Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Early Detection of Pancreatic Ductal Adenocarcinoma through Layered Multimodal Risk Enrichment of Routine Clinical Data

Che Ngufor

ngufor.che@mayo.edu

Mayo Clinic

Shrinath Patel

Mayo Clinic

Alex D. Weston

Mayo Clinic

Alex M. Bogdanski

Leiden University Medical Center

Koen J.H. Wijsman

Mayo Clinic

Jianfu Li

Mayo Clinic in Florida

Michael B. Wallace

Mayo Clinic in Florida

Yan Bi

Mayo Clinic in Florida

Article

Keywords: pancreatic ductal adenocarcinoma, early detection, machine learning, electronic health records, risk prediction, multimodal risk enrichment

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11265358/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Pancreatic ductal adenocarcinoma (PDAC) has a five-year survival rate of approximately 8%, largely because most patients are diagnosed at an advanced stage. Effective population-level screening strategies remain unavailable, and existing approaches are costly, invasive, or impractical for broad implementation.

Objective

To develop and temporally validate a layered multimodal risk stratification framework for early PDAC detection using routinely collected clinical data.

Methods

We conducted a retrospective cohort study using longitudinal electronic health record (EHR) data from Mayo Clinic. Adults with histologically confirmed PDAC diagnosed between 2000 and 2020 were propensity-score matched to cancer-free controls using age, sex, race, year of encounter, and healthcare utilization. Machine learning models progressively incorporated demographics, vital signs and laboratory measurements, medications, and Clinical Classifications Software (CCS) diagnostic categories to predict PDAC at 3, 6, 12, and 36 months before diagnosis. Models were developed using patients diagnosed between 2000 and 2013 and evaluated in a temporally held-out validation cohort from 2014 to 2020. Two complementary analyses were conducted according to laboratory data availability: the Biomarker-Cohort Analysis included patients with at least one blood-based biomarker measurement within the observation window, while the Full-Cohort Analysis retained eligible patients regardless of laboratory availability and used extreme gradient boosting (XGBoost) to natively accommodate missing values. Additional analyses evaluated ensemble learning, model explainability using Shapley Additive Explanations (SHAP), and CT-derived body composition as an opportunistic imaging enrichment layer.

Results

At the 3-month prediction horizon, the Biomarker-Cohort Analysis included 11,965 patients (1,570 PDAC cases and 10,395 controls), while the Full-Cohort Analysis included 17,720 patients (3,332 cases and 14,388 controls). Progressive enrichment of structured EHR information improved PDAC discrimination, with the largest incremental gain occurring after incorporation of CCS diagnostic history (ΔAUC, 0.13–0.17 across horizons). The best-performing structured EHR enrichment layer, incorporating demographics, vital signs, laboratory measurements, medications, and CCS diagnostic history, achieved temporal validation AUCs of 0.84, 0.86, 0.81, and 0.79 at 3, 6, 12, and 36 months, respectively, when

modeled using XGBoost. Corresponding AUCs in the Full-Cohort Analysis were 0.83, 0.82, 0.80, and 0.77 despite substantial laboratory missingness. Ensemble learning further increased AUCs to 0.88, 0.91, 0.89, and 0.89, corresponding to absolute improvements of 0.04, 0.05, 0.08, and 0.10 over the individual XGBoost models. SHAP analysis identified diagnostic history as a prominent source of predictive information, with additional contributions from demographic, metabolic, physiologic, hematologic, and treatment-related features. Adding CT-derived body composition to structured EHR data provided further multimodal enrichment, increasing AUC by approximately 0.02–0.03 across prediction horizons.

Conclusions

PDAC risk can be progressively enriched using increasingly informative layers of routinely acquired healthcare data across prediction horizons extending to 36 months before diagnosis. Structured EHR information provided substantial discrimination, with additional gains from ensemble learning and complementary information from opportunistically available CT-derived body composition. This approach provides a scalable risk-enrichment strategy using data already generated during routine care to identify individuals who may benefit from targeted PDAC evaluation. External validation and prospective evaluation are required before clinical implementation.

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is among the deadliest human malignancies.[1] Despite decades of research, the five-year survival rate remains at approximately 8%, a figure that has improved only marginally over the past generation.[2–5] The primary driver of this dismal prognosis is late diagnosis: 80–90% of patients present with locally advanced or metastatic disease, at which point curative resection is no longer feasible.[6] Among the minority of patients diagnosed with resectable disease, five-year survival rises substantially, underscoring that the major determinant of outcome is the timing of detection.[7]

Yet no established screening program exists for the general population. The US Preventive Services Task Force recommends against routine pancreatic cancer screening in asymptomatic adults, a position grounded in PDAC's low population prevalence (approximately 13 per 100,000 person-years), the absence of a sufficiently accurate and inexpensive screening test, and the unacceptable false-positive burden that would result from applying imperfect tests to large unselected populations.[8] Current high-risk surveillance programs, guided by the International Cancer of the Pancreas Screening Consortium, are restricted to individuals with defined familial or genetic predispositions and rely on endoscopic ultrasound or MRI, modalities that are expensive, invasive, and unsuitable for broad deployment.[9]

This creates a clinical paradox: early detection dramatically improves survival, but practical population-level early detection tools remain unavailable for the vast majority of patients who will develop PDAC.[10,11] Addressing this gap requires a fundamentally different approach that leverages data already

collected during routine clinical care, without the need for additional tests, procedures, or specialized infrastructure.[12]

Several lines of evidence suggest that subtle, measurable changes occur months to years before a clinical PDAC diagnosis. New-onset hyperglycemia and diabetes, which affect up to 85% of PDAC patients, frequently manifest two to three years before cancer detection.[13–15] Unintentional weight loss, rising blood glucose, alterations in lipid profiles, increasing CA19-9 levels, and shifts in body composition have all been documented in the pre-diagnostic period.[16–20] Patients accumulate gastrointestinal, hepatobiliary, and metabolic diagnoses at rates exceeding those of matched controls.[21] These signals, while individually nonspecific, may collectively form a distinctive pre-diagnostic signature when captured from electronic health records (EHRs).

Recent work has demonstrated the potential of machine learning applied to EHR data for PDAC risk prediction.[12] Placido et al. modeled longitudinal disease-code trajectories from Danish and US registries and demonstrated discrimination up to 36 months before diagnosis; the Prism model used large-scale structured EHR data from 55 US healthcare organizations to predict PDAC risk 6–18 months before diagnosis; and Zhu et al. used demographic, diagnostic, and laboratory information from a large health-system EHR to identify patients at increased risk of pancreatic cancer 2.5–3 years before diagnosis.[21–23] These studies established that routinely collected clinical data contain meaningful prediagnostic signals. However, existing approaches have generally focused on a single representation of the clinical record, such as diagnosis trajectories or a fixed set of structured EHR features. Consequently, less is known about how different routinely acquired clinical domains contribute incrementally to PDAC prediction, whether useful discrimination is retained when laboratory data are sparse, and whether opportunistically available imaging phenotypes can further refine risk. A layered framework that explicitly evaluates these successive sources of information may provide a more clinically interpretable and scalable approach to risk enrichment.

CT-derived body composition represents a particularly attractive enrichment layer. Loss of skeletal muscle and changes in adipose tissue distribution are well-documented precursors to PDAC, and CT scans performed for unrelated clinical indications can be retrospectively analyzed for body composition without additional radiation exposure.[18,24,25]

In this study, we developed and temporally validated a layered multimodal risk-enrichment framework for early PDAC detection. Our central hypothesis was that PDAC risk can be progressively enriched as increasingly informative layers of routinely acquired healthcare data become available. We systematically quantified the incremental contribution of demographics, vital signs and laboratory measurements, medications, and diagnostic history across 3-, 6-, 12-, and 36-month prediction horizons; evaluated performance in a complementary Full-Cohort Analysis without requiring laboratory availability; and examined whether opportunistically derived CT body-composition phenotypes provided additional enrichment beyond structured EHR information. This design was intended to move beyond development of a single prediction model toward a clinically scalable risk-enrichment strategy in which broadly

available EHR data establish an initial risk estimate and additional information can refine that estimate when available.

METHODS

Ethics Statement

This study was approved by the Mayo Clinic institutional review board (IRB ID: 24-007380), which classified the study as minimal risk and waived the requirement for informed consent. Only patients who had authorized the use of their medical records for research were included. EHR data used for model development and evaluation were processed in accordance with institutional data governance standards and applicable privacy regulations.

The study was reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence (TRIPOD + AI) reporting guidelines.[26,27]

Study Design and Data Source

We conducted a retrospective matched case-control study using longitudinal EHR data from the three Mayo Clinic sites (Rochester, Minnesota; Jacksonville, Florida; and Phoenix, Arizona) to develop and temporally validate machine learning models for early prediction of PDAC.

The case cohort included adults (≥ 18 years) with histologically confirmed PDAC diagnosed between January 1, 2000, and December 31, 2020. A total of 3,332 PDAC cases were identified from a previously characterized PDAC cohort.[18] Candidate controls were drawn from adults receiving care at Mayo Clinic between January 1, 2000, and December 31, 2021. The control-source period was extended through 2021 to allow identification of eligible control encounters temporally proximal to cases diagnosed near the end of 2020. Patients with a history of cirrhosis, bariatric surgery, or heart failure were excluded from the candidate control population because these conditions can independently produce substantial metabolic, anthropometric, physiologic, and healthcare-utilization changes that overlap with clinical features evaluated as potential prediagnostic PDAC signals.[28–30] These exclusions were not applied to the PDAC case cohort. Among the 3,332 PDAC cases, only 126 (3.8%) had at least one of these conditions documented before PDAC diagnosis. Because PDAC cases were substantially less numerous than potential controls and these conditions affected only a small proportion of the established case cohort, these cases were retained in the main study cohort. To evaluate whether this asymmetric eligibility criterion materially influenced model performance, we performed a sensitivity analysis in the Full-Cohort Analysis after excluding the 126 affected PDAC cases.

