the dataset query date. We also required the cutoff date of a control patient to be at least at 38.5 years old (40 years - 18 months). In the temporal validation experiments, we limited the cutoff date of control patients to earlier than 18 months before the data split dates, to simulate model training with datasets queried on the data split dates.

During training and testing, we further filtered patients based on the availability of EHR data before the cutoff dates. We only worked with patients with sufficient medical history up to the cutoff date. We empirically defined any patient with at least 16 diagnosis, medication, or lab entries in total within 2 years before their cutoff date and whose first entry is at least 3 months earlier than their last entry before the cutoff date to have *sufficient medical history*. We excluded patients that did not have sufficient medical history given determined cutoff dates.

For each patient, we defined six basic features including age, whether age is known, sex, whether sex is known, number of diagnosis, medication, or lab entries in the EHR up to 18 months before cutoff (the recent entries), and number of diagnosis, medication, or lab entries in the EHR within 18 months and five years before cutoff (the early entries). Age is calculated based on the birth date entry and the cutoff date, linearly normalised to $[0, 1]$ with 40 years old being 0 and 90 years old being 1.

Besides the basic features, we included features that correspond to individual diagnosis, medication, or lab codes, with the code empirically included for feature extraction if it appeared in the EHR of at least 1% of the patients in the cancer group of the training set.

We manually grouped 827 commonly used diagnosis codes into 39 groups. For ungrouped codes, we used the ICD-10 category plus the first digit of the subcategory. We derived 3 features for each diagnosis code or group: whether or not it exists $\{0, 1\}$, its first and last date (encoding for first and last dates: linear between 0 and 1; 0 for greater or equal to 4 years before cutoff; 1 for at cutoff). To use past ICD-9 data to train the model for use on current and future ICD-10 data, we mapped all ICD-9 codes to their ICD-10 equivalents. For ICD-9 codes that could be mapped to more than one ICD-10 code, the feature vector included all the mapped ICD-10 codes.

We manually grouped 67 medication codes into 8 different medication classes. Ungrouped codes were used as they are. We derived 4 features for each medication code: whether or not it exists $\{0, 1\}$, its frequency (i.e., number of times it appears in the EHR), span (time between first and last appearance of a medication code, linearly encoded up to 1 when the time is 4 years), and last date (same encoding as diagnosis first/last date). Medication frequency was linearly encoded to $[0, 1]$ by dividing by the 96% quantile of frequencies of the medication code on the control training set.

For lab features, we used a grouping provided by TriNetX for similar lab tests, which had 98 groups for 462 codes. Ungrouped codes were used as they are. For each lab code or group, we derived 8 features: existence, frequency, first date of a valid value (same encoding as diagnosis first date), last date of a valid value (same encoding as diagnosis last date), most recent valid value, whether a valid value is known, slope, and whether slope can be computed. The frequency was the number of lab results within five years before cutoff. Lab frequency was linearly encoded to $[0, 1]$ by dividing by the 96% quantile of frequencies of the lab code on the control training set. Slope was measured by calculating the yearly change in lab test values within four years before cutoff, normalised by population standard deviation.

The number of recent entries, number of early entries, and values of a lab code have a large dynamic range with a long tail, for which normal distribution is a poor approximation. Therefore, we used quantile normalisation to encode the values. The quantile normalisation maps a value to its quantile on the control training set, which is equivalent to the percentage-point function of the value distribution. The quantile normalisation was calculated by precomputing 16 equally spaced quantiles of number of entries or lab values on the control training set and estimating the inverse cumulative distribution function of a to-be-normalised value with linear interpolation based on the precomputed quantiles. For lab codes with less than 128 valid values on the control training set, we did not extract the value-related lab features.

All features corresponding to individuals diagnosis, medication, and lab codes, along with the six basic features, were concatenated to form a feature vector.

A.3 Model training and evaluation details

All of our code was implemented with `Python 3.10.6`. We used a cloud instance with 16 AMD EPYC 7R13 CPU cores, 123 GiB of RAM, and Ubuntu 22.04 to conduct all our training and evaluation experiments.

PrismNN has three fully connected layers. Each layer has 64, 20, and 1 output neurons. Hidden layers use the `tanh` nonlinearity. Using a vanilla neural network on all the 5459 features, the model achieved training set AUC $0{\cdot}989$ (95% CI: $0{\cdot}988$ to $0{\cdot}989$) and test set AUC $0{\cdot}764$ (95% CI: $0{\cdot}757$ to $0{\cdot}771$), suggesting severe overfitting. To ameliorate overfitting, we used sparse weights computed by the recently developed `BinMask`

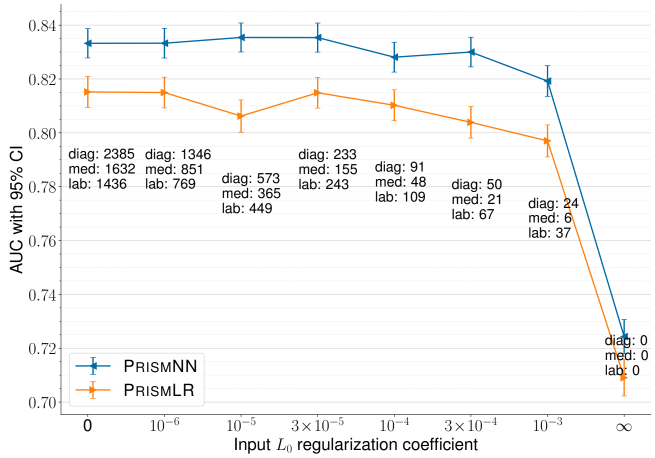

Fig. A1: Model performance with different numbers of features selected by input L_0 regularisation. All models also used the six `basic` features besides the indicated number of features. The label `diag` refers to diagnosis features, `med` to medication features, and `lab` to lab features.

sparsification technique[1]. We used balanced numbers of PDAC and control patients in each mini-batch. PrismLR used the SAGA solver[2] with balanced class weights during training.

The training procedure used a softened version of medical history sufficiency requirement to better utilise available training data. The training set does not exclude patients based on the medical history sufficiency requirement. Given a cutoff date C for a patient, model training used a continuous score $S(C)$ that represents the sufficiency of the medical history of this patient. Let t_i denote the date of medical record entry i (either a diagnosis, medication, or a lab entry). We define $S(C) \stackrel{\text{def}}{=} P(C)Q(C)$, where:

$$P(C) \stackrel{\text{def}}{=} \min\left\{1,\ \frac{1}{16}\sum_{i:\, t_i \le C} \min\left\{1,\ \exp\left(-(\log 100)\left(\frac{C-t_i}{365}-2\right)\right)\right\}\right\}$$

$$Q(C) \stackrel{\text{def}}{=} \min\left\{1,\ \frac{1}{90}\left(\max\left\{t_i \mid t_i \le C\right\} - \min\left\{t_i \mid t_i \le C\right\}\right)\right\}$$

Note that $S(C) = 1$ is equivalent to having sufficient medical history defined above. All test datasets or deployment only included patients with $S(C) = 1$ as stated earlier. The components $P(C)$ and $Q(C)$ are smoothed versions of the two requirements of sufficient medical history. For example, if a patient has lots of EHR entries at 2.1 years before the cutoff date, they may still be included in training (with probably a smaller sample weight), but not in testing.

In each iteration during PrismNN training, we randomly sampled patients in the training set, sampled new cutoff dates for each patient, and put patients with $S(C) \ge 0.5$ into the mini-batch until the desired mini-batch size was achieved. The value of $S(C)$ is also the weight for an individual instance in the cross-entropy loss. When training PrismLR, we sampled a cutoff date for every patient in the training set and removed patients with $S(C) < 0.5$ to produce features for training. During testing, we sampled a cutoff date for every patient in the test set and kept patients with $S(C) = 1$ to produce the same sets of features used by both PrismNN and PrismLR.

We trained the PrismNN models with `pytorch 1.12.1`. For each training task, we trained one dense model and three sparse models with `BinMask` L_0 regularisation coefficients being 2×10^{-5}, 3×10^{-5}, and 4×10^{-5}, respectively. We selected the model with the highest partial AUC with up to 6% FPR on the validation set. We used the AdamW optimiser[3] with a weight decay of 10^{-2} on the weights of fully-connected layers. Each mini-batch consisted of 256 cases from the PDAC group and 256 cases from the control group, with randomly sampled cutoff dates per patient. Each epoch had 1000 mini-batches. The models were trained for 16 epochs (which is about two full iterations over the control group). We used the cosine learning rate annealing[4] to schedule learning rate from 2×10^{-3} to 5×10^{-5}. During training, we also used data augmentation to improve the robustness against typical noises in EHR databases. Data augmentation included randomly perturbing numerical values, randomly masking out EHR entries, and randomly removing demographics information (sex and birth date).

We trained the PrismLR models with `sklearn 1.1.1`. For each training task, we trained four models with L_2 regularisation coefficients 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2}, respectively. We selected the model with the highest partial AUC with up to 6% FPR on the validation set. We used balanced class weights. We used the SAGA solver[2] with 12 iterations when the total training set size (i.e., number of training instances multiplied with number of features per instance) exceeds $2^{30} = 1073741824$, and the L-BFGS solver[5] with 500 iterations when the training set size is below 2^{30}. With our default training/test split, L-BFGS was used when there are at most 665 features. Implementations of both solvers were provided by `sklearn`. We used the default values for other hyperparameters.

For feature selection, we used the same network architecture as PrismNN. We set the L_0 regularisation coefficients for network weights to 5×10^{-6}. We then applied a L_0-regularised binary mask to the inputs, with the regularisation coefficient $\lambda \in \{0,\, 10^{-6},\, 10^{-5},\, 3 \times 10^{-5},\, 10^{-4},\, 3 \times 10^{-4},\, 10^{-3},\, \infty\}$. A larger regularisation coefficient results in a smaller number of features selected. Fig. A1 shows model AUCs with different numbers of input features. The performance decreased with larger regularisation coefficients. Based on those results, we chose $\lambda = 3 \times 10^{-4}$, which delivered PrismNN AUC 0·830 (95% CI: 0·824 to 0·836) and PrismLR AUC 0·804 (95% CI: 0·798 to 0·810) with 144 features, compared to PrismNN AUC 0·833 (95% CI: 0·828 to 0·839) and PrismLR AUC 0·815 (95% CI: 0·809 to 0·821) with all the 5459 features. After feature selection via input L_0 regularisation, we further reduced the number of input features by iteratively removing features. We evaluated the model AUCs on the training set with each single feature removed, and removed the features that resulted in lowest AUC drop. After removing each feature, we finetuned the bias parameter in the first fully connected layer to compensate for bias shift due to less active inputs. We repeated feature removal until the model AUC on the training set was 0·007 times lower than the original AUC. For each training task (e.g., in an external validation setting), we used independent feature selection on the training set.

