MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories

Placido, D., Yuan, B., Hjaltelin, J.X. et al.

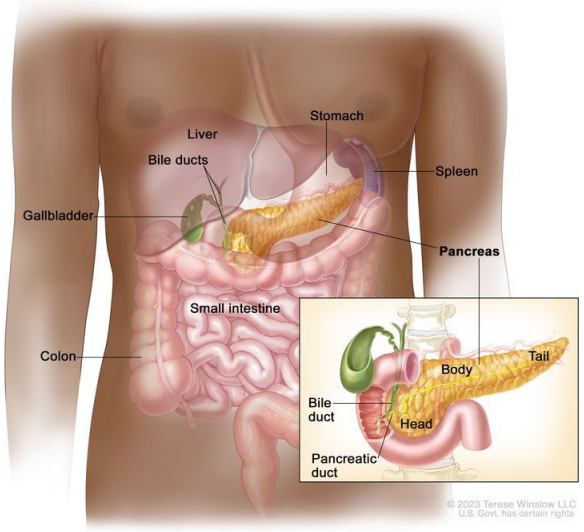
Hongyi Yang | EECS 598-007 FA23

Outline

- **Research Background**
- Methods
- Experiments
- Results
- Discussion

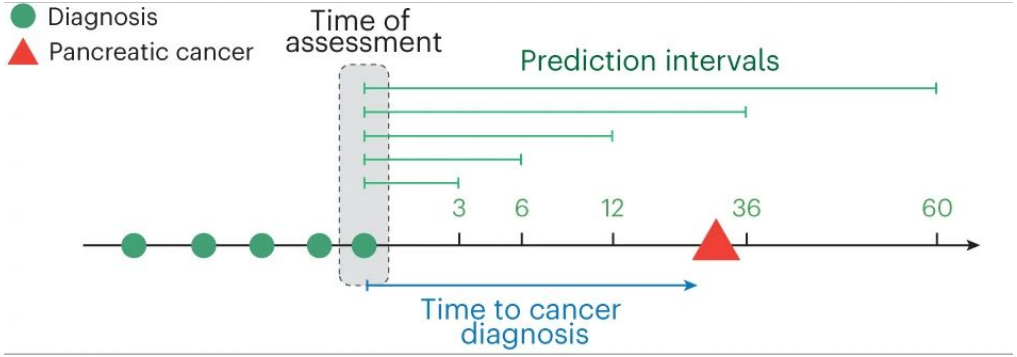
Research Background

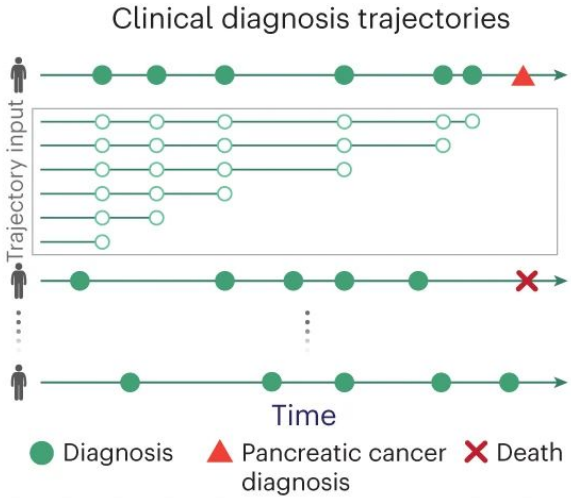
- Pancreatic cancer
  - 4-th leading cause of cancer-related deaths
  - Hard to detect
    - ~80% detected at a late stage, <10% survivals in 5 years
    - lack of high-penetrance risk factors
    - low incidence rate
- Risk assessment methods
  - Traditional: family history, behavioral and clinical risk factors
    - subjective and inaccurate
  - Bioinformatics: circulating biomarkers and genetic predisposition
    - expensive and limited
  - This paper: clinical records
    - affordable and general



Research Background

- Disease trajectories
  - Longitudinal clinical records (list)
    - Diagnosis (categorical)
    - Date (timestamp)
  - Case (Positive)
    - Patients with pancreatic cancer
    - Assess at the last diagnosis before pancreatic cancer
  - Control (Negative)
    - Patients without pancreatic cancer
    - Assess at 2 years previous to the end of data