Controls were selected using a propensity score-based matching framework incorporating age, sex, race (White, Black, Asian, or Other/Unknown), year of encounter, and number of healthcare encounters.

Propensity scores were estimated using logistic regression, and nearest-neighbor matching without replacement was performed. For each case, up to 10 candidate controls were selected based on proximity of the estimated propensity scores, prioritizing encounters temporally closest to the corresponding case's diagnosis date. This initial matching procedure yielded 3,332 cases and 29,990 matched controls.

Additional eligibility criteria were then applied to the matched control cohort. Controls with any diagnosis code from a prespecified comprehensive cancer ICD list (**Supplementary Table S1**) were excluded, leaving 21,995 cancer-free controls. Controls were additionally required to have an eligible encounter within ± 3 years of the diagnosis date of the corresponding case to maintain temporal alignment, resulting in 14,388 controls and 3,332 PDAC cases. The encounter closest to the corresponding case's diagnosis date was assigned as the control's pseudo-diagnosis date.

To evaluate prediction at clinically relevant lead times, separate datasets were constructed for prediction horizons of 3, 6, 12, and 36 months before diagnosis. For each horizon, the prediction index date was defined relative to the PDAC diagnosis date for cases and the corresponding pseudo-diagnosis date for controls. Predictor variables were derived exclusively from data available during the five years preceding each horizon-specific index date (Fig. 1).

Additional details regarding study procedures, data preparation, model development, and secondary analyses are provided in the **Supplementary Methods**.

Covariates

Structured EHR covariates included demographics (age, sex, race); vital signs (systolic and diastolic blood pressure, pulse rate, temperature, height, weight, and body mass index [BMI]); laboratory results (see “Blood-Based Biomarker Assessment” below); medications, represented as binary (yes/no) indicators for 50 therapeutic drug categories or classes (e.g., analgesics, antibiotics, antihyperglycemics, cardiac drugs, gastrointestinal agents); diagnosis codes, in which ICD-9 and ICD-10 codes were summarized using the Clinical Classifications Software (CCS),[31] and operationalized as binary presence indicators; and procedures, encoded as binary indicators for relevant procedures (CT abdomen, MRI, ERCP, EUS, and selected surgical procedures). CCS category definitions used in this study are provided in **Supplementary Table S18**.

Blood-Based Biomarker Assessment

Blood-based biomarkers were prioritized because of their established relationship to nutrition, metabolism, and systemic inflammation, processes known to change during PDAC development.[18] The biomarker panel comprised albumin, prealbumin, C-reactive protein (CRP), high-sensitivity CRP (hs-CRP), fasting blood glucose (FBG), hemoglobin, hemoglobin A1c (HbA1c), high-density lipoprotein (HDL)

cholesterol, low-density lipoprotein (LDL) cholesterol, total cholesterol, triglycerides, non-HDL cholesterol, the cholesterol/HDL ratio, and white blood cell count.

Study Cohorts: Biomarker-Cohort and Full-Cohort Analyses

Two complementary analyses were conducted to evaluate the layered risk-enrichment framework under differing levels of laboratory data availability. The **Biomarker-Cohort Analysis** was restricted to patients with at least one blood-based biomarker measurement during the observation period preceding the horizon-specific index date. This cohort enabled systematic evaluation of how progressively richer layers of structured EHR data contribute to PDAC risk discrimination in patients with available laboratory information.

A second analysis used the full cohort (**Full-Cohort Analysis**), retaining all eligible patients regardless of laboratory data availability. This analysis addressed the substantial proportion of patients without available blood-based biomarker measurements and assessed PDAC discrimination under patterns of missing laboratory data encountered in routine clinical practice. XGBoost was used to accommodate missing values natively, allowing patients with partially or entirely missing laboratory measurements to remain in the analysis without external imputation.

Imaging Enrichment Analysis

From the matched cohort, we identified patients who received at least one CT examination during the five-year period preceding the index date, as detailed in **Supplementary Methods S6**. From these examinations, we extracted subcutaneous adipose tissue area (SAT), visceral adipose tissue area (VAT), skeletal muscle area (SMA) and radiodensity, and vertebral bone area and radiodensity from eligible abdominal axial CT series using a previously described deep-learning image segmentation algorithm.[32,33] Body composition was assessed from a three-dimensional region of interest extending from the top of the L2 vertebra to the bottom of L4; tissue volume was divided by the height of the region of interest to report measurements in units of area, consistent with prior studies.[32]

To evaluate body composition as an opportunistic imaging enrichment layer, CT-derived body composition measures were incorporated into the Biomarker-Cohort without requiring body composition (or blood-based biomarker) data to be available for every patient. Models integrating structured EHR features with body composition were compared with corresponding structured EHR models to quantify the incremental predictive value of imaging-derived phenotypes across prediction horizons. XGBoost was used to accommodate missing body composition measurements natively, allowing patients without available CT-derived measurements to remain in the analysis. A complementary analysis restricted to patients with available body composition measurements was also performed and is reported in the **Supplementary Results**.

Layered Risk-Enrichment Framework

To quantify the incremental value of routinely collected healthcare data for PDAC risk stratification, we developed a layered risk-enrichment framework representing progressively richer sources of patient information (Fig. 2). This framework was designed to emulate a clinically scalable approach to early detection, whereby risk estimates are refined as additional data become available during routine clinical care.

The first layer consisted of demographic characteristics alone (Baseline Clinical Model), including age, sex, and race. The second layer incorporated vital signs and laboratory measurements (Core Clinical Model), representing routinely collected physiologic and biochemical indicators. The third layer added medication exposures (Clinical + Treatment Model), capturing treatment patterns and healthcare utilization that may reflect evolving disease processes. The fourth layer incorporated diagnostic history summarized using Clinical Classifications Software (CCS) categories (Extended Clinical Model), representing accumulated comorbidity burden and clinical manifestations documented throughout routine care. To evaluate the robustness of diagnostic information, we additionally constructed a Refined Clinical Model that excluded PDAC-related CCS categories (**Supplementary Table S2**). A further Comorbidity-Enriched Model combined demographics, medications, and CCS categories without vital signs or laboratory results to characterize performance in settings where laboratory measurements are entirely absent.

Recognizing that additional clinically relevant information may become available for a subset of patients, we further evaluated a fifth layer consisting of opportunistically acquired imaging biomarkers derived from abdominal CT scans as described above. In patients with available imaging, CT-derived body composition measures were integrated with the structured EHR models to determine whether imaging-derived phenotypes provided incremental predictive value beyond routinely collected clinical data alone.

Together, the layered framework provides a systematic evaluation of how successive domains of clinical information contribute to PDAC risk prediction and reflects a pragmatic clinical paradigm in which increasingly informative data sources can be leveraged to refine risk estimates and identify individuals who may benefit from targeted diagnostic evaluation.

Handling Missing Data

Medication, procedure, and CCS diagnosis variables were defined by presence and had no missing values. Laboratory data availability varied considerably across patients and horizons (range, 21% to 100% missingness, **Supplementary Figure S1**). In the Biomarker-Cohort Analysis, patients with no laboratory data within the observation window were excluded, laboratory variables with more than 40% missingness were dropped, and remaining missing laboratory and vital-sign values were imputed using the random-forest-based MissForest algorithm.[34] In the Full-Cohort Analysis, all patients were retained irrespective of laboratory availability, and XGBoost was used throughout to natively accommodate missing values without external imputation. XGBoost was also used throughout the body-composition analyses to natively accommodate missing body-composition and laboratory measurements.

Both analyses were derived from the same matched source cohort and used the same feature definitions, while differing in eligibility based on laboratory availability and in missing-data handling.

Feature Engineering

Horizon-specific cross-sectional datasets were constructed for each of the 3-, 6-, 12-, and 36-month prediction horizons. Medications, procedures, and CCS diagnostic categories were represented as binary indicators reflecting presence before the index date. Vital signs, laboratory, and body composition measurements were summarized as both the most recent value and the median value observed within the five-year lookback window, to capture complementary information on current physiologic status and longitudinal trends. Continuous variables were normalized using min-max scaling; categorical variables were one-hot-encoded; and binary features with a prevalence below 0.01% were excluded to reduce sparsity.

Development and Temporal Validation Cohorts

To emulate prospective deployment and evaluate model generalizability, the matched cohort was partitioned into temporally distinct development and validation cohorts based on the year of diagnosis. Patients diagnosed between 2000 and 2013 were assigned to the development cohort and used for model training, hyperparameter optimization, and internal cross-validation. Patients diagnosed between 2014 and 2020 were reserved as an independent temporal validation cohort and were not used during any stage of model development (Fig. 1). This temporal separation was applied consistently across all prediction horizons, feature groups, and analytic cohorts.

Machine Learning Model Development

We developed and compared four machine learning approaches: regularized logistic regression (LASSO / Elastic Net),[35,36] random forest,[37] extreme gradient boosting (XGBoost),[38] and TabNet.[39] These algorithms were selected to represent complementary modeling strategies, including linear methods, tree-based ensemble learning, and deep learning for tabular data.

For each prediction horizon and layer of the risk-enrichment framework, models were trained using identical input features and preprocessing procedures. Hyperparameters for each algorithm were optimized exclusively within the development cohort using stratified 5-fold cross-validation. Automated Bayesian optimization was employed to identify parameter combinations that maximized discrimination while minimizing overfitting.[40] The final model configuration for each algorithm was selected based on cross-validated discrimination and subsequently evaluated without further tuning in the temporally independent validation cohort. Detailed model-training procedures and hyperparameter search spaces are provided in the **Supplementary Methods S7 and Supplementary Table S3**. The best-performing algorithm for each prediction horizon was subsequently carried forward for additional analyses.

Ensemble Learning

To determine whether combining complementary machine learning models could further improve predictive performance, we evaluated several ensemble learning strategies that combined predictions from models trained on the best-performing structured EHR enrichment layer (Extended Clinical Model). Ensemble learning was evaluated only within the Biomarker-Cohort Analysis.