To calibrate the risk scores, we adopted a modified Platt algorithm[6]. We learned a function $f_\theta(s) = \theta_1 \min(s - s_0,\, 0) + \theta_2 \max(s - s_0,\, 0) + \theta_3$ to map the model output score s to the logits of person-year risk, where $\theta \in \mathbb{R}^3$ is the learnable parameter and s_0 is the median of model output scores on the validation set. We accounted for the unbalanced sampling of control group and estimated the risk on the whole population in calibration by assigning weights inversely proportional to the sampling ratio of control cases in the dataset. We fitted the values of θ for each model independently. We minimised the cross-entropy loss of $f_\theta(s)$ on the validation set to solve θ, which is a linear logistic regression problem. Note that $f_\theta(s)$ is monotonic with respect to s when $\theta_1 > 0$ and $\theta_2 > 0$. Therefore, our risk calibration does not impact the discriminatory power of the model. During testing, We chose 16 risk groups for calibration evaluation as a geometric sequence between the 85% percentile of predicted risk on the test set and the maximum predicted risk.

A.4 Simulated deployment details

The goal of simulated deployment is to simulate model deployment in a clinical study setting to obtain a more accurate estimation of model performance. We trained the model only on data available prior to Apr 11, 2020 (70% percentile of diagnosis dates) in the same way as temporal validation. Then for each date D separated by 90 days after Apr 11, 2020, we

1. Enrolled a new patient into the simulated deployment if the patient had a known age, was at least 40 years old on date D, and had sufficient medical history on D for the first time. We call the date D the *enrolment date* for such a patient.
2. For each enrolled patient, we checked if that patient still had sufficient medical history on D. If so, we evaluated their PDAC risk by our model, with the cutoff date set at D. We call the date D when risk evaluation was performed a *check date* for such a patient.
3. For each enrolled patient who had a PDAC diagnosis, we stopped evaluating this patient's PDAC risk if the date D is within 6 months before the diagnosis date.
4. Stopped enrolling or checking patients if the date D is within 18 months before dataset query date.

We excluded patients who were diagnosed with PDAC either before enrolment or within 6 months after enrolment, patients who had no medical entries between first and last check dates, and patients with a known death but no PDAC diagnosis within 18 months after enrolment. We started following up a patient 6 months after their enrolment date. We stopped following up a patient 18 months after the last check date. During the followup period, we defined the following outcomes:

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

The performance metrics reported in the body text were based on the above definition of outcomes. The sensitivity and specificity confidence intervals were estimated using the exact Clopper-Pearson method[7]. PPV was derived from the sensitivity, specificity, and prevalence. PPV's confidence interval was estimated using Monte Carlo simulation with 100,000 samples from the joint distributions of sensitivity, specificity, and prevalence (assumed to be independent). TriNetX population-level estimation of PPV was calculated by enlarging the size of control group with the sampling ratio, assuming the same false positive characteristics, with a single binomial distribution to model the uncertainty of the combined effect of sampling and any patient exclusion process.

The SIR was calculated as the following. Let H denote the set of high-risk patients predicted by the model (i.e., all false positives and true positives). Let $T : H \mapsto \{0, 1\}$ denote if a patient x is a true positive ($T(x) = 1$) or a false positive ($T(x) = 0$). Let $\mathrm{race}(x)$, $\mathrm{birth}(x)$, $\mathrm{sex}(x)$ denote the race, birth year, and sex of the patient x, respectively. Let $B(x)$ denote the year of first high-risk prediction date of patient x plus six months, and $E(x)$ the year of last followup date of x. Let $\mathrm{SEER}(\cdot)$ denote the SEER database that maps a demographics profile (race, age, sex, calendar year) to a person-year PDAC risk. The SIR is defined as:

$$\mathrm{SIR} \stackrel{\text{def}}{=} \frac{\sum_{x \in H} T(x)}{\sum_{x \in H} \sum_{B(x) \le i \le E(x)} \mathrm{SEER}(\mathrm{race}(x),\, i - \mathrm{birth}(x),\, \mathrm{sex}(x),\, i)}$$

SIR uncertainty estimation included the uncertainty of the SEER risk estimation and the uncertainty introduced by extrapolating the results to the whole TriNetX population assuming the same false positive characteristics.

To calculate population SIR, we replaced H with the set of all the enrolled patients, T the ground truth of PDAC diagnosis, and $\mathrm{B}(x)$ the first followup date of patient x (i.e., 6 months after enrolment). Population

SIRs were $1{\cdot}00$ (95% CI: $0{\cdot}98$ to $1{\cdot}01$) for temporal split and $1{\cdot}03$ (95% CI: $1{\cdot}02$ to $1{\cdot}04$) for random split. Values close to one indicate that the dataset matches the overall PDAC incidence rate of the United States after our patient exclusion.

We report the complete simulated deployment results in Table A2. We also conducted a relaxed version of simulated deployment on the normal test set, where we split the training and test data randomly but not temporally. We chose the first enrolment date to be Jan 1, 2017 (before 2017, yearly PDAC cases were increasing; after 2017, the numbers were mostly stable, which suggests that earlier records might not be fully computerised). As shown in Table A3, both models exhibited only slightly better overall performance compared to the temporal split case, indicating that Prism models have favourable temporal generalizability.

We further analyse the breakdown of simulated deployment performance by race, age, sex, and geographical locations, with high-risk threshold chosen at SIR ≈ 5. We chose one global threshold based on the whole test population. Tables A4 to A9 present the results. The difference between subgroup PPV and overall PPV is a rough assessment of risk calibration within each subgroup. The results signify the need for recalibration if the model is to be deployed to a specific subgroup, especially for certain race and age subgroups. The AUC values, which are independent of recalibration, suggest that Prism models maintain good discrimination power for different subgroups.

A.5 The impact of training set size on model performance

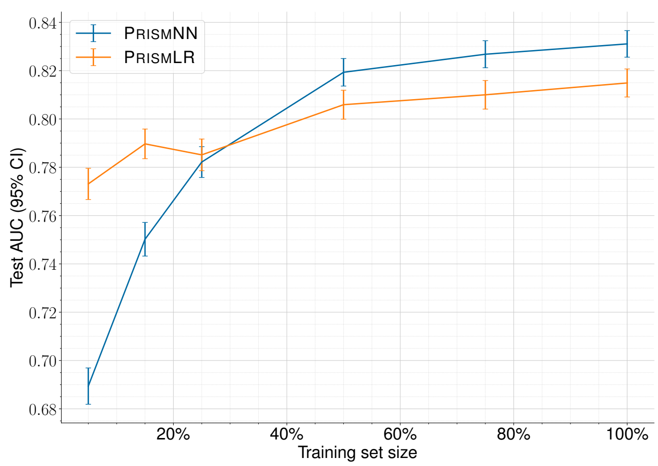

Fig. A2: Model performance with different training set sizes

While neural networks have shown impressive performance on many tasks, some research suggests that they may be less favourable compared to other models on tabular data[8]. It is also suggested that in the majority of clinical cases, NN only provides marginal improvement over LR[9].

We have observed that PrismNN outperformed PrismLR consistently and significantly, especially in the high-specificity region in the simulated deployment. We attribute part of the reason to be our large training data size enabled by the use of a federated EHR network. To investigate how training size impacted the performance of Prism models, we trained PrismNN and PrismLR models on randomly sampled subsets of training data, and evaluated their performance on the same held-out test set. We used all the features without learning-based feature selection (i.e., only selecting codes based on the 1% criteria without using

L_0-regularisation for feature selection) to reduce uncontrolled variables. As shown in Fig. A2, PrismNN outperformed PrismLR when the training set was sufficiently large.

A.6 The impact of data from external locations on model performance

We observed modest performance drop in location-based internal-external validation, which implies that there exist systematic heterogeneity in the EHR data from HCOs in different geographic locations. If such heterogeneity is significant enough, a better strategy is to use location-specific models (i.e., a model trained only on data from one location) to make predictions for patients with known locations, rather than using a unified model trained on all data for everyone.

To test if inter-location heterogeneity is significant enough to warrant location-specific models, we conducted further experiments to investigate how adding data from external locations impacts model performance on each location. As shown in Table A10, both PrismNN and PrismLR benefited from training on data from external locations even when tested on one location. The model obtained by training on all data is not significantly worse than the best model selected a posterior in each case. Therefore, we should use a unified model instead of location-specific models.

A.7 More strict PDAC diagnosis criteria

Our study used one diagnostic code entry in the EHR as the criteria for PDAC diagnosis. It is possible that some patients with suspected PDAC but did not actually develop PDAC had one of those codes. To investigate the impact of using a potentially more specific PDAC definition, we experimented with a more strict PDAC inclusion criteria by requiring two diagnostic code entries or one diagnostic code entry and one tumour registry entry.

After applying this more strict criteria, there were 24,372 patients left in the PDAC group, 31% less than the 35,387 patients reported in the body text. Model performance results exhibited no statistically significant difference. The test AUCs were 0·824 (95% CI: 0·814 to 0·834) vs 0·826 (95% CI: 0·824 to 0·828) for PrismNN and 0·802 (95% CI: 0·790 to 0·814) vs 0·800 (95% CI: 0·798 to 0·802) for PrismLR. Location-based external validation average test AUCs were 0·743 (95% CI: 0·712 to 0·773) vs 0·740 (95% CI: 0·716 to 0·764) for PrismNN and 0·760 (95% CI: 0·744 to 0·775) vs 0·744 (95% CI: 0·727 to 0·762) for PrismLR. Race-based external validation average test AUCs were 0·819 (95% CI: 0·717 to 0·922) vs 0·828 (95% CI: 0·744 to 0·912) for PrismNN and 0·806 (95% CI: 0·713 to 0·900) vs 0·814 (95% CI: 0·740 to 0·888) for PrismLR. Temporal external validation average test AUCs were 0·788 (95% CI: 0·772 to 0·803) vs 0·789 (95% CI: 0·762 to 0·816) for PrismNN and 0·779 (95% CI: 0·764 to 0·795) vs 0·780 (95% CI: 0·763 to 0·798) for PrismLR.

However, the population SIR in simulated deployment dropped from 1·00 (95% CI: 0·98 to 1·01) to 0·62 (95% CI: 0·61 to 0·62). The SIR being much lower than 1 indicates that the filtering is too restrictive for the PDAC group (and asymmetric for the control group) so that there were much fewer cancer cases than expected in the dataset.