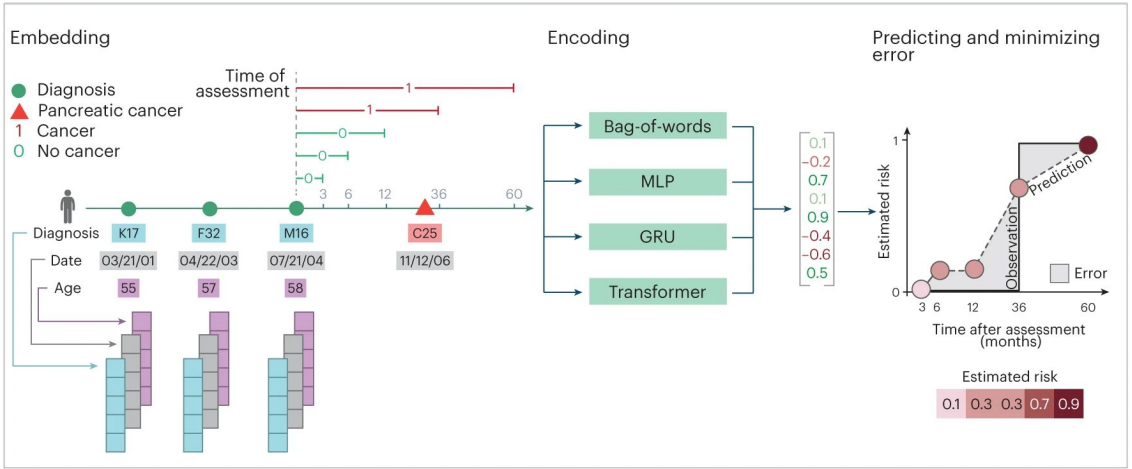


Outline

- Research Background
- **Methods**
- Experiments
- Results
- Discussion

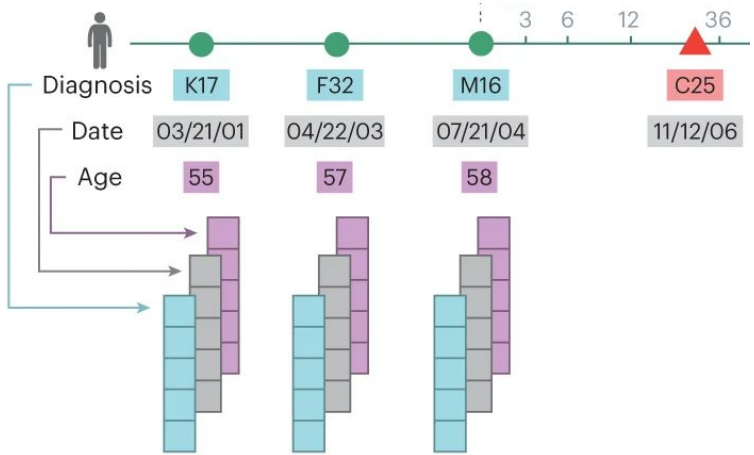
Methods - Overview

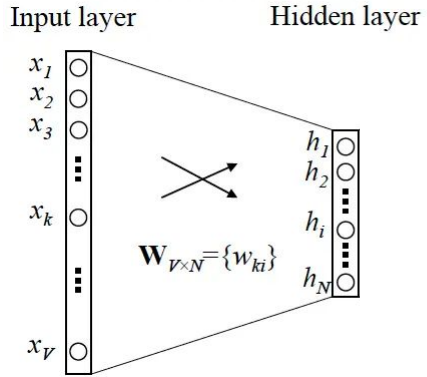
- Embedding
  - Dimension reduction of categorical disease codes and timestamps
- Encoding
  - Feature extraction of longitudinal embedded sequences
- Predicting
  - Predict the risk of pancreatic cancer for five different prediction intervals



Methods - Embedding

- Diagnosis
  - ICD-10 level-3 codes, e.g.
    - F32=Depressive episode
    - C25=Malignant neoplasm of pancreas
  - Different settings
    - one-hot
    - embedding layer