Ensemble meta-learners were trained exclusively within the development cohort using out-of-fold (OOF) predicted probabilities generated by the base models (regularized logistic regression, random forest, XGBoost, and TabNet) trained on the Extended Clinical Model.[41] Specifically, each patient's base-model probability was produced via the cross-validation framework described previously, ensuring the prediction was derived from a model trained only on the other folds. The OOF predicted probabilities were then used as inputs to four ensemble strategies: (1) XGBoost stacking, in which the four probabilities were combined using an XGBoost model; (2) XGBoost stacking augmented with percent-change laboratory features (computed between the earliest and latest available measurements within the observation window for hemoglobin, leukocytes, MCV, glucose, cholesterol, HDL, LDL, triglycerides, HbA1c, non-HDL cholesterol, prealbumin, hs-CRP, CRP, albumin, cholesterol/HDL ratio, and fasting glucose) computed from the OOF; (3) simple probability averaging; and (4) soft-voting ensembles. Once trained, the ensemble models were applied without further fitting, optimization, or recalibration to the corresponding base-model predictions and laboratory-change features in the temporally independent validation cohort.

Longitudinal percent-change laboratory features were evaluated only at the ensemble stage as an additional model-enhancement strategy. The base models retained the prespecified feature representation used throughout the layered risk-enrichment framework. Given the substantially greater missingness of percent-change features, incorporating them into all base models would also have altered the analytic feature space and missing-data burden used for the primary comparisons. We consequently evaluated multiple preprocessing strategies that differed in the extent of feature filtering and missing-data handling before ensemble model development. These strategies assessed the tradeoff between retaining longitudinal information and limiting the influence of highly incomplete features. Detailed descriptions of the preprocessing strategies are provided in **Supplementary Methods S8**.

Performance Metrics

Using the predicted probability for each model across each prediction horizon and layer of the risk-enrichment framework and ground truth in the development cohort, we computed the sensitivity and specificity across possible cutoff points or thresholds and constructed ROC curves. The classification threshold for each model was defined as the predicted probability above which the model classified patient as high risk for PDAC. We selected the threshold corresponding to the point on the ROC curve with minimum distance to the upper-left corner (0, 1). Using this threshold and predicted probabilities in the temporal validation cohort, we computed sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and F1 score, and the threshold independent area under the ROC curve (AUC). Because PDAC prevalence was determined by the matched case-control sampling

design, PPV and NPV reflect the observed prevalence within the analytic cohorts and should not be interpreted as population-level predictive values.

Model calibration was assessed in the temporal validation cohort by grouping patients according to deciles of predicted risk and comparing the mean predicted PDAC probability with the observed event frequency within each group. Calibration plots were generated separately for each prediction horizon, with the 45-degree reference line representing perfect agreement between predicted and observed probabilities. Calibration was evaluated for the Extended Clinical Model in both the Biomarker-Cohort and Full-Cohort analyses. Because the case prevalence was determined by the sampling design, these analyses evaluate calibration within the sampled analytic cohorts and do not establish calibration of absolute PDAC risk in the underlying clinical population.

Interpretability

To characterize the clinical features contributing to PDAC risk predictions, we performed post-hoc interpretability analysis using Shapley Additive Explanations (SHAP).[42] SHAP values were calculated for the best-performing structured EHR model to quantify the contribution of individual features to model predictions and the direction of their associations with predicted PDAC risk. We report the top 20 features ranked by mean absolute SHAP value and visualize their distributions and directional effects on model predictions (**Supplementary Methods S9**).

Statistical Analysis

Cohort characteristics are reported using mean and standard deviation (SD) for continuous variables and counts and percentages for categorical variables. We report the mean the mean and 95% confidence intervals (CI) of the performance measures across the 5-fold cross-validation on the development cohort and the point estimates on the temporal validation cohort. No sex-based analyses were performed.

Statistical analyses were performed using R (version 4.4.1), with machine-learning analyses conducted in Python. Python packages used for model development and analysis included scikit-learn (version 1.8.0), SciPy (version 1.17.1), XGBoost (version 3.0.0), Optuna (version 4.2.1) and PyTorch TabNet (version 4.1.0).[38–40,43,44]

RESULTS

Cohort Characteristics

At the 3-month prediction horizon, the Biomarker-Cohort Analysis included 11,965 patients, comprising 1,570 individuals with PDAC and 10,395 matched cancer-free controls (Table 1). The cohort was predominantly White, and slightly more than half were male. Compared with controls, patients with PDAC were older (mean age, 71.9 vs. 66.8 years) and had higher HbA1c (6.77% vs. 6.16%), fasting glucose (118.1 vs. 106.7 mg/dL), and random glucose (140.5 vs. 127.4 mg/dL). PDAC cases also had more

recorded CCS diagnosis categories (mean, 11.8 vs. 9.0) and procedures (mean, 0.48 vs. 0.19), including greater use of abdominal CT, MRI, and ERCP. Differences in vital signs and BMI were comparatively smaller (Table 1). Cohort characteristics at the 6-, 12-, and 36-month prediction horizons are provided in **Supplementary Tables S4-S6**.

Table 1

Characteristics of the Biomarker-Cohort Analysis at the 3-Month Prediction Horizon. Selected demographics, clinical, laboratory, diagnostic, and procedural characteristics of PDAC cases and matched cancer-free controls included in the Biomarker-Cohort Analysis at the 3-month prediction horizon. Continuous variables are presented as mean (SD) and categorical variables as n (%); SMD represents the standardized mean difference between PDAC cases and controls; Missing (%) represents the proportion of patients without a recorded value for the corresponding variable. Characteristics shown represent a subset of the variables used for model development.

Characteristic	Controls (n = 10395)	PDAC (n = 1570)	SMD	Missing (%)
Male sex, n (%)	5588 (53.8)	877 (55.9)	0.042	0.0
White race, n (%)	10156 (97.7)	1481 (94.3)	0.179	0.0
Black race, n (%)	48 (0.5)	28 (1.8)		0.0
Asian race, n (%)	37 (0.4)	14 (0.9)		0.0
Other race, n (%)	74 (0.7)	20 (1.3)		0.0
Unknown race, n (%)	80 (0.8)	27 (1.7)		0.0
Age, mean (SD), years	66.80 (15.21)	71.93 (10.64)	0.391	0.0
Systolic BP, mean (SD), mmHg	129.48 (19.92)	129.09 (17.49)	0.021	34.3
Diastolic BP, mean (SD), mmHg	73.10 (13.27)	72.68 (11.59)	0.034	34.4
Pulse rate, mean (SD), bpm	74.55 (13.87)	72.33 (12.35)	0.169	68.9
BMI, mean (SD), kg/m²	30.31 (10.61)	30.72 (11.57)	0.037	47.7
Hemoglobin, mean (SD), g/dL	13.12 (1.87)	13.27 (1.78)	0.079	29.7
Glucose, mean (SD), mg/dL	127.41 (40.87)	140.45 (43.01)	0.311	76.5
Fasting glucose, mean (SD), mg/dL	106.71 (27.87)	118.09 (39.63)	0.332	92.7
HbA1c, mean (SD), %	6.16 (1.17)	6.77 (1.52)	0.451	85.9
Total cholesterol, mean (SD), mg/dL	183.96 (43.55)	176.96 (44.06)	0.16	56.3
LDL, mean (SD), mg/dL	103.01 (36.09)	98.34 (34.63)	0.132	62.8
HDL, mean (SD), mg/dL	52.77 (17.90)	51.91 (18.14)	0.048	62.2
Albumin, mean (SD), g/dL	4.10 (0.49)	4.15 (0.45)	0.115	87.1
CRP, mean (SD), mg/L	25.26 (44.40)	31.25 (54.45)	0.121	93
No. medications, mean (SD)	5.63 (8.02)	5.13 (7.94)	0.063	0.0

The Full-Cohort Analysis at the 3-month prediction horizon included 17,720 patients, comprising 3,332 PDAC cases and 14,388 controls (**Supplementary Table S7**). Cases and controls were similar in sex

distribution (56.3% vs. 53.3% male; SMD, 0.061), while cases were older (mean age, 69.2 vs. 67.2 years; SMD, 0.161) and included a lower proportion of White patients (87.3% vs. 96.9%; SMD, 0.366). Laboratory measurements were frequently unavailable in this cohort, with missingness exceeding 80% for several metabolic and inflammatory biomarkers. Visual examination of missingness patterns did not reveal an obvious common pattern across laboratory measures (**Supplementary Figure S1**). Among patients with available measurements, PDAC cases had higher glucose, fasting glucose, and HbA1c levels than controls. Characteristics of the Full-Cohort Analytic at the remaining prediction horizons are provided in **Supplementary Tables S8-S10**.

The body-composition cohort included 2,912 patients with available CT-derived measurements, comprising 691 PDAC cases and 2,221 controls (**Supplementary Table S11**). Compared with controls, PDAC cases had higher mean skeletal muscle area (153.2 vs. 144.8 cm^2), subcutaneous adipose tissue area (223.5 vs. 205.3 cm^2), and visceral adipose tissue area (150.1 vs. 121.4 cm^2), with the largest difference observed for visceral adipose tissue (SMD, 0.284).

Model Development and Validation

Models were developed separately for each prediction horizon and structured EHR enrichment layer, with algorithm-specific hyperparameters optimized within the development cohort. Final selected hyperparameters for each machine learning algorithm, enrichment layer, and prediction horizon in the Biomarker-Cohort Analysis are provided in **Supplementary Table S12**.

Performance of Structured EHR Models: Biomarker-Cohort Analysis

We first evaluated the layered risk-enrichment framework in the Biomarker-Cohort Analysis, beginning with demographic characteristics and progressively incorporating vital signs and laboratory measurements, medication exposures, and diagnostic history. Across prediction horizons, demographics alone provided limited discrimination, with temporal validation AUCs ranging from 0.58 at 3 months to 0.60 at 36 months (Table 2). Incorporating vital signs and laboratory measurements increased AUCs to 0.63–0.66, and the addition of medication exposures further increased AUCs to 0.67–0.69 across horizons.