Therefore, we did not use this more strict inclusion for the final model.

Table A2: Complete simulated deployment results with temporal training/test split

(a) Study statistics (with 95% CI when applicable)

First enrolment date	Apr 11, 2020	Mean age at enrolment (SD)	61·62 (11·98)
Last check date	Apr 6, 2021	Mean age at PDAC (SD)	69·75 (10·37)
Patients enrolled	185,932	Mean years of followup (SD)	1·82 (0·31)
PDAC cases	7,095	PrismNN single-point test AUC	0·791 (0·787 to 0·796)
PDAC prevalence	3·82% (3·73 to 3·90)	PrismNN sim-dep test AUC	0·793 (0·788 to 0·799)
SIR (TrxPop. Est.)	1·00 (0·98 to 1·01)	PrismLR single-point test AUC	0·780 (0·776 to 0·785)
Total person-years of followup	337 894·14	PrismLR sim-dep test AUC	0·787 (0·781 to 0·792)

(b) PrismNN performance statistics (with 95% CI) at different thresholds

Thresh	Sensitivity	Specificity	PPV	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)
89·00%	54·6% (53·4 to 55·8)	85·3% (85·1 to 85·5)	12·8% (12·5 to 13·2)	0·28% (0·27 to 0·29)	2·38 (2·34 to 2·41)
90·00%	52·8% (51·6 to 54·0)	86·6% (86·5 to 86·8)	13·5% (13·1 to 13·9)	0·30% (0·29 to 0·31)	2·52 (2·48 to 2·56)
91·00%	50·8% (49·6 to 52·0)	87·9% (87·7 to 88·0)	14·3% (13·8 to 14·7)	0·32% (0·31 to 0·32)	2·69 (2·64 to 2·72)
92·00%	48·7% (47·5 to 49·9)	89·1% (89·0 to 89·3)	15·1% (14·6 to 15·6)	0·34% (0·33 to 0·35)	2·87 (2·82 to 2·91)
94·00%	43·7% (42·5 to 44·9)	91·9% (91·7 to 92·0)	17·6% (17·0 to 18·1)	0·40% (0·39 to 0·42)	3·47 (3·42 to 3·52)
96·00%	37·7% (36·5 to 38·8)	94·6% (94·4 to 94·7)	21·5% (20·8 to 22·3)	0·52% (0·50 to 0·54)	4·54 (4·47 to 4·61)
96·20%	37·1% (36·0 to 38·2)	94·8% (94·7 to 94·9)	22·0% (21·3 to 22·8)	0·54% (0·52 to 0·55)	4·69 (4·61 to 4·75)
96·40%	36·6% (35·5 to 37·7)	95·1% (95·0 to 95·2)	22·8% (22·0 to 23·6)	0·56% (0·54 to 0·58)	4·91 (4·83 to 4·98)
96·60%	35·9% (34·8 to 37·1)	95·3% (95·2 to 95·4)	23·4% (22·6 to 24·2)	0·58% (0·56 to 0·60)	5·10 (5·02 to 5·18)
96·80%	35·3% (34·2 to 36·4)	95·6% (95·5 to 95·7)	24·2% (23·3 to 25·0)	0·60% (0·58 to 0·63)	5·36 (5·28 to 5·44)
97·00%	34·5% (33·4 to 35·6)	95·9% (95·8 to 96·0)	24·9% (24·1 to 25·8)	0·63% (0·61 to 0·65)	5·62 (5·53 to 5·70)
97·20%	33·9% (32·8 to 35·0)	96·1% (96·1 to 96·2)	25·8% (25·0 to 26·8)	0·66% (0·63 to 0·69)	5·92 (5·83 to 6·01)
97·40%	33·2% (32·1 to 34·3)	96·4% (96·3 to 96·5)	26·7% (25·8 to 27·6)	0·69% (0·66 to 0·72)	6·24 (6·14 to 6·33)
97·60%	32·3% (31·2 to 33·4)	96·7% (96·6 to 96·8)	27·8% (26·8 to 28·8)	0·73% (0·70 to 0·76)	6·65 (6·54 to 6·74)
97·80%	31·4% (30·4 to 32·5)	96·9% (96·9 to 97·0)	28·9% (27·9 to 29·9)	0·77% (0·74 to 0·80)	7·07 (6·96 to 7·17)
98·00%	30·5% (29·4 to 31·6)	97·2% (97·1 to 97·3)	30·2% (29·1 to 31·3)	0·82% (0·78 to 0·85)	7·59 (7·47 to 7·70)
98·50%	27·9% (26·8 to 28·9)	97·8% (97·8 to 97·9)	33·9% (32·7 to 35·1)	0·97% (0·92 to 1·02)	9·10 (8·96 to 9·23)
99·00%	25·0% (24·0 to 26·0)	98·5% (98·5 to 98·6)	39·9% (38·5 to 41·4)	1·25% (1·19 to 1·32)	12·3 (12·1 to 12·4)
99·50%	20·0% (19·1 to 21·0)	99·2% (99·2 to 99·3)	50·9% (49·0 to 52·7)	1·94% (1·81 to 2·07)	19·8 (19·5 to 20·1)
99·70%	17·3% (16·4 to 18·2)	99·5% (99·5 to 99·6)	60·3% (58·1 to 62·4)	2·81% (2·59 to 3·06)	29·0 (28·6 to 29·5)
99·90%	13·4% (12·6 to 14·2)	99·8% (99·8 to 99·8)	75·2% (72·8 to 77·6)	5·47% (4·86 to 6·18)	58·0 (57·0 to 58·9)
99·95%	11·9% (11·2 to 12·7)	99·9% (99·9 to 99·9)	83·2% (80·7 to 85·4)	8·62% (7·42 to 10·0)	96·0 (94·3 to 97·6)

(c) PrismLR performance statistics (with 95% CI) at different thresholds

Thresh	Sensitivity	Specificity	PPV	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)
89·00%	52·3% (51·1 to 53·4)	86·2% (86·1 to 86·4)	13·1% (12·7 to 13·5)	0·29% (0·28 to 0·29)	2·22 (2·19 to 2·25)
90·00%	50·6% (49·4 to 51·8)	87·4% (87·3 to 87·6)	13·8% (13·3 to 14·2)	0·30% (0·30 to 0·31)	2·35 (2·31 to 2·38)
91·00%	48·4% (47·2 to 49·6)	88·7% (88·5 to 88·8)	14·5% (14·0 to 15·0)	0·32% (0·31 to 0·33)	2·48 (2·44 to 2·52)
92·00%	46·4% (45·2 to 47·6)	89·9% (89·7 to 90·0)	15·4% (14·9 to 15·9)	0·35% (0·34 to 0·36)	2·66 (2·61 to 2·70)
94·00%	40·9% (39·8 to 42·1)	92·3% (92·1 to 92·4)	17·3% (16·8 to 17·9)	0·40% (0·39 to 0·41)	3·06 (3·01 to 3·10)
96·00%	34·2% (33·1 to 35·3)	94·7% (94·6 to 94·8)	20·4% (19·6 to 21·1)	0·49% (0·47 to 0·50)	3·73 (3·67 to 3·79)
96·20%	33·4% (32·3 to 34·5)	94·9% (94·8 to 95·0)	20·7% (20·0 to 21·5)	0·50% (0·48 to 0·51)	3·81 (3·75 to 3·87)
96·40%	32·7% (31·6 to 33·8)	95·2% (95·1 to 95·3)	21·3% (20·5 to 22·0)	0·51% (0·49 to 0·53)	3·95 (3·88 to 4·01)
96·60%	31·8% (30·7 to 32·9)	95·4% (95·3 to 95·5)	21·7% (20·9 to 22·5)	0·53% (0·50 to 0·55)	4·06 (3·99 to 4·12)
96·80%	31·0% (29·9 to 32·1)	95·7% (95·6 to 95·8)	22·1% (21·3 to 23·0)	0·54% (0·52 to 0·56)	4·17 (4·11 to 4·24)
97·00%	30·0% (29·0 to 31·1)	96·0% (95·9 to 96·0)	22·8% (21·9 to 23·6)	0·56% (0·54 to 0·58)	4·32 (4·25 to 4·39)
97·20%	29·2% (28·1 to 30·3)	96·2% (96·1 to 96·3)	23·3% (22·5 to 24·2)	0·58% (0·55 to 0·60)	4·48 (4·41 to 4·55)
97·40%	28·1% (27·1 to 29·2)	96·5% (96·4 to 96·5)	24·0% (23·1 to 24·9)	0·60% (0·57 to 0·63)	4·67 (4·59 to 4·74)
97·60%	27·2% (26·2 to 28·3)	96·7% (96·6 to 96·8)	24·8% (23·9 to 25·8)	0·63% (0·60 to 0·65)	4·89 (4·80 to 4·96)
97·80%	26·4% (25·3 to 27·4)	97·0% (96·9 to 97·1)	25·8% (24·8 to 26·8)	0·66% (0·63 to 0·69)	5·17 (5·09 to 5·25)
98·00%	25·1% (24·1 to 26·1)	97·3% (97·2 to 97·4)	26·9% (25·8 to 27·9)	0·70% (0·66 to 0·73)	5·49 (5·40 to 5·58)
98·50%	21·8% (20·8 to 22·8)	98·0% (97·9 to 98·0)	29·8% (28·6 to 31·1)	0·80% (0·76 to 0·85)	6·47 (6·37 to 6·57)
99·00%	17·5% (16·6 to 18·4)	98·6% (98·6 to 98·7)	33·2% (31·7 to 34·8)	0·94% (0·88 to 1·00)	7·59 (7·46 to 7·71)
99·50%	12·1% (11·4 to 12·9)	99·3% (99·2 to 99·3)	40·2% (38·1 to 42·3)	1·26% (1·16 to 1·37)	10·6 (10·4 to 10·7)
99·70%	8·95% (8·30 to 9·64)	99·6% (99·5 to 99·6)	44·5% (42·0 to 47·1)	1·51% (1·36 to 1·67)	13·1 (12·9 to 13·3)
99·90%	5·20% (4·70 to 5·74)	99·9% (99·8 to 99·9)	59·1% (55·2 to 63·0)	2·69% (2·30 to 3·14)	23·1 (22·7 to 23·5)
99·95%	2·93% (2·55 to 3·35)	99·9% (99·9 to 99·9)	60·8% (55·5 to 66·0)	2·88% (2·32 to 3·56)	24·2 (23·7 to 24·7)

Notes

- Single-point test AUC was obtained by testing the model on one sampled cutoff date per patient; sim-dep test AUC was obtained by varying the high-risk threshold in the simulated deployment.
- Thresh: The threshold for high-risk patients, corresponding to the specificity on validation set.
- PPV: Positive Predictive Value
- SIR: Standardised Incidence Ratio
- TrxPop. Est.: Since we used all the PDAC cases in the TriNetX database but sampled a subset of control patients, we needed to account for this imbalance to estimate the PPV and SIR that would be obtained if we had evaluated the model on the full TriNetX population.