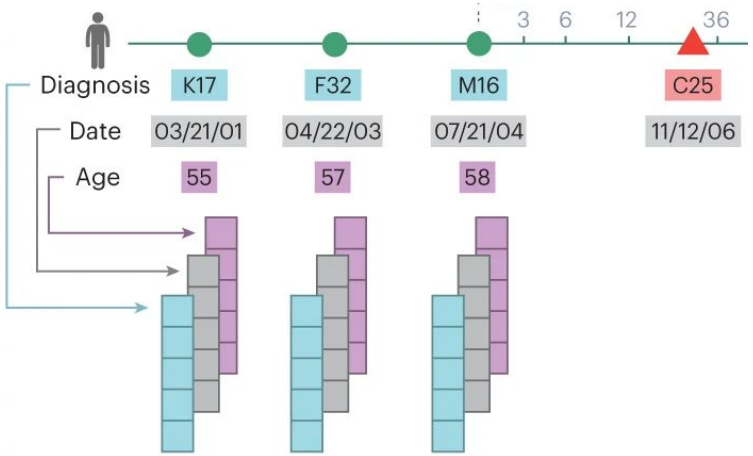
Methods - Embedding

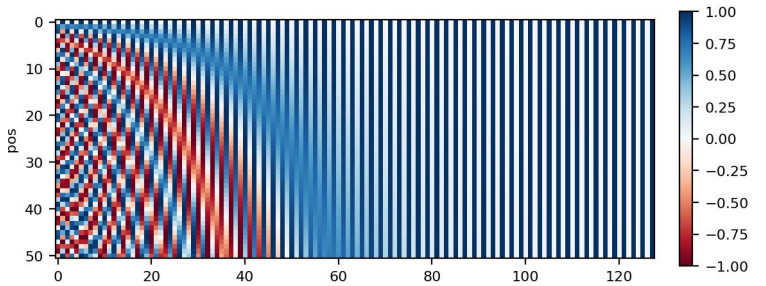
- Date & Age
  - Different settings
    - not embedded
    - append age neuron to the diagnosis vector
    - positional encoding
      - Depth: d = 128
      - Date: pos = time interval to last record
      - Age: pos = age at diagnosis
      - PE further embeded
        - PE_scale
        - PE_add

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

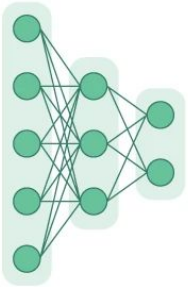
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

- Embedding = Diagnosis * PE_scale + PE_add

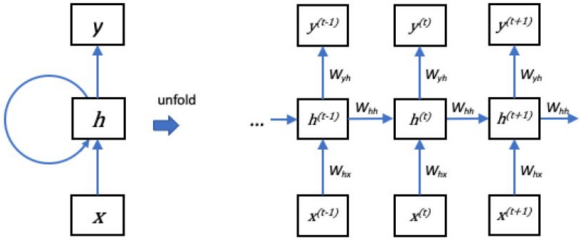


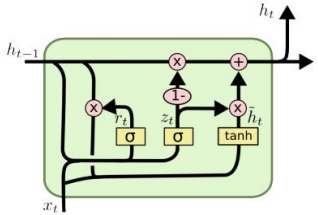


Methods - Encoding



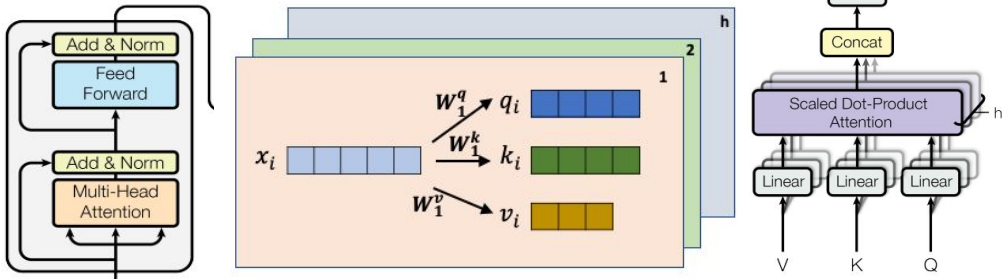
- Non-time-sequential models: date & age not embedded
  - Bag-of-words (pooling layers)
  - Multilayer Perceptron (MLP)
- Gated Recurrent Unit (GRU): sequential layers