Table 2

Incremental Contribution of Structured EHR Information Layers in the Biomarker-Cohort Analysis. Temporal validation performance of XGBoost models incorporating progressively richer layers of structured EHR information at 3-, 6-, 12-, and 36-month prediction horizons. Values are point estimates from the temporal validation cohort. **Supplementary Table S14** presents complete results for all machine learning models.

Clinical information layer	AUC	Sensitivity	Specificity	PPV	NPV	F1
3 Months						
Baseline Clinical Model: demographics only	0.58	0.74	0.40	0.26	0.82	0.36
Core Clinical Model: demographics + vitals + labs	0.63	0.55	0.64	0.28	0.86	0.40
Clinical + Treatment Model: Core Clinical Model + medications	0.67	0.55	0.72	0.32	0.86	0.43
Extended Clinical Model: Clinical + Treatment Model + CCS diagnosis codes	**0.84**	0.68	0.82	0.43	0.92	0.56
Refined Clinical Model: Extended Clinical Model with PDAC related CCS codes removed	0.82	0.71	0.78	0.41	0.92	0.54
Comorbidity-Enriched Model: demographics + medications + CCS codes	0.81	0.83	0.65	0.45	0.92	0.59
6 Months						
Baseline Clinical Model: demographics only	0.59	0.73	0.41	0.27	0.81	0.37
Core Clinical Model: demographics + vitals + labs	0.66	0.56	0.69	0.30	0.87	0.43
Clinical + Treatment Model: Core Clinical Model + medications	0.69	0.65	0.65	0.32	0.88	0.45
Extended Clinical Model: Clinical + Treatment Model + CCS diagnosis codes	**0.86**	0.74	0.81	0.45	0.94	0.58
Refined Clinical Model: Extended Clinical Model with PDAC related CCS codes removed	0.84	0.79	0.75	0.39	0.94	0.54
Comorbidity-Enriched Model: demographics + medications + CCS codes	0.78	0.84	0.59	0.41	0.91	0.56
12 Months						
Baseline Clinical Model: demographics only	0.59	0.77	0.36	0.27	0.82	0.36

Clinical information layer	AUC	Sensitivity	Specificity	PPV	NPV	F1
Core Clinical Model: demographics + vitals + labs	0.64	0.68	0.54	0.30	0.85	0.41
Clinical + Treatment Model: Core Clinical Model + medications	0.68	0.68	0.60	0.33	0.86	0.43
Extended Clinical Model: Clinical + Treatment Model + CCS diagnosis codes	**0.81**	0.72	0.74	0.45	0.90	0.55
Refined Clinical Model: Extended Clinical Model with PDAC related CCS codes removed	0.78	0.71	0.71	0.41	0.89	0.52
Comorbidity-Enriched Model: demographics + medications + CCS codes	0.77	0.81	0.59	0.41	0.88	0.53
36 Months						
Baseline Clinical Model: demographics only	0.60	0.67	0.49	0.27	0.83	0.39
Core Clinical Model: demographics + vitals + labs	0.64	0.81	0.40	0.29	0.84	0.40
Clinical + Treatment Model: Core Clinical Model + medications	0.67	0.67	0.60	0.31	0.87	0.43
Extended Clinical Model: Clinical + Treatment Model + CCS diagnosis codes	**0.79**	0.80	0.65	0.41	0.91	0.53
Refined Clinical Model: Extended Clinical Model with PDAC related CCS codes removed	0.78	0.80	0.63	0.39	0.89	0.51
Comorbidity-Enriched Model: demographics + medications + CCS codes	0.72	0.88	0.46	0.37	0.85	0.49

The largest incremental gain occurred after incorporation of CCS diagnosis categories, which raised AUCs to 0.84, 0.86, 0.81, and 0.79 at 3, 6, 12, and 36 months, respectively (**Supplementary Figure S2**). This represented an increase of approximately 0.13–0.17 over models containing demographics, vital signs, laboratory measurements, and medications and indicates that diagnostic history represents the dominant structured-data layer for PDAC risk enrichment. Importantly, this performance gain is unlikely to reflect information leakage. All diagnosis codes used to derive CCS features preceded the horizon-specific prediction index date by design; the most recent diagnosis code occurred at least 90, 180, 365, and 1,095 days before PDAC diagnosis at the 3-, 6-, 12-, and 36-month horizons, respectively (**Supplementary Table S13**). The Refined Clinical Model, which excluded CCS categories directly related to PDAC diagnosis or diagnostic evaluation, showed modest attenuation in performance, with AUCs of 0.82, 0.84, 0.78, and 0.78 across the four horizons. The Comorbidity-Enriched Model, comprising

demographics, medications, and CCS categories without vital signs or laboratory measurements, achieved AUCs of 0.81, 0.78, 0.77, and 0.72, respectively.

Across the tested machine learning algorithms, XGBoost consistently achieved the highest temporal validation performance across prediction horizons (**Supplementary Table S14**). Random forest and TabNet showed lower and less consistent discrimination across horizons.

Calibration plots showed that predicted probabilities generally tracked observed PDAC frequencies across prediction horizons, although the models tended to underestimate observed risk in several intermediate- and higher-risk groups (**Supplementary Figure S3**).

Performance of Structured EHR Models: Full-Cohort Analysis

The layered risk-enrichment framework was also evaluated in the Full-Cohort Analysis, which retained eligible patients regardless of laboratory data availability and used XGBoost to accommodate missing values natively. Here, we report performance of the fully enriched structured EHR feature set (Extended Clinical Model), modeled using XGBoost. Temporal validation AUCs were 0.83, 0.82, 0.80, and 0.77 at 3, 6, 12, and 36 months, respectively (**Supplementary Table S15**). Temporal validation ROC curves demonstrated consistent discrimination across prediction horizons, with some attenuation at longer prediction horizons while remaining informative at 36 months (**Supplementary Figure S4**). Except for the 6-month horizon, these values were similar to the corresponding AUCs of 0.84, 0.86, 0.81, and 0.79 achieved by the XGBoost Extended Clinical Model in the Biomarker-Cohort Analysis, despite substantially greater laboratory missingness in the Full-Cohort Analysis.

Calibration plots similarly showed generally good correspondence between predicted and observed PDAC frequencies across prediction horizons, with some underestimation of observed risk at intermediate predicted probabilities (**Supplementary Figure S5**).

To evaluate whether the asymmetric exclusion of cirrhosis, bariatric surgery, and heart failure from the control population influenced model performance, we repeated the Full-Cohort Analysis after excluding 126 PDAC cases (3.8%) with at least one of these conditions before diagnosis. Temporal-validation AUCs remained essentially unchanged at 0.82, 0.82, 0.80, and 0.76 at the 3-, 6-, 12-, and 36-month horizons, respectively, compared with 0.83, 0.82, 0.80, and 0.76 in the original analysis (**Supplementary Table S19**).

Ensemble Model Performance

Ensemble learning was evaluated within the Biomarker-Cohort Analysis using predictions from base models trained on the best-performing structured EHR enrichment layer, the Extended Clinical Model. The best-performing ensemble was an XGBoost stacking model that combined predicted probabilities from the four base models with the full set of longitudinal percent-change laboratory features, with

missing values handled natively by XGBoost (**Supplementary Methods S8**). This model achieved temporal validation AUCs of 0.88, 0.91, 0.89, and 0.89 at 3, 6, 12, and 36 months, respectively (Table 3), corresponding to absolute AUC improvements of 0.04, 0.06, 0.08, and 0.10 over the individual XGBoost models, with progressively larger gains at longer prediction horizons. Alternative ensemble configurations using different preprocessing and missing-data strategies for the longitudinal laboratory features, as well as probability averaging and soft voting, are reported in **Supplementary Table S16**.

Table 3

Temporal Validation Performance of Ensemble Learning and CT-Derived Imaging Enrichment in the Biomarker-Cohort Analysis. Temporal validation performance of the ensemble-learning and imaging-enrichment analyses across the 3-, 6-, 12-, and 36-month prediction horizons. Ensemble methods combined predictions from base models trained on the best-performing structured EHR enrichment layer (Extended Clinical Model). For XGBoost stacking with longitudinal laboratory change, results are shown for the best-performing missing-data strategy, which retained the full set of percent-change laboratory features and used native XGBoost missing-value handling. The imaging-enrichment analysis evaluated CT-derived body-composition features without requiring body-composition measurements to be available for all patients; missing imaging features were handled natively by XGBoost. Additional ensemble preprocessing strategies and complementary imaging-subcohort analyses are reported in the **Supplementary Material**.

Model / Feature Set	Horizon (months)	AUC	Sensitivity	Specificity	PPV	NPV	F1
Ensemble learning							
XGBoost stacking + longitudinal laboratory change	3	**0.88**	0.82	0.81	0.56	0.94	0.67
	6	**0.91**	0.83	0.85	0.65	0.93	0.73
	12	**0.89**	0.79	0.86	0.62	0.93	0.69
	36	**0.89**	0.90	0.74	0.51	0.96	0.65
Probability averaging	3	0.84	0.71	0.81	0.52	0.91	0.60
	6	0.86	0.76	0.81	0.54	0.92	0.63
	12	0.80	0.73	0.72	0.43	0.90	0.54
	36	0.80	0.74	0.73	0.44	0.91	0.55
Soft-voting classifier	3	0.85	0.83	0.70	0.44	0.93	0.57
	6	0.86	0.83	0.73	0.48	0.94	0.61
	12	0.82	0.68	0.79	0.48	0.90	0.56
	36	0.81	0.80	0.68	0.42	0.92	0.55
Imaging enrichment							
Structured EHR + body composition	3	**0.85**	0.77	0.78	0.47	0.93	0.59
	6	**0.84**	0.70	0.84	0.43	0.93	0.57
	12	**0.83**	0.72	0.79	0.44	0.91	0.56
	36	**0.82**	0.80	0.69	0.42	0.93	0.55
Body composition + demographics	3	0.64	0.48	0.77	0.28	0.84	0.39
	6	0.65	0.46	0.78	0.30	0.83	0.40
	12	0.64	0.43	0.80	0.29	0.83	0.39

Model / Feature Set	Horizon (months)	AUC	Sensitivity	Specificity	PPV	NPV	F1
Ensemble learning							
	36	0.66	0.45	0.80	0.30	0.84	0.41

Imaging Enrichment Analysis

We next evaluated CT-derived body composition as an opportunistic imaging enrichment layer within the Biomarker-Cohort Analysis, without requiring body composition measurements to be available for every patient. Addition of body composition measures to structured EHR features yielded temporal validation AUCs of 0.85, 0.84, 0.83, and 0.82 at 3, 6, 12, and 36 months, respectively (Table 3). Compared with the corresponding structured EHR models, incorporation of body composition was associated with modestly higher discrimination across prediction horizons.