Table A3: Simulated deployment with random (not temporal) training/test split

(a) Study statistics (with 95% CI when applicable)

First enrolment date	Jan 1, 2017	Mean age at enrolment (SD)	59·81 (11·95)
Last check date	Jun 9, 2021	Mean age at PDAC (SD)	68·72 (10·50)
Patients enrolled	274,067	Mean years of followup (SD)	3·63 (1·48)
PDAC cases	3,278	PrismNN single-point test AUC	0·825 (0·819 to 0·830)
PDAC prevalence	1·20% (1·16 to 1·24)	PrismNN sim-dep test AUC	0·803 (0·795 to 0·810)
SIR (TrxPop. Est.)	1·03 (1·02 to 1·04)	PrismLR single-point test AUC	0·798 (0·793 to 0·804)
Total person-years of followup	996 022·39	PrismLR sim-dep test AUC	0·781 (0·773 to 0·789)

(b) PrismNN performance statistics (with 95% CI) at different thresholds

Thresh	Sensitivity	Specificity	PPV	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)
89·00%	60·1% (58·4 to 61·7)	82·2% (82·0 to 82·3)	3·92% (3·75 to 4·09)	0·52% (0·50 to 0·53)	2·14 (2·11 to 2·16)
90·00%	57·8% (56·1 to 59·5)	83·5% (83·4 to 83·7)	4·08% (3·90 to 4·26)	0·54% (0·52 to 0·55)	2·23 (2·20 to 2·25)
91·00%	56·1% (54·4 to 57·8)	85·0% (84·8 to 85·1)	4·32% (4·13 to 4·52)	0·57% (0·55 to 0·59)	2·36 (2·33 to 2·38)
92·00%	54·2% (52·5 to 55·9)	86·5% (86·4 to 86·6)	4·63% (4·43 to 4·85)	0·61% (0·59 to 0·63)	2·54 (2·51 to 2·56)
94·00%	48·6% (46·9 to 50·4)	89·5% (89·4 to 89·6)	5·31% (5·06 to 5·57)	0·71% (0·68 to 0·73)	2·93 (2·89 to 2·96)
96·00%	41·5% (39·9 to 43·3)	92·7% (92·6 to 92·8)	6·43% (6·10 to 6·77)	0·87% (0·83 to 0·90)	3·63 (3·58 to 3·67)
96·20%	41·0% (39·3 to 42·7)	93·0% (92·9 to 93·1)	6·61% (6·27 to 6·96)	0·89% (0·85 to 0·93)	3·75 (3·70 to 3·79)
96·40%	40·3% (38·6 to 42·0)	93·3% (93·2 to 93·4)	6·79% (6·44 to 7·15)	0·92% (0·88 to 0·96)	3·86 (3·81 to 3·90)
96·60%	39·5% (37·8 to 41·2)	93·6% (93·5 to 93·7)	6·97% (6·61 to 7·35)	0·94% (0·90 to 0·99)	3·98 (3·93 to 4·03)
96·80%	38·9% (37·2 to 40·6)	94·0% (93·9 to 94·1)	7·29% (6·91 to 7·69)	0·99% (0·95 to 1·04)	4·19 (4·14 to 4·24)
97·00%	38·2% (36·6 to 39·9)	94·3% (94·3 to 94·4)	7·56% (7·16 to 7·97)	1·03% (0·98 to 1·08)	4·38 (4·33 to 4·43)
97·20%	37·2% (35·6 to 38·9)	94·7% (94·6 to 94·8)	7·82% (7·40 to 8·25)	1·07% (1·02 to 1·12)	4·57 (4·51 to 4·62)
97·40%	36·4% (34·7 to 38·1)	95·0% (94·9 to 95·1)	8·07% (7·64 to 8·51)	1·10% (1·05 to 1·16)	4·76 (4·70 to 4·81)
97·60%	35·6% (33·9 to 37·2)	95·3% (95·3 to 95·4)	8·45% (7·99 to 8·93)	1·16% (1·10 to 1·22)	5·02 (4·96 to 5·08)
97·80%	34·6% (32·9 to 36·2)	95·7% (95·6 to 95·8)	8·87% (8·38 to 9·38)	1·22% (1·16 to 1·28)	5·33 (5·26 to 5·39)
98·00%	33·5% (31·9 to 35·1)	96·1% (96·0 to 96·2)	9·38% (8·85 to 9·92)	1·30% (1·23 to 1·37)	5·69 (5·61 to 5·75)
98·50%	30·7% (29·1 to 32·3)	97·0% (96·9 to 97·1)	11·1% (10·4 to 11·7)	1·56% (1·47 to 1·64)	6·99 (6·90 to 7·07)
99·00%	26·7% (25·2 to 28·2)	98·0% (97·9 to 98·1)	13·9% (13·1 to 14·8)	2·01% (1·89 to 2·14)	9·42 (9·31 to 9·53)
99·50%	21·9% (20·5 to 23·4)	98·9% (98·8 to 98·9)	19·2% (17·9 to 20·4)	2·93% (2·72 to 3·14)	14·6 (14·4 to 14·7)
99·70%	19·8% (18·5 to 21·2)	99·3% (99·2 to 99·3)	24·3% (22·7 to 25·9)	3·92% (3·62 to 4·24)	20·1 (19·8 to 20·3)
99·90%	15·3% (14·1 to 16·6)	99·7% (99·7 to 99·7)	38·3% (35·7 to 40·9)	7·31% (6·61 to 8·06)	37·9 (37·5 to 38·4)
99·95%	13·6% (12·4 to 14·8)	99·8% (99·8 to 99·8)	47·8% (44·5 to 51·1)	10·4% (9·31 to 11·7)	56·5 (55·7 to 57·1)

(c) PrismLR performance statistics (with 95% CI) at different thresholds

Thresh	Sensitivity	Specificity	PPV	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)
89·00%	54·8% (53·0 to 56·5)	83·3% (83·1 to 83·4)	3·81% (3·64 to 3·99)	0·50% (0·48 to 0·52)	1·95 (1·92 to 1·97)
90·00%	52·7% (51·0 to 54·4)	84·6% (84·4 to 84·7)	3·97% (3·79 to 4·16)	0·52% (0·51 to 0·54)	2·02 (2·00 to 2·05)
91·00%	50·4% (48·7 to 52·2)	85·9% (85·8 to 86·1)	4·16% (3·97 to 4·36)	0·55% (0·53 to 0·57)	2·12 (2·10 to 2·15)
92·00%	48·1% (46·4 to 49·8)	87·4% (87·3 to 87·5)	4·42% (4·20 to 4·64)	0·58% (0·56 to 0·61)	2·25 (2·22 to 2·28)
94·00%	42·7% (41·0 to 44·4)	90·4% (90·3 to 90·5)	5·11% (4·85 to 5·38)	0·68% (0·65 to 0·71)	2·61 (2·57 to 2·64)
96·00%	35·6% (33·9 to 37·2)	93·3% (93·2 to 93·4)	6·07% (5·74 to 6·42)	0·82% (0·78 to 0·86)	3·13 (3·09 to 3·17)
96·20%	35·1% (33·5 to 36·8)	93·7% (93·6 to 93·7)	6·27% (5·93 to 6·64)	0·84% (0·80 to 0·89)	3·24 (3·20 to 3·28)
96·40%	34·4% (32·8 to 36·1)	93·9% (93·9 to 94·0)	6·44% (6·08 to 6·82)	0·87% (0·83 to 0·91)	3·34 (3·30 to 3·38)
96·60%	33·8% (32·2 to 35·4)	94·2% (94·1 to 94·3)	6·61% (6·24 to 6·99)	0·89% (0·85 to 0·94)	3·43 (3·39 to 3·47)
96·80%	33·0% (31·4 to 34·6)	94·6% (94·5 to 94·7)	6·88% (6·49 to 7·28)	0·93% (0·88 to 0·98)	3·59 (3·54 to 3·63)
97·00%	31·8% (30·2 to 33·4)	94·9% (94·8 to 95·0)	7·02% (6·61 to 7·44)	0·95% (0·90 to 1·00)	3·68 (3·64 to 3·73)
97·20%	30·9% (29·3 to 32·5)	95·2% (95·2 to 95·3)	7·29% (6·86 to 7·73)	0·99% (0·94 to 1·04)	3·84 (3·79 to 3·88)
97·40%	30·1% (28·6 to 31·7)	95·6% (95·5 to 95·6)	7·59% (7·14 to 8·05)	1·03% (0·98 to 1·09)	4·03 (3·97 to 4·08)
97·60%	29·3% (27·7 to 30·9)	95·9% (95·8 to 95·9)	7·89% (7·41 to 8·38)	1·08% (1·02 to 1·14)	4·20 (4·15 to 4·26)
97·80%	28·0% (26·4 to 29·5)	96·2% (96·1 to 96·3)	8·20% (7·70 to 8·72)	1·12% (1·06 to 1·19)	4·39 (4·33 to 4·44)
98·00%	26·8% (25·3 to 28·3)	96·5% (96·5 to 96·6)	8·56% (8·03 to 9·12)	1·18% (1·11 to 1·25)	4·61 (4·55 to 4·67)
98·50%	23·6% (22·2 to 25·1)	97·3% (97·3 to 97·4)	9·64% (9·00 to 10·3)	1·34% (1·25 to 1·43)	5·35 (5·28 to 5·42)
99·00%	19·6% (18·2 to 21·0)	98·2% (98·2 to 98·3)	11·7% (10·9 to 12·6)	1·66% (1·54 to 1·78)	6·84 (6·75 to 6·92)
99·50%	14·1% (12·9 to 15·3)	99·1% (99·0 to 99·1)	15·7% (14·4 to 17·1)	2·32% (2·11 to 2·54)	10·2 (10·1 to 10·3)
99·70%	10·2% (9·20 to 11·3)	99·5% (99·4 to 99·5)	18·7% (17·0 to 20·6)	2·85% (2·54 to 3·18)	13·0 (12·9 to 13·2)
99·90%	4·76% (4·06 to 5·54)	99·8% (99·8 to 99·8)	24·5% (21·2 to 28·0)	3·96% (3·32 to 4·70)	19·0 (18·7 to 19·2)
99·95%	3·48% (2·88 to 4·16)	99·9% (99·9 to 99·9)	27·5% (23·3 to 32·0)	4·59% (3·72 to 5·63)	22·8 (22·4 to 23·1)

See notes under Table A2.