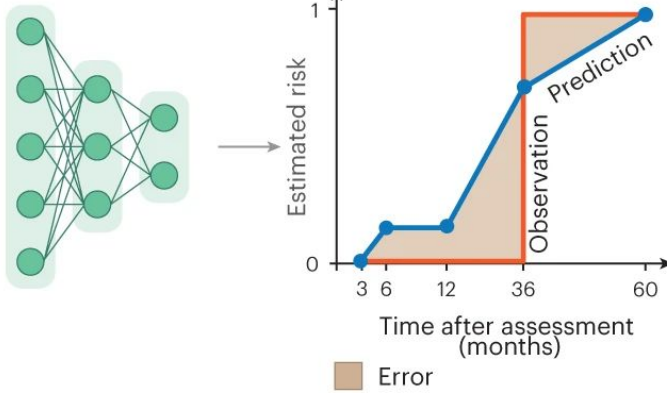
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

- Transformer: attention layers



Methods - Predicting

- Fully connected feedforward network
  - Predicts risks for five prediction intervals
    - 3, 6, 12, 36, 60 months
    - 36 months: reasonable in clinical practice
  - L2-regularized cross-entropy loss
    - average across valid predictions
    - use only meaningful timepoints
      - case: before cancer and the first include cancer
      - control: before death/end of data



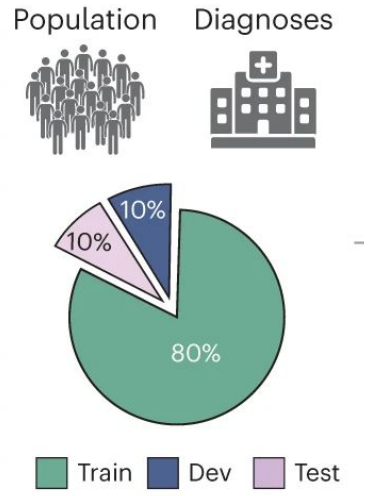
$$\text{loss}\ \frac{1}{N}\frac{1}{N_T}\sum_{i,t} -\left[y_{i,t}\log\left[\hat{p}_{\Theta,t}(x_i)\right] + (1-y_{i,t})\log\left[1-\hat{p}_{\Theta,t}(x_i)\right]\right] + \lambda_2\|\Theta\|_2$$

Outline

- Research Background
- Methods
- **Experiments**
- Results
- Discussion

Experiments - Data

- Datasets
  - DK: entire population in Denmark (Main)
    - event rate: 0.4% (24K/6.2M)
  - US-VA: military veterans in the US (External, shorter but more dense)
    - event rate: 0.2% (3.4K/1.9M)
- Data splits
  - 80% Training
    - train models
    - resample during training to address the inbalance
  - 10% Development (Validation)
    - select hyperparameters
  - 10% Test
    - evaluate models



Experiments - Evaluation

- Select best hyperparameters on Development set with
  - Area under the precision recall curve (AUPRC)
- Evaluate selected models on Test set with
  - Area under the receiver operating characteristic (AUROC)
  - Relative risk ratio (RR): threshold = top 0.1% (1K/1M)