Models based on body composition and demographic characteristics alone achieved AUCs of 0.64–0.66 across horizons, indicating substantially lower discrimination than models integrating body composition with structured EHR information. A complementary analysis restricted to patients with available CT-derived body composition measurements was also conducted, with results reported in **Supplementary Table S17**.

Explainability: Drivers of Predicted PDAC Risk

SHAP analysis of the Extended Clinical Model in the Biomarker-Cohort Analysis demonstrated contributions from diagnostic, demographic, physiologic, hematologic, medication, and anthropometric features, with their relative importance varying across prediction horizons (Fig. 3). Other screening for suspected conditions (CCS 258) was consistently among the most influential features, while age and overall diagnostic burden, represented by the total number of CCS categories, were prominent across horizons. Several diagnostic categories also contributed substantially, including congestive heart failure (CCS 108), diabetes mellitus without complication (CCS 49), pancreatic disorders other than diabetes (CCS 152), and other and unspecified benign neoplasms (CCS 47). Continuous measures including blood pressure, body weight, BMI, hemoglobin, leukocyte count, and mean corpuscular volume contributed with varying importance across prediction horizons.

The Full-Cohort Analysis showed both shared and distinct patterns (**Supplementary Figure S6**). Overall diagnostic burden, CCS 258, and age remained among the most influential features, supporting the persistence of these signals when patients were retained regardless of laboratory data availability. The Full-Cohort models also placed greater prominence on routinely available physiologic and metabolic measures, including temperature, blood pressure, cholesterol, fasting glucose, body weight, and height. Temperature measures appeared among the leading features at multiple horizons, a pattern less evident in the Biomarker-Cohort Analysis.

Across both analyses, SHAP distributions demonstrated substantial heterogeneity in feature contributions. For several continuous variables, both high and low values contributed to increased predicted risk in different patients, consistent with nonlinear effects and interactions captured by XGBoost. Accordingly, these explainability findings characterize how features contributed to model predictions and should not be interpreted as independent or causal associations with PDAC development.

DISCUSSION

In this retrospective cohort study, we developed and temporally validated a layered multimodal risk-enrichment framework for early prediction of PDAC using routinely acquired clinical data. The central finding is that PDAC risk can be progressively enriched as increasingly informative layers of routinely collected healthcare data become available. In the Biomarker-Cohort Analysis, the best-performing individual structured EHR model achieved temporal validation AUCs of 0.84, 0.86, 0.81, and 0.79 at 3, 6, 12, and 36 months before diagnosis, respectively. Similar discrimination was observed in the complementary Full-Cohort Analysis, which retained patients regardless of laboratory data availability. Ensemble learning further improved discrimination, particularly at longer prediction horizons, while opportunistically available CT-derived body composition provided complementary imaging information. Together, these findings support further evaluation of a risk-enrichment paradigm that leverages data already generated during routine care to identify individuals who may benefit from targeted diagnostic evaluation.

The structured EHR findings suggest that the prediagnostic PDAC signal is distributed across multiple clinical domains and emerges from the accumulation of multiple abnormalities. Demographic characteristics alone provided limited discrimination, and the addition of vital signs and laboratory measurements produced incremental gains. Cases in the Biomarker-Cohort Analysis also exhibited higher glucose, fasting glucose, and HbA1c levels than controls, consistent with the metabolic disturbances known to precede clinical PDAC diagnosis. Yet metabolic and physiologic features alone were insufficient to achieve the discrimination observed after progressively incorporating medications and diagnostic history. These findings support the premise that prediagnostic PDAC risk may emerge from the accumulation of multiple individually weak or nonspecific signals across routinely collected healthcare data. Machine learning may be particularly suited to this setting because it can integrate such signals across clinical domains and identify combinations that would be difficult to recognize from any individual feature.

The contribution of diagnostic history was particularly notable. Incorporating CCS diagnosis categories produced the largest incremental gain among the structured EHR information layers, increasing AUC by approximately 0.13 to 0.17 beyond models containing demographics, vital signs, laboratory measurements, and medications. Some of this information may reflect evolving clinical evaluation as an occult malignancy approaches diagnosis. However, all diagnosis codes preceded the horizon-specific prediction index date by design, and discrimination was only modestly attenuated in the Refined Clinical

Model after excluding categories directly related to PDAC or its diagnostic evaluation. Furthermore, a model containing demographics, medications, and diagnostic history without laboratory or vital-sign information retained meaningful discrimination. Collectively, these findings suggest that accumulated diagnostic history contains substantial predictive information beyond isolated metabolic abnormalities or codes directly associated with pancreatic cancer workup.

The SHAP analyses provide additional insight into how the model used this information across clinical domains and how these patterns differed between the two complementary analyses. In the Biomarker-Cohort Analysis, other screening for suspected conditions (CCS 258), age, and overall diagnostic burden were consistently influential across prediction horizons, alongside contributions from diagnostic categories, hematologic indices, blood pressure, medications, and anthropometric measures. Diabetes mellitus without complication (CCS 49) was also influential at the 3- and 6-month horizons, consistent with the established relationship between dysglycemia and prediagnostic PDAC.[13,18] The Full-Cohort Analysis showed substantial overlap in these leading predictors, particularly for age, CCS 258, and overall diagnostic burden, but also identified prominent contributions from routinely collected physiologic and metabolic measures, including temperature, blood pressure, cholesterol, fasting glucose, body weight, and height. Temperature measures were particularly notable across multiple horizons in the Full-Cohort Analysis but were less prominent in the Biomarker-Cohort Analysis. These differences suggest that the relative information used by the model can shift with cohort composition and patterns of data availability, while diagnostic history and demographic information remain important sources of predictive signal.

Feature importance also varied across prediction horizons. Diagnostic and healthcare-utilization signals were prominent closer to diagnosis, whereas age, overall diagnostic burden, hematologic and physiologic measures, and a broader range of diagnoses contributed at longer prediction horizons. SHAP distributions further demonstrated considerable heterogeneity across individuals, with both higher and lower values of several continuous features contributing to increased predicted risk in different patients. These patterns are consistent with nonlinear relationships and feature interactions captured by XGBoost and support the broader premise that prediagnostic PDAC risk may emerge from patient-specific combinations of multiple, individually nonspecific clinical signals. SHAP values characterize model behavior, however, and should not be interpreted as independent associations or evidence of causal or biological relationships with PDAC development.

The present study builds on a rapidly expanding literature demonstrating that EHR-derived information can identify individuals at increased risk for PDAC before clinical diagnosis.[45] Placido et al. showed that deep learning applied to longitudinal disease-code trajectories could predict pancreatic cancer up to 36 months before diagnosis; Prism demonstrated that a fixed set of structured EHR features could support risk prediction across a large multi-institutional US population; and Zhu et al. used demographic, diagnostic, and laboratory information from a large health-system EHR to identify patients at increased risk of pancreatic cancer 2.5–3 years before diagnosis.[21–23] The contribution of the present study is different. Our primary objective was not simply to optimize discrimination from a single feature

representation, but to determine how PDAC risk changes as successive clinical information domains are added. This design demonstrated that diagnostic history provided the largest structured-data enrichment, quantified the additional contribution of ensemble learning and longitudinal laboratory change, showed that discrimination was largely maintained when laboratory availability was not required, and extended the framework beyond structured EHR data through opportunistically available CT-derived body composition. Together, these analyses demonstrate how heterogeneous information generated during routine care can be progressively integrated into a scalable PDAC risk-enrichment pathway and provide a foundation for future large-scale, multimodal approaches to early PDAC detection.

Beyond enrichment of the underlying clinical information, the ensemble analysis demonstrated that integrating complementary model predictions with longitudinal laboratory changes provided a further, substantial gain in discrimination. XGBoost stacking of predictions from the four base models together with longitudinal percent-change laboratory features increased AUC from 0.84 to 0.88 at 3 months, 0.86 to 0.91 at 6 months, 0.81 to 0.89 at 12 months, and 0.79 to 0.89 at 36 months. The progressively larger absolute gains at longer prediction horizons may be clinically relevant because identifying elevated risk farther from clinical diagnosis could provide a longer interval for subsequent risk assessment and diagnostic evaluation. These results also suggest that complementary information captured by different learning algorithms and explicit longitudinal laboratory changes may not be fully represented by a single model. Nevertheless, whether these improvements in discrimination translate into clinically useful risk thresholds, or improved patient outcomes requires further evaluation.

The complementary Biomarker-Cohort and Full-Cohort analyses address an important challenge in EHR-based prediction: clinical data availability is inherently uneven. Requiring laboratory measurements creates a data-rich population in which the incremental contribution of successive information layers can be evaluated, but it also selects patients who have undergone laboratory testing. The Full-Cohort Analysis therefore retained eligible patients irrespective of laboratory availability and used XGBoost's native handling of missing values. Discrimination remained similar across prediction horizons despite the substantial laboratory missingness in this broader cohort. These findings suggest that the framework can extract useful risk information even when laboratory measurements are unavailable for many patients and support its potential applicability across clinical settings with heterogeneous patterns of testing. At the same time, the two analyses should be viewed as complementary: the Biomarker-Cohort Analysis provides a richer assessment of successive clinical information layers, whereas the Full-Cohort Analysis evaluates model performance without conditioning eligibility on laboratory availability. Together, these analyses support the robustness of the risk-enrichment strategy across differing patterns of laboratory availability.