Table A4: Performance breakdown by location and race of simulated deployment of PrismNN

Location Race	TP	FP	Cancer	Control	Sensitivity	Specificity	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)	AUC
All	2,550	8,352	7,095	178,837	35·9% (34·8 to 37·1)	95·3% (95·2 to 95·4)	0·58% (0·56 to 0·60)	5·10 (5·02 to 5·18)	0·793 (0·788 to 0·799)
Midwest	740	2,573	1,300	22,648	56·9% (54·2 to 59·6)	88·6% (88·2 to 89·0)	0·55% (0·51 to 0·58)	5·21 (5·12 to 5·29)	0·828 (0·816 to 0·840)
AIAN	3	8	6	70	50·0% (11·8 to 88·2)	88·6% (78·7 to 94·9)	0·71% (0·14 to 1·92)	9·98 (4·92 to 14·0)	0·831 (0·655 to 1·000)
Asian	8	40	12	357	66·7% (34·9 to 90·1)	88·8% (85·1 to 91·9)	0·38% (0·19 to 0·60)	6·27 (5·59 to 6·90)	0·925 (0·877 to 0·972)
Black	71	242	155	3,015	45·8% (37·8 to 54·0)	92·0% (90·9 to 92·9)	0·56% (0·44 to 0·69)	5·33 (5·04 to 5·59)	0·800 (0·762 to 0·839)
NHPI	0	2	0	22	N/A	90·9% (70·8 to 98·9)	N/A	0·00 (0·00 to 0·00)	N/A
White	649	2,192	1,097	18,078	59·2% (56·2 to 62·1)	87·9% (87·4 to 88·3)	0·56% (0·53 to 0·60)	5·28 (5·17 to 5·37)	0·833 (0·820 to 0·846)
Unknown	9	89	30	1,106	30·0% (14·7 to 49·4)	92·0% (90·2 to 93·5)	0·19% (0·09 to 0·33)	2·21 (2·13 to 2·27)	0·719 (0·624 to 0·814)
Northeast	780	2,868	2,552	54,416	30·6% (28·8 to 32·4)	94·7% (94·5 to 94·9)	0·52% (0·48 to 0·55)	4·22 (4·15 to 4·28)	0·765 (0·755 to 0·774)
AIAN	3	16	3	124	100·0% (29·2 to 100·0)	87·1% (79·9 to 92·4)	0·36% (0·10 to 0·58)	4·52 (2·59 to 6·20)	0·938 (0·879 to 0·997)
Asian	11	52	33	1,281	33·3% (18·0 to 51·8)	95·9% (94·7 to 97·0)	0·40% (0·20 to 0·68)	4·78 (4·33 to 5·18)	0·803 (0·727 to 0·879)
Black	76	318	285	6,892	26·7% (21·6 to 32·2)	95·4% (94·9 to 95·9)	0·45% (0·36 to 0·57)	3·61 (3·41 to 3·79)	0·770 (0·743 to 0·797)
NHPI	0	1	1	37	0·00% (0·00 to 97·5)	97·3% (85·8 to 99·9)	0·00% (0·00 to 4·55)	0·00 (0·00 to 0·00)	0·946 (0·872 to 1·000)
White	625	2,132	1,994	40,047	31·3% (29·3 to 33·4)	94·7% (94·5 to 94·9)	0·56% (0·51 to 0·60)	4·53 (4·44 to 4·61)	0·764 (0·753 to 0·775)
Unknown	65	349	236	6,035	27·5% (21·9 to 33·7)	94·2% (93·6 to 94·8)	0·35% (0·28 to 0·45)	2·85 (2·79 to 2·91)	0·755 (0·726 to 0·784)
South	677	2,272	2,565	82,186	26·4% (24·7 to 28·1)	97·2% (97·1 to 97·3)	0·56% (0·52 to 0·61)	5·00 (4·91 to 5·07)	0·770 (0·761 to 0·780)
AIAN	4	3	11	254	36·4% (10·9 to 69·2)	98·8% (96·6 to 99·8)	2·48% (0·49 to 12·2)	36·0 (22·5 to 45·2)	0·793 (0·633 to 0·953)
Asian	12	49	50	2,152	24·0% (13·1 to 38·2)	97·7% (97·0 to 98·3)	0·46% (0·24 to 0·81)	5·65 (5·11 to 6·12)	0·805 (0·751 to 0·859)
Black	145	418	523	16,402	27·7% (23·9 to 31·8)	97·5% (97·2 to 97·7)	0·66% (0·55 to 0·78)	6·14 (5·84 to 6·39)	0·786 (0·766 to 0·806)
NHPI	0	5	1	125	0·00% (0·00 to 97·5)	96·0% (90·9 to 98·7)	0·00% (0·00 to 0·48)	0·00 (0·00 to 0·00)	0·832 (0·766 to 0·898)
White	499	1,653	1,860	56,696	26·8% (24·8 to 28·9)	97·1% (96·9 to 97·2)	0·57% (0·52 to 0·63)	4·90 (4·81 to 4·99)	0·766 (0·755 to 0·777)
Unknown	17	144	120	6,557	14·2% (8·47 to 21·7)	97·8% (97·4 to 98·1)	0·22% (0·13 to 0·36)	2·24 (2·19 to 2·28)	0·738 (0·695 to 0·781)
West	331	424	590	15,424	56·1% (52·0 to 60·2)	97·3% (97·0 to 97·5)	1·47% (1·30 to 1·65)	13·7 (13·4 to 13·9)	0·893 (0·878 to 0·908)
AIAN	1	3	5	194	20·0% (0·51 to 71·6)	98·5% (95·5 to 99·7)	0·63% (0·02 to 4·81)	5·61 (2·56 to 9·73)	0·875 (0·768 to 0·983)
Asian	7	24	18	322	38·9% (17·3 to 64·3)	92·5% (89·1 to 95·2)	0·55% (0·23 to 1·08)	5·42 (4·84 to 5·99)	0·731 (0·573 to 0·890)
Black	15	22	25	771	60·0% (38·7 to 78·9)	97·1% (95·7 to 98·2)	1·28% (0·71 to 2·19)	9·60 (8·82 to 10·4)	0·907 (0·852 to 0·961)
NHPI	0	0	0	23	N/A	100·0% (85·2 to 100·0)	N/A	N/A	N/A
White	259	254	419	11,128	61·8% (57·0 to 66·5)	97·7% (97·4 to 98·0)	1·91% (1·65 to 2·20)	17·4 (17·0 to 17·8)	0·913 (0·897 to 0·930)
Unknown	49	121	123	2,986	39·8% (31·1 to 49·1)	95·9% (95·2 to 96·6)	0·77% (0·57 to 1·01)	7·85 (7·66 to 8·04)	0·827 (0·788 to 0·866)
Unknown	22	215	88	4,163	25·0% (16·4 to 35·4)	94·8% (94·1 to 95·5)	0·19% (0·13 to 0·28)	1·69 (1·65 to 1·73)	0·750 (0·698 to 0·802)
AIAN	0	0	0	0	N/A	N/A	N/A	N/A	N/A
Asian	1	1	2	151	50·0% (1·26 to 98·7)	99·3% (96·4 to 100·0)	1·87% (0·02 to 45·8)	28·8 (20·7 to 39·6)	0·930 (0·791 to 1·000)
Black	0	15	6	362	0·00% (0·00 to 45·9)	95·9% (93·3 to 97·7)	0·00% (0·00 to 0·39)	0·00 (0·00 to 0·00)	0·824 (0·729 to 0·918)
NHPI	0	0	0	21	N/A	100·0% (83·9 to 100·0)	N/A	N/A	N/A
White	18	189	68	3,165	26·5% (16·5 to 38·6)	94·0% (93·1 to 94·8)	0·18% (0·11 to 0·27)	1·57 (1·52 to 1·61)	0·746 (0·686 to 0·807)
Unknown	3	10	12	464	25·0% (5·49 to 57·2)	97·8% (96·1 to 99·0)	0·57% (0·11 to 1·71)	6·16 (5·89 to 6·44)	0·699 (0·528 to 0·869)

Notes
- Numbers in brackets indicate 95% CI.
- All evaluations use the same high-risk threshold chosen at $\mathrm{SIR} \approx 5$ on the whole test population.
- TP: total number of true positive predictions in a subpopulation.
- FP: total number of false positive predictions in a subpopulation.
- Cancer: total number of patients with PDAC in a subpopulation.
- Control: total number of patients without PDAC in a subpopulation.
- PPV: Positive Predictive Value
- SIR: Standardised Incidence Ratio
- TrxPop. Est.: Since we used all the PDAC cases in the TriNetX database but sampled a subset of control patients, we needed to account for this imbalance to estimate the PPV and SIR that would be obtained if we had evaluated the model on the full TriNetX population.
- AUC: Area Under the ROC Curve, calculated by varying the threshold for high-risk patients.

Table A5: Performance breakdown by race and sex of simulated deployment of PrismNN