$$RR = \frac{precision}{incidence} = \frac{TP/(TP+FP)}{(TP+FN)/(TP+FP+TN+FN)}$$

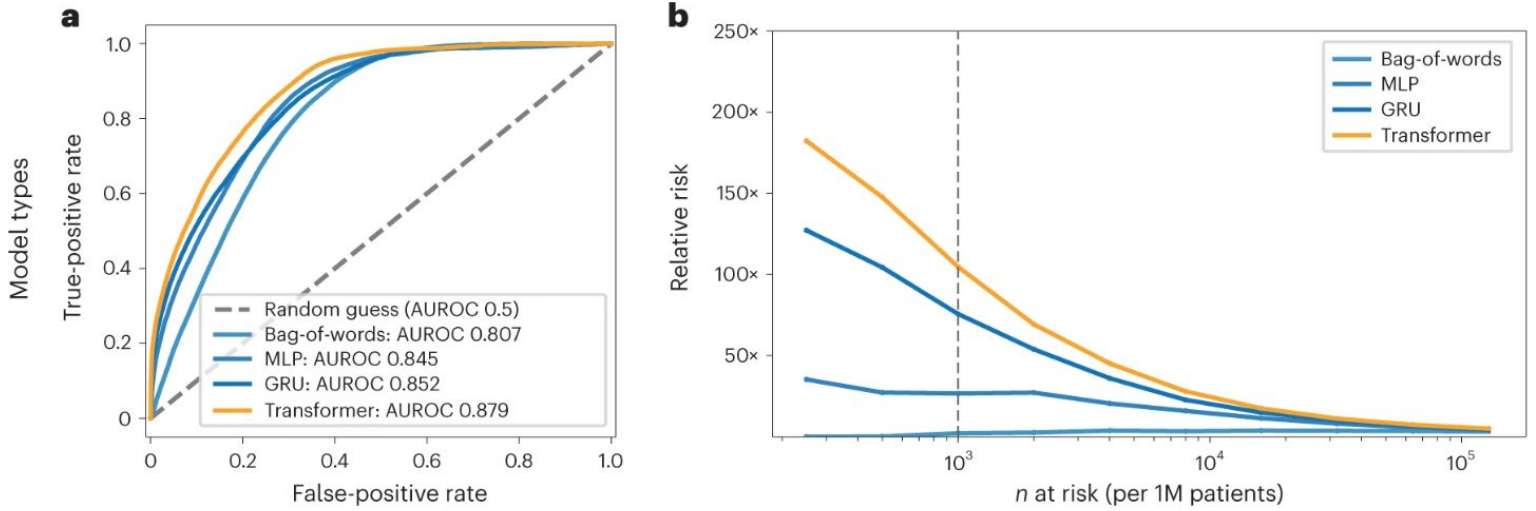
For a prediction interval, PI		Predicted	
		Positive (Risk > Threshold)	Negative (Risk < Threshold)
Actual	Positive (cancer within PI)	TP	FN
	Negative (no cancer within PI)	FP	TN