The imaging enrichment analysis provides further support for the central premise of the layered risk-enrichment framework by demonstrating that predictive information can be incrementally refined as additional data modalities become available. Previous studies have demonstrated loss of skeletal muscle and changes in adipose tissue during the prediagnostic period of PDAC,[18,20] potentially reflecting

evolving metabolic dysfunction and cancer-associated cachexia.[46] In our study, adding CT-derived body-composition measures to the structured EHR feature set was associated with modestly higher discrimination across prediction horizons, with AUC gains of approximately 0.02–0.03. In contrast, body-composition measures combined with demographic characteristics showed substantially lower discrimination, indicating that their contribution was most informative when integrated with the broader clinical profile. These findings extend the layered enrichment concept beyond progressively richer structured EHR information to multimodal data integration: structured EHR data provide a broadly available foundation for risk identification, while imaging-derived phenotypes can further refine risk among patients with existing abdominal CT examinations. Importantly, this approach does not require CT imaging for every patient. Body-composition phenotypes can be extracted opportunistically from clinically acquired CT examinations, potentially providing additional risk enrichment without new imaging or radiation exposure.

A critical implication of this work is the distinction between risk enrichment and population screening. PDAC remains uncommon in the general population, and strong discrimination alone does not ensure adequate positive predictive value when a model is applied indiscriminately at very low disease prevalence. The intended role of this framework is therefore not to diagnose PDAC or trigger invasive testing across the general population, but to progressively concentrate risk. In a potential clinical workflow, structured EHR models could operate passively to identify individuals whose longitudinal clinical profiles warrant further assessment. Opportunistically available imaging-derived phenotypes could provide additional enrichment, and patients exceeding appropriately validated risk thresholds could subsequently undergo more specific evaluation using pancreatic imaging, biomarkers, or specialist assessment. Determining where those thresholds should lie will require assessment of calibration and, where necessary, recalibration in representative external populations, evaluation of false-positive burden, decision-curve analysis, and prospective assessment of downstream clinical consequences.

This study has several strengths. First, the large longitudinal cohort spans more than two decades and uses a temporally separated validation population, providing a more stringent assessment of performance on patients from a later clinical era than random train-test partitioning. Second, the study was designed around a prespecified layered risk-enrichment framework that quantifies the incremental contribution of distinct clinical information domains instead of evaluating only a single fixed predictor set. Third, complementary Biomarker-Cohort and Full-Cohort analyses allowed evaluation of the framework under substantially different patterns of laboratory and vital sign data availability, an important consideration for real-world EHR deployment. Fourth, multiple machine learning algorithms and ensemble strategies were evaluated across four clinically relevant prediction horizons extending to three years before diagnosis. Fifth, model explainability was examined at each horizon, providing insight into how diagnostic, demographic, physiologic, metabolic, hematologic, treatment, and anthropometric information contributed to predicted risk. Finally, opportunistically derived CT body-composition phenotypes were incorporated as an additional enrichment layer, demonstrating how structured EHR risk estimates may be further refined when clinically acquired imaging is already available. Collectively, these design features distinguish the study from approaches centered on a single data domain or prediction

model and more closely reflect how multimodal risk enrichment could operate within routine clinical care.

Study Limitations

Several limitations warrant consideration. First, the study was conducted within a single tertiary healthcare system and the cohorts had limited racial and ethnic diversity, which may restrict generalizability to other populations and healthcare settings. Despite propensity-score matching on age, race, and other characteristics in the parent cohort, differences remained in the final analytic populations, including age and race in the Full-Cohort Analysis. The subsequent restriction defining the Biomarker-Cohort Analysis may have further altered the distribution of matching characteristics; notably, PDAC cases in the 3-month Biomarker-Cohort were older than controls (mean age, 71.9 vs. 66.8 years). The prominent contribution of age in the SHAP analyses, particularly at 36 months, should therefore be interpreted in this context. Second, the retrospective design is susceptible to selection bias, coding variability, changes in clinical practice over time, and informative patterns of healthcare utilization. Laboratory measurements were highly incomplete in the Full-Cohort Analysis, while restricting the Biomarker-Cohort to patients with laboratory data may preferentially select individuals receiving more intensive clinical evaluation. Third, original cohort-construction criteria excluded controls with cirrhosis, prior bariatric surgery, or heart failure because these conditions may independently alter several metabolic, anthropometric, physiologic, and healthcare-utilization features evaluated for PDAC prediction. These exclusions were not applied to the established PDAC case cohort; however, only 126 cases (3.8%) had any of these conditions before diagnosis, and repeating the Full-Cohort Analysis after applying the same exclusions to cases produced essentially unchanged discrimination across prediction horizons. Fourth, diagnostic categories may capture underlying disease processes as well as elements of evolving diagnostic evaluation, particularly at shorter prediction horizons. This concern is relevant to influential features such as screening for suspected conditions, although the persistence of discrimination after removal of PDAC-related and diagnostic-workup CCS categories provides some reassurance. Fifth, SHAP values explain how the fitted models generated predictions but do not establish causal relationships between individual features and PDAC development. Finally, although temporal validation provides a more stringent assessment than random internal splitting, external validation, decision-curve analysis, prospective workflow evaluation, and health economic assessment are necessary before clinical implementation.

Overall, these findings support a shift away from seeking a single universal PDAC screening marker toward developing scalable systems that progressively enrich risk using information already generated during routine clinical care. The complementary Biomarker-Cohort and Full-Cohort analyses demonstrate that meaningful discrimination can be achieved under different levels of clinical data availability, while ensemble learning and opportunistic imaging provide additional opportunities for refinement. Importantly, the explainability findings demonstrate that predicted PDAC risk reflects contributions from multiple clinical domains, with diagnostic history emerging as the most influential source of predictive information. With external validation and prospective evaluation, such a framework could help identify

individuals outside existing hereditary or familial surveillance programs whose evolving clinical profiles warrant targeted evaluation for PDAC.

The original cohort-construction criteria excluded controls with cirrhosis, prior bariatric surgery, or heart failure because these conditions may independently alter several metabolic, anthropometric, physiologic, and healthcare-utilization features evaluated for PDAC prediction. These exclusions were not applied to the established PDAC case cohort; however, only 126 cases (3.8%) had any of these conditions before diagnosis, and repeating the Full-Cohort Analysis after applying the same exclusions to cases produced essentially unchanged discrimination across prediction horizons.

CONCLUSIONS

We developed and temporally validated a layered multimodal risk-enrichment framework for early PDAC prediction using routinely collected clinical data. Progressive incorporation of structured EHR information improved discrimination across prediction horizons from 3 to 36 months before diagnosis, with diagnostic history providing the largest incremental gain and ensemble learning providing further improvement. Opportunistically available CT-derived body composition provided complementary predictive information, supporting a framework in which risk can be refined as additional clinical information becomes available. This approach is intended as a risk-enrichment strategy that leverages data already generated during routine care to identify individuals who may benefit from targeted PDAC evaluation. External validation, prospective workflow evaluation, and assessment of clinical utility and cost-effectiveness are needed before clinical implementation.

Abbreviations

AUC
area under the receiver operating characteristic curve
BMI
body mass index
CCS
Clinical Classifications Software
CI
confidence interval
CT
computed tomography
EHR
electronic health record
F1
F1 score
ICD
International Classification of Diseases

ML
machine learning
NPV
negative predictive value
PDAC
pancreatic ductal adenocarcinoma
PPV
positive predictive value
ROC
receiver operating characteristic
SAT
subcutaneous adipose tissue
SHAP
Shapley Additive Explanations
SMA
skeletal muscle area
VAT
visceral adipose tissue
XGBoost
extreme gradient boosting

Declarations

ACKNOWLEDGEMENTS

Author Contributions: Conceptualization, C.N. and Y.B.; Methodology, C.N.; Data curation, S.P., and A.D.W.; Formal analysis, S.P. and A.D.W.; Writing – original draft, C.N.; Writing – review and editing, All Authors; Supervision, C.N. and Y.B.; Funding acquisition, C.N. and Y.B.

Funding: This effort was supported by the Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery. The study sponsors had no role in the study design; data collection, management, analysis, or interpretation; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Conflicts of Interest: The authors declare no competing interests

Data Availability: The Mayo Clinic EHR data used in this study are not publicly available due to their sensitive nature and institutional restrictions. Access may be granted to qualified researchers through a formal data sharing agreement, subject to institutional review and approval and demonstration of appropriate training in human subject research and data confidentiality. Researchers interested in obtaining access should contact the corresponding author for information on the application process.

Code Availability: The early PDAC identification models were developed using open-source machine learning libraries, including scikit-learn, XGBoost, Optuna, and PyTorch TabNet. Model-development procedures, including data preprocessing, training, temporal validation, hyperparameter optimization and selection, ensemble learning, and the final selected model configurations, are described in the Methods and Supplementary Methods. Mayo Clinic-specific code used for EHR data extraction, cohort construction, preprocessing, feature engineering, body-composition data integration, and mapping of institutional data elements to model inputs cannot be made publicly available because it contains sensitive institutional and operational information, including internal data mappings, data logic and rules, and associated implementation artifacts. Qualified researchers interested in reproducing the analytic methodology may contact the corresponding author regarding available analytic documentation and code that can be shared within institutional, privacy, and data-governance requirements.