Race / Sex	TP	FP	Cancer	Control	Sensitivity	Specificity	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)	AUC
All	2,550	8,352	7,095	178,837	35·9% (34·8 to 37·1)	95·3% (95·2 to 95·4)	0·58% (0·56 to 0·60)	5·10 (5·02 to 5·18)	0·793 (0·788 to 0·799)
AIAN	11	30	25	642	44·0% (24·4 to 65·1)	95·3% (93·4 to 96·8)	0·70% (0·35 to 1·21)	8·78 (5·47 to 11·2)	0·833 (0·747 to 0·920)
Female	6	13	14	359	42·9% (17·7 to 71·1)	96·4% (93·9 to 98·1)	0·87% (0·31 to 1·94)	12·2 (7·36 to 15·7)	0·831 (0·732 to 0·929)
Male	5	17	11	273	45·5% (16·7 to 76·6)	93·8% (90·2 to 96·3)	0·56% (0·19 to 1·19)	6·56 (3·45 to 9·62)	0·835 (0·676 to 0·995)
Unknown	0	0	0	10	N/A	100·0% (69·2 to 100·0)	N/A	N/A	N/A
Asian	39	166	115	4,263	33·9% (25·3 to 43·3)	96·1% (95·5 to 96·7)	0·45% (0·32 to 0·60)	5·55 (5·12 to 5·92)	0·810 (0·769 to 0·850)
Female	19	76	65	2,426	29·2% (18·6 to 41·8)	96·9% (96·1 to 97·5)	0·47% (0·29 to 0·73)	7·35 (6·61 to 8·00)	0·828 (0·777 to 0·880)
Male	20	90	50	1,779	40·0% (26·4 to 54·8)	94·9% (93·8 to 95·9)	0·42% (0·27 to 0·62)	4·51 (4·04 to 4·93)	0·780 (0·712 to 0·848)
Unknown	0	0	0	58	N/A	100·0% (93·8 to 100·0)	N/A	N/A	N/A
Black	307	1,015	994	27,442	30·9% (28·0 to 33·9)	96·3% (96·1 to 96·5)	0·57% (0·51 to 0·64)	5·07 (4·82 to 5·29)	0·790 (0·776 to 0·805)
Female	184	531	587	16,769	31·3% (27·6 to 35·3)	96·8% (96·6 to 97·1)	0·66% (0·56 to 0·76)	6·26 (5·86 to 6·62)	0·806 (0·788 to 0·825)
Male	123	483	407	10,639	30·2% (25·8 to 34·9)	95·5% (95·0 to 95·8)	0·48% (0·40 to 0·57)	3·96 (3·67 to 4·21)	0·764 (0·741 to 0·787)
Unknown	0	1	0	34	N/A	97·1% (84·7 to 99·9)	N/A	0·00 (0·00 to 0·00)	N/A
NHPI	0	8	2	228	0·00% (0·00 to 84·2)	96·5% (93·2 to 98·5)	0·00% (0·00 to 0·48)	0·00 (0·00 to 0·00)	0·851 (0·708 to 0·994)
Female	0	5	1	132	0·00% (0·00 to 97·5)	96·2% (91·4 to 98·8)	0·00% (0·00 to 0·47)	0·00 (0·00 to 0·00)	0·917 (0·869 to 0·964)
Male	0	3	1	94	0·00% (0·00 to 97·5)	96·8% (91·0 to 99·3)	0·00% (0·00 to 0·95)	0·00 (0·00 to 0·00)	0·723 (0·632 to 0·814)
Unknown	0	0	0	2	N/A	100·0% (15·8 to 100·0)	N/A	N/A	N/A
White	2,050	6,420	5,438	129,114	37·7% (36·4 to 39·0)	95·0% (94·9 to 95·1)	0·61% (0·58 to 0·63)	5·27 (5·17 to 5·36)	0·795 (0·788 to 0·801)
Female	985	3,109	2,726	70,821	36·1% (34·3 to 38·0)	95·6% (95·5 to 95·8)	0·60% (0·56 to 0·64)	6·22 (6·06 to 6·37)	0·790 (0·781 to 0·799)
Male	1,065	3,264	2,712	54,627	39·3% (37·4 to 41·1)	94·0% (93·8 to 94·2)	0·62% (0·58 to 0·66)	4·67 (4·55 to 4·79)	0·788 (0·779 to 0·798)
Unknown	0	47	0	3,666	N/A	98·7% (98·3 to 99·1)	N/A	0·00 (0·00 to 0·00)	N/A
Unknown	143	713	521	17,148	27·4% (23·7 to 31·5)	95·8% (95·5 to 96·1)	0·38% (0·32 to 0·45)	3·47 (3·41 to 3·53)	0·773 (0·753 to 0·792)
Female	65	345	259	9,745	25·1% (19·9 to 30·8)	96·5% (96·1 to 96·8)	0·36% (0·28 to 0·45)	3·83 (3·74 to 3·91)	0·767 (0·738 to 0·796)
Male	78	368	262	7,299	29·8% (24·3 to 35·7)	95·0% (94·4 to 95·4)	0·40% (0·32 to 0·50)	3·22 (3·14 to 3·30)	0·772 (0·744 to 0·800)
Unknown	0	0	0	104	N/A	100·0% (96·5 to 100·0)	N/A	N/A	N/A

See notes under Table A4

Table A6: Performance breakdown by age and sex of simulated deployment of PrismNN

Age / Sex	TP	FP	Cancer	Control	Sensitivity	Specificity	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)	AUC
All	2,550	8,352	7,095	178,837	35·9% (34·8 to 37·1)	95·3% (95·2 to 95·4)	0·58% (0·56 to 0·60)	5·10 (5·02 to 5·18)	0·793 (0·788 to 0·799)
40 - 50	96	448	407	38,305	23·6% (19·5 to 28·0)	98·8% (98·7 to 98·9)	0·41% (0·33 to 0·50)	40·7 (37·6 to 43·1)	0·774 (0·748 to 0·799)
Female	62	288	256	23,229	24·2% (19·1 to 29·9)	98·8% (98·6 to 98·9)	0·41% (0·31 to 0·52)	49·3 (43·8 to 53·7)	0·763 (0·729 to 0·797)
Male	34	156	151	14,349	22·5% (16·1 to 30·0)	98·9% (98·7 to 99·1)	0·41% (0·29 to 0·58)	32·0 (28·6 to 34·5)	0·786 (0·746 to 0·825)
Unknown	0	4	0	727	N/A	99·4% (98·6 to 99·8)	N/A	0·00 (0·00 to 0·00)	N/A
50 - 60	329	1,027	1,097	45,528	30·0% (27·3 to 32·8)	97·7% (97·6 to 97·9)	0·61% (0·54 to 0·68)	16·0 (15·4 to 16·5)	0·767 (0·751 to 0·783)
Female	165	549	584	25,502	28·3% (24·6 to 32·1)	97·8% (97·7 to 98·0)	0·57% (0·49 to 0·66)	18·4 (17·4 to 19·3)	0·768 (0·746 to 0·790)
Male	164	477	513	19,034	32·0% (27·9 to 36·2)	97·5% (97·3 to 97·7)	0·65% (0·55 to 0·76)	14·2 (13·5 to 14·8)	0·757 (0·733 to 0·781)
Unknown	0	1	0	992	N/A	99·9% (99·4 to 100·0)	N/A	0·00 (0·00 to 0·00)	N/A
60 - 70	788	2,382	2,302	49,374	34·2% (32·3 to 36·2)	95·2% (95·0 to 95·4)	0·63% (0·58 to 0·67)	7·29 (7·10 to 7·46)	0·747 (0·736 to 0·758)
Female	379	1,146	1,136	26,467	33·4% (30·6 to 36·2)	95·7% (95·4 to 95·9)	0·63% (0·57 to 0·69)	8·53 (8·22 to 8·83)	0·755 (0·739 to 0·771)
Male	409	1,219	1,166	21,783	35·1% (32·3 to 37·9)	94·4% (94·1 to 94·7)	0·64% (0·58 to 0·70)	6·50 (6·27 to 6·72)	0·725 (0·709 to 0·741)
Unknown	0	17	0	1,124	N/A	98·5% (97·6 to 99·1)	N/A	0·00 (0·00 to 0·00)	N/A
70 - 80	939	3,081	2,377	34,354	39·5% (37·5 to 41·5)	91·0% (90·7 to 91·3)	0·58% (0·54 to 0·61)	4·02 (3·92 to 4·12)	0·729 (0·717 to 0·740)
Female	446	1,412	1,185	18,667	37·6% (34·9 to 40·5)	92·4% (92·0 to 92·8)	0·60% (0·55 to 0·66)	4·87 (4·69 to 5·04)	0·720 (0·703 to 0·737)
Male	493	1,651	1,192	14,875	41·4% (38·5 to 44·2)	88·9% (88·4 to 89·4)	0·57% (0·52 to 0·62)	3·51 (3·38 to 3·63)	0·724 (0·708 to 0·740)
Unknown	0	18	0	812	N/A	97·8% (96·5 to 98·7)	N/A	0·00 (0·00 to 0·00)	N/A
> 80	398	1,414	912	11,276	43·6% (40·4 to 46·9)	87·5% (86·8 to 88·1)	0·53% (0·49 to 0·58)	2·94 (2·84 to 3·05)	0·737 (0·719 to 0·755)
Female	207	684	491	6,387	42·2% (37·7 to 46·7)	89·3% (88·5 to 90·0)	0·57% (0·50 to 0·65)	3·36 (3·20 to 3·51)	0·735 (0·710 to 0·759)
Male	191	722	421	4,670	45·4% (40·5 to 50·3)	84·5% (83·5 to 85·6)	0·50% (0·44 to 0·57)	2·63 (2·48 to 2·77)	0·732 (0·705 to 0·758)
Unknown	0	8	0	219	N/A	96·3% (92·9 to 98·4)	N/A	0·00 (0·00 to 0·00)	N/A

See notes under Table A4

Table A7: Performance breakdown by location and race of simulated deployment of PrismLR