Outline

- Research Background
- Methods
- Experiments
- **Results**
- Discussion

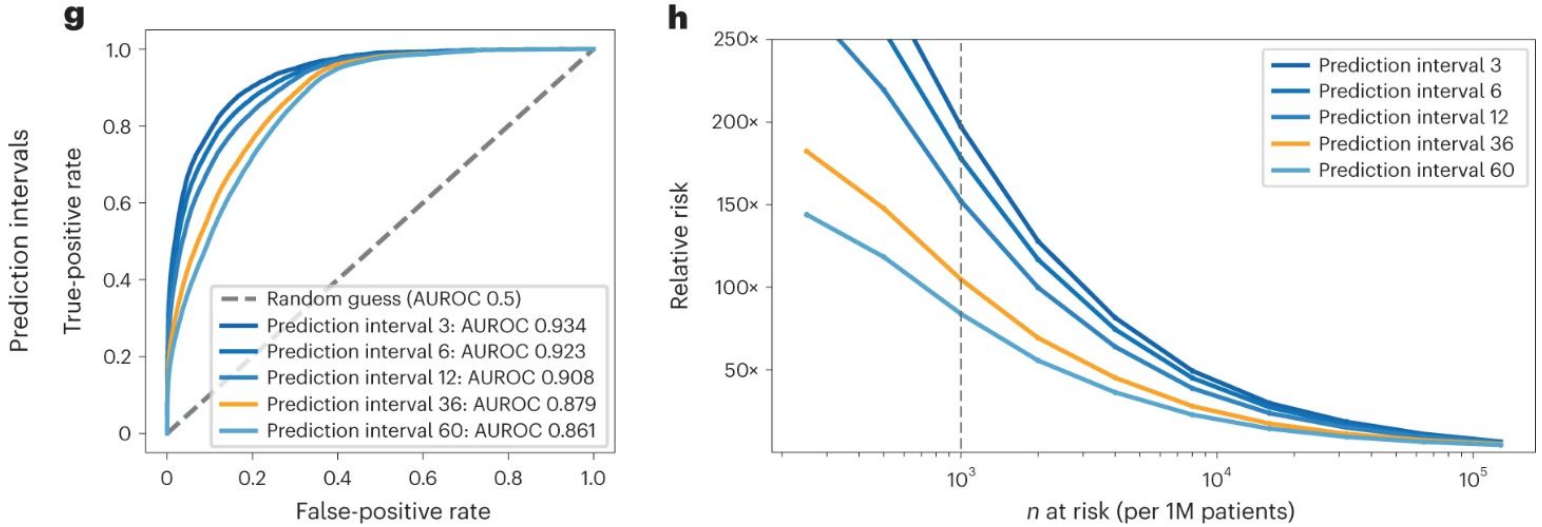
Results - Compare Models

- Prediction interval = 36 months
- Time information help a lot with prediction



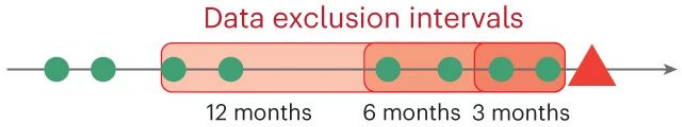
Results - Compare Prediction Intervals

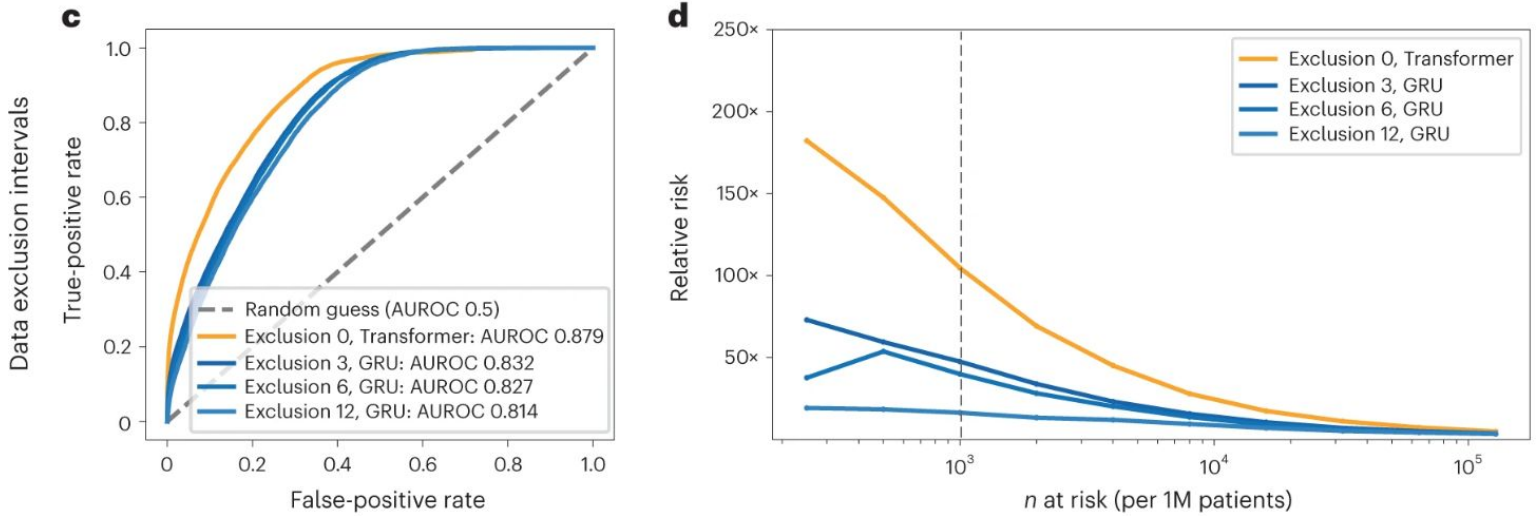
- Prediction is more challenging for longer intervals