References

1. Halbrook CJ, Lyssiotis CA, Di Magliano MP, Maitra A. Pancreatic cancer: Advances and challenges. *Cell*. 2023;186:1729–1754.
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA: a cancer journal for clinicians*. 2023;73:17–48.
3. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer research*. 2014;74:2913–2921.
4. American Cancer S. Cancer Facts & Figs. 2026. In: Atlanta: American Cancer Society; 2026.
5. Siegel RL, Kratzer TB, Wagle NS, Sung H, Jemal A. Cancer statistics, 2026. *Ca*. 2026;76:e70043.
6. Park W, Chawla A, O'Reilly EM. Pancreatic Cancer: A Review. *JAMA*. 2021;326:851–862. doi: 10.1001/jama.2021.13027
7. Singhi AD, Koay EJ, Chari ST, Maitra A. Early detection of pancreatic cancer: opportunities and challenges. *Gastroenterology*. 2019;156:2024–2040.
8. Force UPST, Owens DK, Davidson KW, Krist AH, Barry MJ, Cabana M, Caughey AB, Curry SJ, Doubeni CA, Epling Jr JW. Screening for pancreatic cancer: US preventive services task force reaffirmation recommendation statement. *Jama*. 2019;322:438–444.
9. Goggins M, Overbeek KA, Brand R, Syngal S, Del Chiaro M, Bartsch DK, Bassi C, Carrato A, Farrell J, Fishman EK. Management of patients with increased risk for familial pancreatic cancer: updated recommendations from the International Cancer of the Pancreas Screening (CAPS) Consortium. *Gut*. 2020;69:7–17.
10. Blackford AL, Canto MI, Dbouk M, Hruban RH, Katona BW, Chak A, Brand RE, Syngal S, Farrell J, Kastrinos F, et al. Pancreatic Cancer Surveillance and Survival of High-Risk Individuals. *JAMA Oncology*. 2024;10:1087–1096. doi: 10.1001/jamaoncol.2024.1930
11. Lucas AL, Kastrinos F. Screening for Pancreatic Cancer. *JAMA*. 2019;322:407–408. doi: 10.1001/jama.2019.9690

12. Mukund A, Afridi MA, Karolak A, Park MA, Permuth JB, Rasool G. Pancreatic ductal adenocarcinoma (PDAC): a review of recent advancements enabled by artificial intelligence. *Cancers*. 2024;16:2240.
13. Pannala R, Basu A, Petersen GM, Chari ST. New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer. *The lancet oncology*. 2009;10:88–95.
14. Chari ST, Leibson CL, Rabe KG, Timmons LJ, Ransom J, De Andrade M, Petersen GM. Pancreatic cancer–associated diabetes mellitus: prevalence and temporal association with diagnosis of cancer. *Gastroenterology*. 2008;134:95–101.
15. Sharma A, Kandlakunta H, Nagpal SJS, Feng Z, Hoos W, Petersen GM, Chari ST. Model to determine risk of pancreatic cancer in patients with new-onset diabetes. *Gastroenterology*. 2018;155:730–739. e733.
16. Yuan C, Bao Y, Wu C, Kraft P, Ogino S, Ng K, Qian ZR, Rubinson DA, Stampfer MJ, Giovannucci EL. Prediagnostic body mass index and pancreatic cancer survival. *Journal of Clinical Oncology*. 2013;31:4229–4234.
17. Hoyt M, Song Y, Gao S, O'Palka J, Zhang J. Prediagnostic BMI trajectories in relation to pancreatic cancer risk in the prostate, lung, colorectal, and ovarian cancer screening trial. *Obesity*. 2022;30:2275–2285.
18. Klatte DC, Weston A, Ma Y, Sledge H, Bali A, Bolan C, Engels M, van Hooft JE, van Leerdam ME, Ouni A. Temporal trends in body composition and metabolic markers prior to diagnosis of pancreatic ductal adenocarcinoma. *Clinical Gastroenterology and Hepatology*. 2024;22:1830–1838. e1839.
19. Fahrmann JF, Schmidt CM, Mao X, Irajizad E, Loftus M, Zhang J, Patel N, Vykoukal J, Dennison JB, Long JP. Lead-time trajectory of CA19-9 as an anchor marker for pancreatic cancer early detection. *Gastroenterology*. 2021;160:1373–1383. e1376.
20. Bogdanski AM, Klatte DCF, van Hooft JE, Patel S, van Leerdam ME, Ngufor CG, Kuanar SP, Weston AD, Wallace MB, Bi Y. Longitudinal Body Composition Changes Distinguish Pancreatic Ductal Adenocarcinoma From Controls in a Pre-Diagnostic CT Cohort. *Pancreatology*. 2026. doi: https://doi.org/10.1016/j.pan.2026.09.008
21. Placido D, Yuan B, Hjaltelin JX, Zheng C, Haue AD, Chmura PJ, Yuan C, Kim J, Umeton R, Antell G. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature medicine*. 2023;29:1113–1122.
22. Jia K, Kundrot S, Palchuk MB, Warnick J, Haapala K, Kaplan ID, Rinard M, Appelbaum L. A pancreatic cancer risk prediction model (Prism) developed and validated on large-scale US clinical data. *EBioMedicine*. 2023;98.
23. Zhu W, Chen L, Aphinyanaphongs Y, Kastrinos F, Simeone DM, Pochapin M, Stender C, Razavian N, Gonda TA. Identification of patients at risk for pancreatic cancer in a 3-year timeframe based on machine learning algorithms. *Scientific Reports*. 2025;15:11697. doi: 10.1038/s41598-025-89607-8
24. Khristenko E, Sinitsyn V, Rieden T, Girod P, Kauczor H-U, Mayer P, Klauss M, Lyadov V. CT-based screening of sarcopenia and its role in cachexia syndrome in pancreatic cancer. *Plos one*. 2024;19:e0291185.

25. Bundred J, Kamarajah SK, Roberts KJ. Body composition assessment and sarcopenia in patients with pancreatic cancer: a systematic review and meta-analysis. *HPB*. 2019;21:1603–1612.
26. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *The lancet*. 2007;370:1453–1457.
27. Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, Van Smeden M. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*. 2024;385.
28. Periyalwar P, Dasarathy S. Malnutrition in cirrhosis: contribution and consequences of sarcopenia on metabolic and clinical responses. *Clin Liver Dis*. 2012;16:95–131. doi: 10.1016/j.cld.2011.12.009
29. To VT, Hüttl TP, Lang R, Piotrowski K, Parhofer KG. Changes in body weight, glucose homeostasis, lipid profiles, and metabolic syndrome after restrictive bariatric surgery. *Exp Clin Endocrinol Diabetes*. 2012;120:547–552. doi: 10.1055/s-0032-1323738
30. Vest AR, Chan M, Deswal A, Givertz MM, Lekavich C, Lennie T, Litwin SE, Parsly L, Rodgers JE, Rich MW. Nutrition, obesity, and cachexia in patients with heart failure: a consensus statement from the Heart Failure Society of America Scientific Statements Committee. *Journal of cardiac failure*. 2019;25:380–400.
31. Agency for Healthcare R, Quality. Clinical Classifications Software (CCS) for ICD-10-PCS (beta version). 2025.
32. Weston AD, Sauver JLS, Moerer RT, Grossardt BR, Kline TL, Chinthalapelly PR, Garner HW, Erickson BJ, Rule AD. Volumetric assessment from L2 to L4 improves the reliability and validity of body composition by CT. *European Journal of Radiology*. 2026:113076.
33. Weston AD, Korfiatis P, Kline TL, Philbrick KA, Kostandy P, Sakinis T, Sugimoto M, Takahashi N, Erickson BJ. Automated Abdominal Segmentation of CT Scans for Body Composition Analysis Using Deep Learning. *Radiology*. 2019;290:669–679. doi: 10.1148/radiol.2018181432
34. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28:112–118.
35. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2005;67:301–320.
36. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 1996;58:267–288.
37. Breiman L. Random forests. *Machine learning*. 2001;45:5–32.
38. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper/Poster presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016;
39. Arik SÖ, Pfister T. Tabnet: Attentive interpretable tabular learning. Paper/Poster presented at: Proceedings of the AAAI conference on artificial intelligence; 2021;

40. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. Paper/Poster presented at: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining; 2019;
41. Breiman L. Stacked regressions. *Machine learning*. 1996;24:49–64.
42. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*. 2022;214:106584.
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.
44. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. 2020;17:261–272. doi: 10.1038/s41592-019-0686-2
45. Mishra AK, Chong B, Arunachalam SP, Oberg AL, Majumder S. Machine Learning Models for Pancreatic Cancer Risk Prediction Using Electronic Health Record Data-A Systematic Review and Assessment. *Am J Gastroenterol*. 2024;119:1466–1482. doi: 10.14309/ajg.0000000000002870
46. Yu SY, Luan Y, Dong R, Abazarikia A, Kim SY. Adipose Tissue Wasting as a Determinant of Pancreatic Cancer-Related Cachexia. *Cancers (Basel)*. 2022;14. doi: 10.3390/cancers14194754

Figures

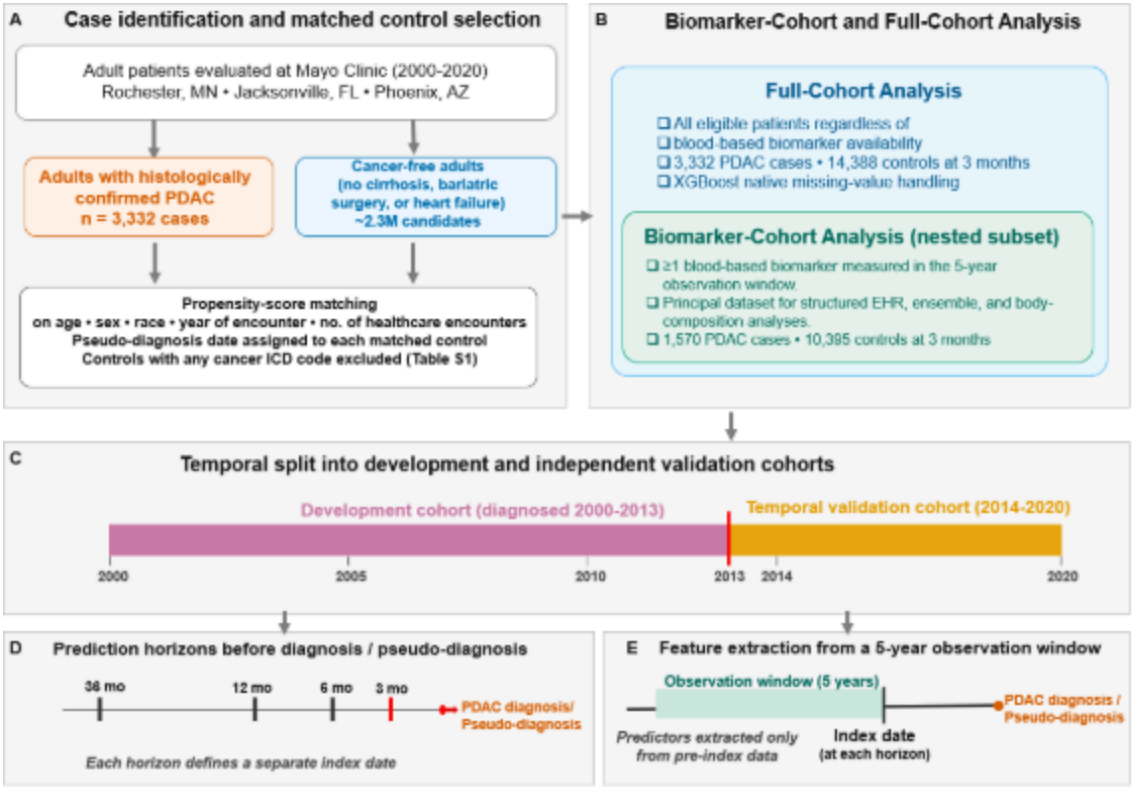


Figure 1

Study Design and Cohort Construction. *Schematic illustrating the study design, including: (A) case identification and propensity-score-matched control selection; (B) definition of the Biomarker-Cohort and Full-Cohort analyses; (C) temporal data split; (D) prediction horizons; and (E) feature extraction from the five-year window preceding the index date.*