Location Race	TP	FP	Cancer	Control	Sensitivity	Specificity	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)	AUC
All	1,871	5,392	7,095	178,837	26·4% (25·3 to 27·4)	97·0% (96·9 to 97·1)	0·66% (0·63 to 0·69)	5·17 (5·08 to 5·25)	0·787 (0·781 to 0·792)
Midwest	495	1,440	1,300	22,648	38·1% (35·4 to 40·8)	93·6% (93·3 to 94·0)	0·65% (0·60 to 0·71)	5·37 (5·27 to 5·47)	0·802 (0·789 to 0·814)
AIAN	3	7	6	70	50·0% (11·8 to 88·2)	90·0% (80·5 to 95·9)	0·81% (0·16 to 2·32)	10·3 (4·70 to 15·0)	0·783 (0·591 to 0·976)
Asian	7	23	12	357	58·3% (27·7 to 84·8)	93·6% (90·5 to 95·9)	0·58% (0·25 to 1·04)	6·71 (5·93 to 7·43)	0·878 (0·796 to 0·961)
Black	49	126	155	3,015	31·6% (24·4 to 39·6)	95·8% (95·0 to 96·5)	0·74% (0·54 to 0·98)	6·44 (6·06 to 6·77)	0·803 (0·767 to 0·839)
NHPI	0	1	0	22	N/A	95·5% (77·2 to 99·9)	N/A	0·00 (0·00 to 0·00)	N/A
White	428	1,251	1,097	18,078	39·0% (36·1 to 42·0)	93·1% (92·7 to 93·4)	0·65% (0·59 to 0·71)	5·25 (5·14 to 5·35)	0·802 (0·789 to 0·816)
Unknown	8	32	30	1,106	26·7% (12·3 to 45·9)	97·1% (95·9 to 98·0)	0·47% (0·20 to 0·92)	5·08 (4·85 to 5·29)	0·680 (0·568 to 0·792)
Northeast	625	2,059	2,552	54,416	24·5% (22·8 to 26·2)	96·2% (96·1 to 96·4)	0·58% (0·53 to 0·62)	4·35 (4·27 to 4·42)	0·775 (0·766 to 0·784)
AIAN	1	11	3	124	33·3% (0·84 to 90·6)	91·1% (84·7 to 95·5)	0·17% (0·01 to 0·59)	2·09 (1·03 to 3·13)	0·774 (0·524 to 1·000)
Asian	9	29	33	1,281	27·3% (13·3 to 45·5)	97·7% (96·8 to 98·5)	0·59% (0·26 to 1·12)	6·52 (5·86 to 7·13)	0·785 (0·700 to 0·870)
Black	65	279	285	6,892	22·8% (18·1 to 28·1)	96·0% (95·5 to 96·4)	0·44% (0·34 to 0·56)	3·35 (3·17 to 3·52)	0·767 (0·739 to 0·795)
NHPI	0	1	1	37	0·00% (0·00 to 97·5)	97·3% (85·8 to 99·9)	0·00% (0·00 to 4·55)	0·00 (0·00 to 0·00)	0·946 (0·872 to 1·000)
White	505	1,495	1,994	40,047	25·3% (23·4 to 27·3)	96·3% (96·1 to 96·5)	0·64% (0·58 to 0·70)	4·78 (4·69 to 4·88)	0·778 (0·768 to 0·788)
Unknown	45	244	236	6,035	19·1% (14·3 to 24·7)	96·0% (95·4 to 96·4)	0·35% (0·26 to 0·47)	2·68 (2·62 to 2·73)	0·754 (0·726 to 0·783)
South	462	1,526	2,565	82,186	18·0% (16·5 to 19·6)	98·1% (98·0 to 98·2)	0·57% (0·52 to 0·63)	4·54 (4·46 to 4·61)	0·759 (0·749 to 0·768)
AIAN	2	3	11	254	18·2% (2·28 to 51·8)	98·8% (96·6 to 99·8)	1·26% (0·12 to 8·00)	23·1 (9·86 to 38·9)	0·785 (0·639 to 0·930)
Asian	7	23	50	2,152	14·0% (5·82 to 26·7)	98·9% (98·4 to 99·3)	0·58% (0·22 to 1·27)	6·91 (6·11 to 7·63)	0·770 (0·708 to 0·833)
Black	86	229	523	16,402	16·4% (13·4 to 19·9)	98·6% (98·4 to 98·8)	0·71% (0·56 to 0·90)	5·77 (5·45 to 6·04)	0·774 (0·753 to 0·795)
NHPI	0	3	1	125	0·00% (0·00 to 97·5)	97·6% (93·1 to 99·5)	0·00% (0·00 to 0·95)	0·00 (0·00 to 0·00)	0·968 (0·937 to 0·999)
White	356	1,155	1,860	56,696	19·1% (17·4 to 21·0)	98·0% (97·8 to 98·1)	0·58% (0·52 to 0·65)	4·52 (4·43 to 4·60)	0·755 (0·744 to 0·766)
Unknown	11	113	120	6,557	9·17% (4·67 to 15·8)	98·3% (97·9 to 98·6)	0·19% (0·09 to 0·33)	1·62 (1·58 to 1·65)	0·736 (0·690 to 0·782)
West	283	334	590	15,424	48·0% (43·9 to 52·1)	97·8% (97·6 to 98·1)	1·59% (1·39 to 1·82)	13·3 (12·9 to 13·5)	0·869 (0·851 to 0·886)
AIAN	1	6	5	194	20·0% (0·51 to 71·6)	96·9% (93·4 to 98·9)	0·32% (0·01 to 1·63)	2·86 (1·09 to 4·49)	0·845 (0·724 to 0·967)
Asian	6	18	18	322	33·3% (13·3 to 59·0)	94·4% (91·3 to 96·7)	0·63% (0·23 to 1·37)	7·33 (6·32 to 8·31)	0·744 (0·600 to 0·888)
Black	11	11	25	771	44·0% (24·4 to 65·1)	98·6% (97·5 to 99·3)	1·87% (0·82 to 4·07)	12·0 (10·7 to 13·3)	0·840 (0·755 to 0·925)
NHPI	0	0	0	23	N/A	100·0% (85·2 to 100·0)	N/A	N/A	N/A
White	220	167	419	11,128	52·5% (47·6 to 57·4)	98·5% (98·3 to 98·7)	2·45% (2·05 to 2·92)	20·6 (20·1 to 21·0)	0·885 (0·865 to 0·905)
Unknown	45	132	123	2,986	36·6% (28·1 to 45·7)	95·6% (94·8 to 96·3)	0·65% (0·47 to 0·86)	5·27 (5·14 to 5·39)	0·817 (0·778 to 0·855)
Unknown	6	33	88	4,163	6·82% (2·54 to 14·3)	99·2% (98·9 to 99·5)	0·35% (0·12 to 0·79)	2·01 (1·94 to 2·09)	0·704 (0·646 to 0·762)
AIAN	0	0	0	0	N/A	N/A	N/A	N/A	N/A
Asian	1	0	2	151	50·0% (1·26 to 98·7)	100·0% (97·6 to 100·0)	100·0% (0·04 to 100·0)	1981·0 (1444·8 to 2742·5)	0·944 (0·831 to 1·000)
Black	0	2	6	362	0·00% (0·00 to 45·9)	99·4% (98·0 to 99·9)	0·00% (0·00 to 4·86)	0·00 (0·00 to 0·00)	0·674 (0·486 to 0·863)
NHPI	0	0	0	21	N/A	100·0% (83·9 to 100·0)	N/A	N/A	N/A
White	4	28	68	3,165	5·88% (1·63 to 14·4)	99·1% (98·7 to 99·4)	0·27% (0·07 to 0·73)	1·49 (1·43 to 1·55)	0·713 (0·646 to 0·779)
Unknown	1	3	12	464	8·33% (0·21 to 38·5)	99·4% (98·1 to 99·9)	0·63% (0·01 to 5·69)	5·13 (4·87 to 5·40)	0·649 (0·497 to 0·802)

See notes under Table A4

Table A8: Performance breakdown by race and sex of simulated deployment of PrismLR

Race / Sex	TP	FP	Cancer	Control	Sensitivity	Specificity	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)	AUC
All	1,871	5,392	7,095	178,837	26·4% (25·3 to 27·4)	97·0% (96·9 to 97·1)	0·66% (0·63 to 0·69)	5·17 (5·08 to 5·25)	0·787 (0·781 to 0·792)
AIAN	7	27	25	642	28·0% (12·1 to 49·4)	95·8% (93·9 to 97·2)	0·49% (0·20 to 1·00)	5·80 (3·23 to 7·43)	0·789 (0·702 to 0·875)
Female	3	10	14	359	21·4% (4·66 to 50·8)	97·2% (94·9 to 98·7)	0·57% (0·11 to 1·77)	7·41 (3·94 to 9·81)	0·788 (0·678 to 0·897)
Male	4	17	11	273	36·4% (10·9 to 69·2)	93·8% (90·2 to 96·3)	0·45% (0·12 to 1·05)	4·99 (2·36 to 7·17)	0·789 (0·647 to 0·931)
Unknown	0	0	0	10	N/A	100·0% (69·2 to 100·0)	N/A	N/A	N/A
Asian	30	93	115	4,263	26·1% (18·3 to 35·1)	97·8% (97·3 to 98·2)	0·61% (0·41 to 0·88)	7·05 (6·44 to 7·57)	0·786 (0·743 to 0·830)
Female	21	44	65	2,426	32·3% (21·2 to 45·1)	98·2% (97·6 to 98·7)	0·90% (0·55 to 1·42)	12·0 (10·6 to 13·3)	0·801 (0·741 to 0·861)
Male	9	49	50	1,779	18·0% (8·58 to 31·4)	97·2% (96·4 to 98·0)	0·35% (0·16 to 0·66)	3·59 (3·17 to 3·97)	0·757 (0·690 to 0·823)
Unknown	0	0	0	58	N/A	100·0% (93·8 to 100·0)	N/A	N/A	N/A
Black	211	647	994	27,442	21·2% (18·7 to 23·9)	97·6% (97·5 to 97·8)	0·62% (0·53 to 0·71)	4·92 (4·65 to 5·14)	0·780 (0·766 to 0·795)
Female	117	348	587	16,769	19·9% (16·8 to 23·4)	97·9% (97·7 to 98·1)	0·64% (0·52 to 0·77)	5·37 (5·01 to 5·69)	0·791 (0·772 to 0·810)
Male	94	299	407	10,639	23·1% (19·1 to 27·5)	97·2% (96·9 to 97·5)	0·60% (0·48 to 0·73)	4·45 (4·10 to 4·76)	0·763 (0·739 to 0·787)
Unknown	0	0	0	34	N/A	100·0% (89·7 to 100·0)	N/A	N/A	N/A
NHPI	0	5	2	228	0·00% (0·00 to 84·2)	97·8% (95·0 to 99·3)	0·00% (0·00 to 0·85)	0·00 (0·00 to 0·00)	0·956 (0·926 to 0·987)
Female	0	3	1	132	0·00% (0·00 to 97·5)	97·7% (93·5 to 99·5)	0·00% (0·00 to 0·94)	0·00 (0·00 to 0·00)	0·947 (0·909 to 0·985)
Male	0	2	1	94	0·00% (0·00 to 97·5)	97·9% (92·5 to 99·7)	0·00% (0·00 to 1·72)	0·00 (0·00 to 0·00)	0·968 (0·932 to 1·000)
Unknown	0	0	0	2	N/A	100·0% (15·8 to 100·0)	N/A	N/A	N/A
White	1,513	4,096	5,438	129,114	27·8% (26·6 to 29·0)	96·8% (96·7 to 96·9)	0·70% (0·66 to 0·74)	5·42 (5·31 to 5·52)	0·789 (0·783 to 0·795)
Female	705	2,038	2,726	70,821	25·9% (24·2 to 27·5)	97·1% (97·0 to 97·2)	0·66% (0·61 to 0·71)	5·75 (5·59 to 5·90)	0·784 (0·775 to 0·793)
Male	808	2,055	2,712	54,627	29·8% (28·1 to 31·6)	96·2% (96·1 to 96·4)	0·74% (0·69 to 0·80)	5·17 (5·03 to 5·30)	0·778 (0·769 to 0·787)
Unknown	0	3	0	3,666	N/A	99·9% (99·8 to 100·0)	N/A	0·00 (0·00 to 0·00)	N/A
Unknown	110	524	521	17,148	21·1% (17·7 to 24·9)	96·9% (96·7 to 97·2)	0·40% (0·33 to 0·48)	3·24 (3·18 to 3·30)	0·766 (0·746 to 0·786)
Female	49	280	259	9,745	18·9% (14·3 to 24·2)	97·1% (96·8 to 97·4)	0·33% (0·25 to 0·44)	2·96 (2·89 to 3·03)	0·756 (0·726 to 0·785)
Male	61	244	262	7,299	23·3% (18·3 to 28·9)	96·7% (96·2 to 97·1)	0·47% (0·36 to 0·61)	3·51 (3·42 to 3·60)	0·770 (0·742 to 0·799)
Unknown	0	0	0	104	N/A	100·0% (96·5 to 100·0)	N/A	N/A	N/A