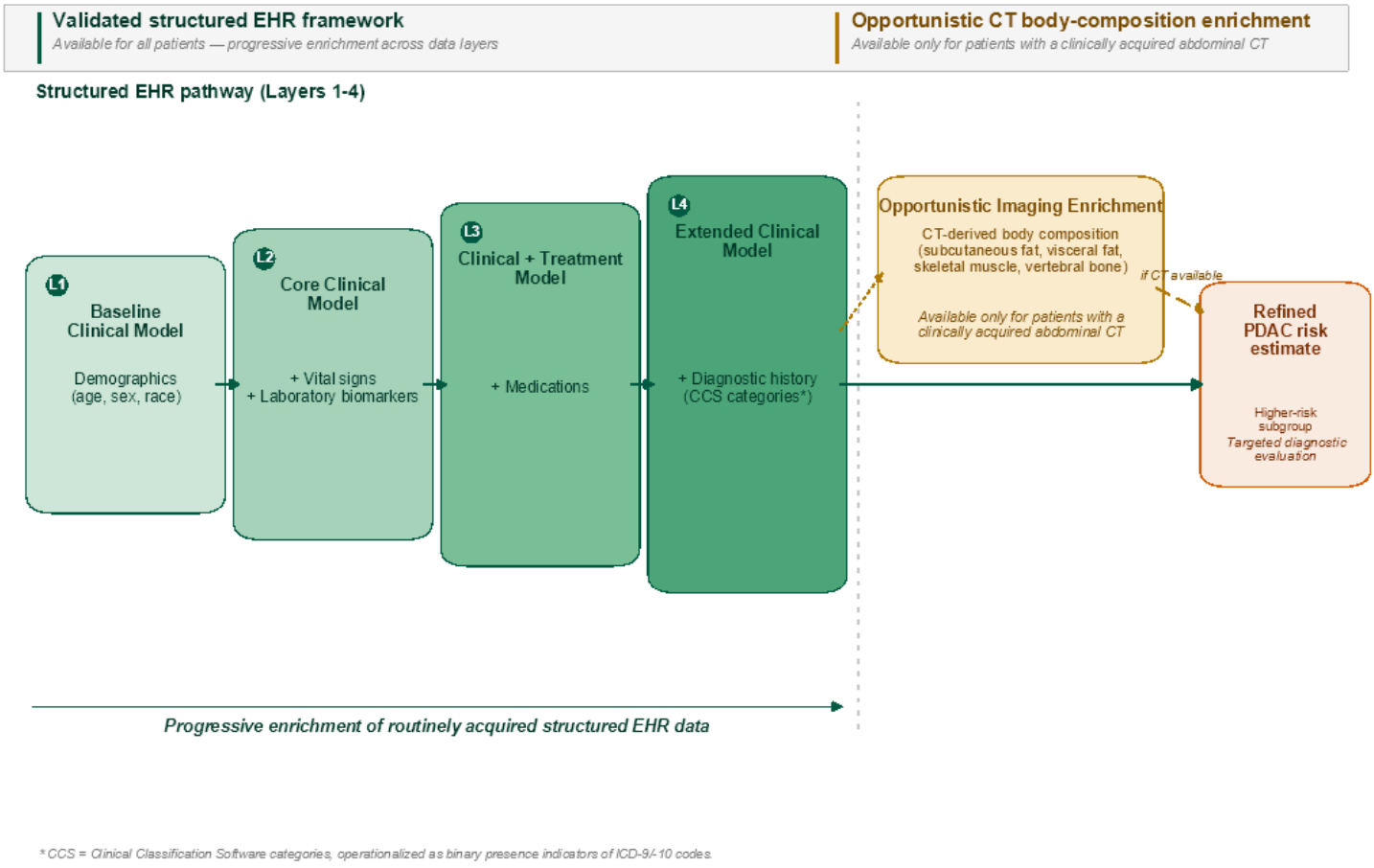


Figure 2

Layered Multimodal Risk-Enrichment Framework. *Conceptual illustration of the layered risk-enrichment framework - showing how progressively richer sources of routinely acquired healthcare data may be leveraged to refine PDAC risk estimates and identify individuals who may benefit from targeted diagnostic evaluation.*

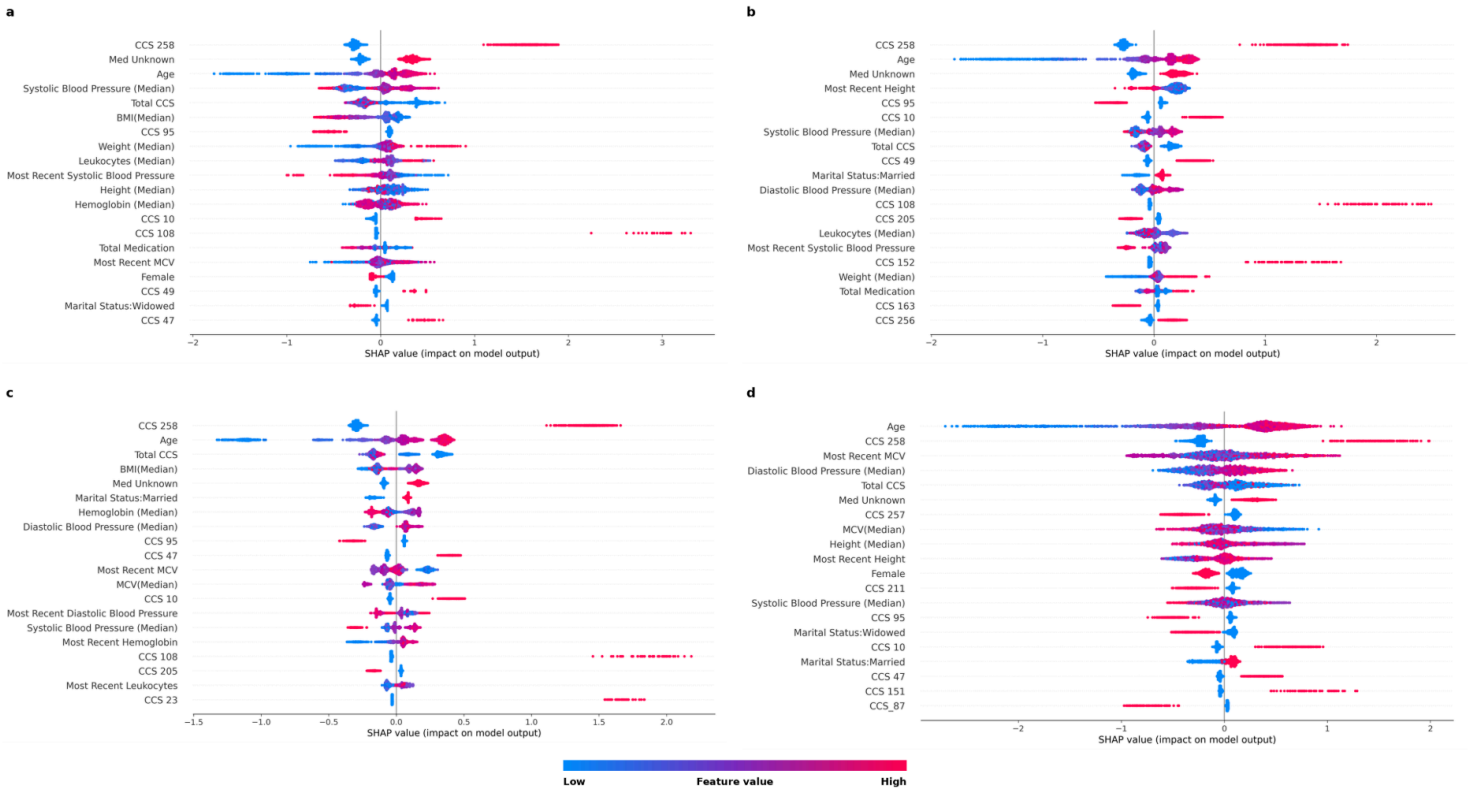


Figure 3

SHAP Summary Plots for the Extended Clinical Model Across Prediction Horizons. *Top 20 features ranked by mean absolute SHAP value for the XGBoost Extended Clinical Model at the (a) 3-month, (b) 6-month, (c) 12-month, and (d) 36-month prediction horizons in the Biomarker-Cohort Analysis. Each point represents an individual patient. SHAP values on the x-axis indicate the contribution of a feature to the model prediction, with positive values shifting the prediction toward higher PDAC risk and negative values shifting it toward lower PDAC risk. Point color represents the underlying feature value, with red indicating higher values and blue indicating lower values. For binary features, color reflects feature presence or absence.*

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- EarlyPDACDetectionsupplement.docx