See notes under Table A4

Table A9: Performance breakdown by age and sex of simulated deployment of PrismLR

Age / Sex	TP	FP	Cancer	Control	Sensitivity	Specificity	PPV (TrxPop. Est.)	SIR (TrxPop. Est.)	AUC
All	1,871	5,392	7,095	178,837	26·4% (25·3 to 27·4)	97·0% (96·9 to 97·1)	0·66% (0·63 to 0·69)	5·17 (5·09 to 5·25)	0·787 (0·781 to 0·792)
40 - 50	41	157	407	38,305	10·1% (7·33 to 13·4)	99·6% (99·5 to 99·7)	0·50% (0·35 to 0·69)	43·8 (39·9 to 46·4)	0·767 (0·740 to 0·794)
Female	25	90	256	23,229	9·77% (6·42 to 14·1)	99·6% (99·5 to 99·7)	0·53% (0·33 to 0·80)	60·5 (52·5 to 66·0)	0·760 (0·725 to 0·795)
Male	16	67	151	14,349	10·6% (6·18 to 16·6)	99·5% (99·4 to 99·6)	0·45% (0·25 to 0·76)	30·6 (27·1 to 33·2)	0·769 (0·728 to 0·811)
Unknown	0	0	0	727	N/A	100·0% (99·5 to 100·0)	N/A	N/A	N/A
50 - 60	188	475	1,097	45,528	17·1% (15·0 to 19·5)	99·0% (98·9 to 99·0)	0·75% (0·64 to 0·88)	19·1 (18·4 to 19·7)	0·755 (0·739 to 0·771)
Female	87	245	584	25,502	14·9% (12·1 to 18·0)	99·0% (98·9 to 99·2)	0·67% (0·53 to 0·85)	21·0 (19·8 to 22·0)	0·753 (0·731 to 0·775)
Male	101	230	513	19,034	19·7% (16·3 to 23·4)	98·8% (98·6 to 98·9)	0·83% (0·66 to 1·03)	17·7 (16·8 to 18·5)	0·743 (0·719 to 0·768)
Unknown	0	0	0	992	N/A	100·0% (99·6 to 100·0)	N/A	N/A	N/A
60 - 70	536	1,275	2,302	49,374	23·3% (21·6 to 25·1)	97·4% (97·3 to 97·6)	0·80% (0·72 to 0·87)	9·01 (8·78 to 9·22)	0·740 (0·728 to 0·751)
Female	256	639	1,136	26,467	22·5% (20·1 to 25·1)	97·6% (97·4 to 97·8)	0·76% (0·66 to 0·86)	9·88 (9·51 to 10·2)	0·744 (0·728 to 0·760)
Male	280	636	1,166	21,783	24·0% (21·6 to 26·6)	97·1% (96·8 to 97·3)	0·83% (0·73 to 0·95)	8·33 (8·04 to 8·61)	0·719 (0·702 to 0·736)
Unknown	0	0	0	1,124	N/A	100·0% (99·7 to 100·0)	N/A	N/A	N/A
70 - 80	731	2,190	2,377	34,354	30·8% (28·9 to 32·7)	93·6% (93·4 to 93·9)	0·63% (0·59 to 0·68)	4·34 (4·23 to 4·45)	0·718 (0·707 to 0·730)
Female	344	1,084	1,185	18,667	29·0% (26·5 to 31·7)	94·2% (93·8 to 94·5)	0·60% (0·54 to 0·67)	4·79 (4·61 to 4·95)	0·708 (0·691 to 0·724)
Male	387	1,104	1,192	14,875	32·5% (29·8 to 35·2)	92·6% (92·1 to 93·0)	0·66% (0·60 to 0·74)	4·02 (3·87 to 4·16)	0·712 (0·695 to 0·728)
Unknown	0	2	0	812	N/A	99·8% (99·1 to 100·0)	N/A	0·00 (0·00 to 0·00)	N/A
> 80	375	1,295	912	11,276	41·1% (37·9 to 44·4)	88·5% (87·9 to 89·1)	0·55% (0·50 to 0·60)	3·04 (2·93 to 3·15)	0·733 (0·715 to 0·751)
Female	183	665	491	6,387	37·3% (33·0 to 41·7)	89·6% (88·8 to 90·3)	0·52% (0·45 to 0·60)	3·01 (2·87 to 3·14)	0·720 (0·695 to 0·745)
Male	192	629	421	4,670	45·6% (40·8 to 50·5)	86·5% (85·5 to 87·5)	0·58% (0·51 to 0·66)	3·08 (2·92 to 3·25)	0·736 (0·710 to 0·762)
Unknown	0	1	0	219	N/A	99·5% (97·5 to 100·0)	N/A	0·00 (0·00 to 0·00)	N/A

See notes under Table A4

Table A10: Model test AUC (95% CI) with training data from different locations

Model	Training set	Test set (PDAC proportion, control proportion) Midwest (21·4%, 15·5%)	Northeast (33·5%, 28·2%)	South (36·2%, 44·1%)	West (7·4%, 8·6%)	All (98·4%, 96·4%)
PrismNN	Midwest	0·818 (±0·012)	0·698 (±0·012)	0·692 (±0·012)	0·736 (±0·027)	0·725 (±0·007)
	Northeast	0·712 (±0·015)	0·821 (±0·009)	0·743 (±0·011)	0·753 (±0·025)	0·768 (±0·006)
	South	0·722 (±0·014)	0·746 (±0·011)	0·812 (±0·010)	0·832 (±0·021)	0·778 (±0·006)
	West	0·600 (±0·016)	0·622 (±0·013)	0·650 (±0·013)	0·884 (±0·020)	0·649 (±0·008)
	Test + 1	0·828 (±0·012)	**0·823** (±0·009)	0·812 (±0·010)	**0·925** (±0·014)	best of 2: M,S
	Test + 1	0·827 (±0·011)	0·820 (±0·009)	**0·818** (±0·010)	0·922 (±0·015)	0·803 (±0·006)
	Test + 1	**0·838** (±0·011)	0·822 (±0·009)	0·817 (±0·010)	0·921 (±0·015)	
	Test + 2	0·824 (±0·012)	0·812 (±0·010)	0·813 (±0·009)	0·905 (±0·018)	
	Test + 2	0·829 (±0·012)	0·820 (±0·009)	0·811 (±0·009)	0·915 (±0·016)	best of 3: M,N,S
	Test + 2	0·832 (±0·011)	0·818 (±0·010)	0·814 (±0·010)	0·913 (±0·017)	0·817 (±0·006)
	Except	0·702 (±0·015)	0·740 (±0·011)	0·723 (±0·011)	0·771 (±0·023)	
	All	0·819 (±0·012)	0·820 (±0·009)	0·812 (±0·009)	0·912 (±0·017)	**0·830** (±0·005)
PrismLR	Midwest	0·823 (±0·012)	0·711 (±0·012)	0·715 (±0·011)	0·836 (±0·020)	0·749 (±0·007)
	Northeast	0·725 (±0·015)	0·810 (±0·009)	0·761 (±0·010)	0·718 (±0·027)	0·771 (±0·006)
	South	0·718 (±0·015)	0·772 (±0·010)	0·801 (±0·010)	0·844 (±0·020)	0·780 (±0·006)
	West	0·657 (±0·017)	0·667 (±0·013)	0·688 (±0·012)	0·913 (±0·015)	0·691 (±0·008)
	Test + 1	0·802 (±0·013)	0·797 (±0·010)	0·793 (±0·010)	**0·921** (±0·015)	best of 2: M,S
	Test + 1	0·800 (±0·013)	0·800 (±0·009)	0·793 (±0·010)	0·909 (±0·016)	0·796 (±0·006)
	Test + 1	**0·823** (±0·012)	**0·810** (±0·009)	**0·801** (±0·010)	0·901 (±0·018)	
	Test + 2	0·788 (±0·013)	0·798 (±0·010)	0·792 (±0·010)	0·906 (±0·017)	
	Test + 2	0·806 (±0·013)	0·797 (±0·009)	0·786 (±0·010)	0·904 (±0·017)	best of 3: M,N,S
	Test + 2	0·802 (±0·013)	0·799 (±0·010)	0·791 (±0·010)	0·893 (±0·018)	0·802 (±0·006)
	Except	0·740 (±0·015)	0·765 (±0·011)	0·749 (±0·011)	0·847 (±0·019)	
	All	0·793 (±0·013)	0·798 (±0·010)	0·785 (±0·010)	0·899 (±0·018)	**0·806** (±0·006)

Notes

- **Bold** numbers are the best performance on each test location of one model family. Underlined numbers are the best performing model on each test location.
- For each test location, `Test + 1` is the model trained on that test location plus another location. There are three ways of choosing one other location and therefore three such rows for each test location. Similarly, the `Test + 2` rows are models trained on two other locations besides the test location.
- The test set `All` contains all data with known locations but excludes patients with unknown locations. Therefore, it does not have 100% data. The `best of 2` and `best of 3` are models trained on two or three locations with the best test AUC on all locations.
- `Except` is training on other locations except the test location. Note that it is different from the external validation results reported in the body text. In the body text, we trained on three locations and tested on all data of the fourth location. Here, we first split data of each location into training and test sets, and used the same test sets in all comparisons in this table. The numbers in this `Except` row were obtained with models on smaller training sets and smaller test sets compared to the models used in the body text.

References

1. Jia K, Rinard M. Effective neural network L_0 regularization with binmask. 2023. arXiv: 2304.11237 [cs.LG].
2. Defazio A, Bach FR, Lacoste-Julien S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in neural information processing systems 27. Ed. by Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ. 2014:1646–54.
3. Loshchilov I, Hutter F. Decoupled weight decay regularization. 2019. arXiv: 1711.05101 [cs.LG].
4. Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts. In: International conference on learning representations. 2017.
5. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical programming 1989;45:503–28.
6. Platt J et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 1999;10:61–74.
7. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26:404–13.
8. Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. Deep neural networks and tabular data: a survey. IEEE Transactions on Neural Networks and Learning Systems 2022.
9. Issitt RW, Cortina-Borja M, Bryant W, Bowyer S, Taylor AM, Sebire N. Classification performance of neural networks versus logistic regression models: evidence from healthcare practice. Cureus 2022;14.