Results - Effect of Data Exclusion

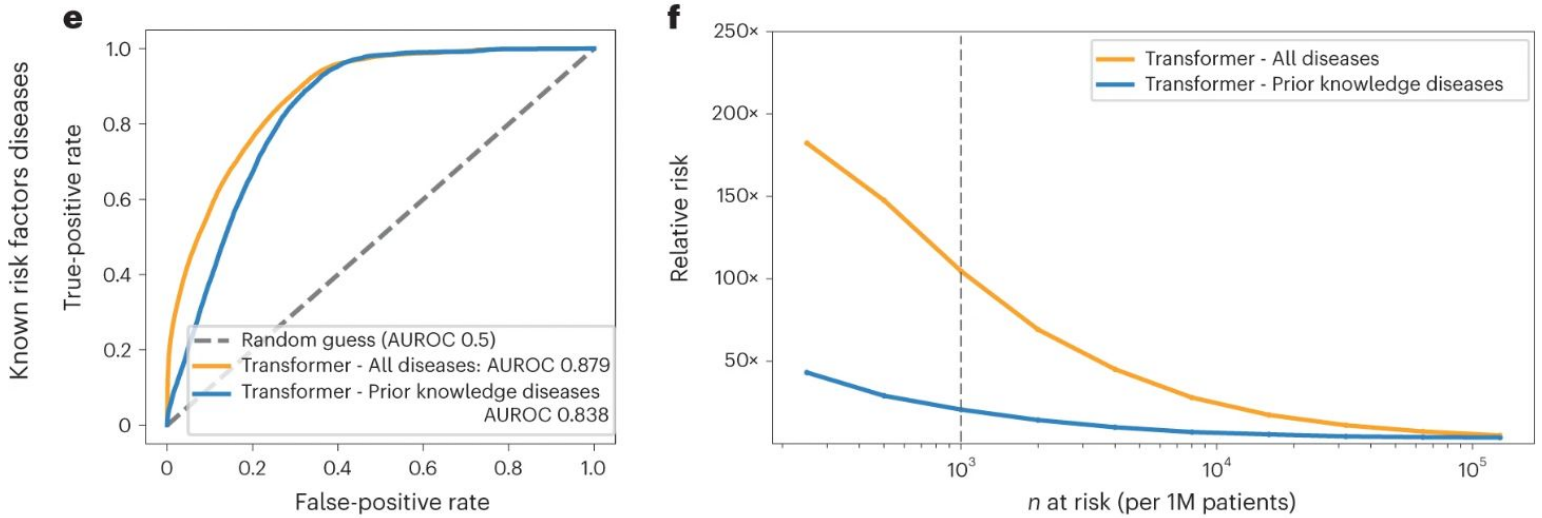
- Prediction interval = 36 months
- Close-to-cancer diagnoses play an important role in prediction





Results - Predictive Features

- Prediction interval = 36 months
- Known risk factors might not cover all the important signs of pancreatic cancer



Results - Risk Factors

- Ranked by integrated gradients
- Most are known risk factors in clinical pratice

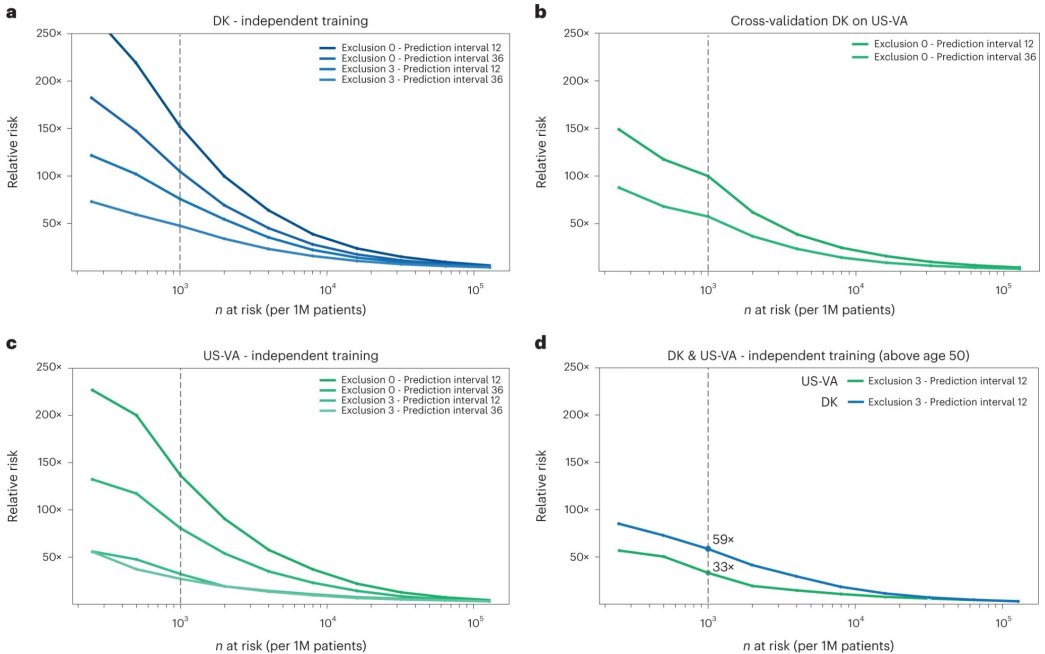
Feature contributions - No exclusion (DK)

	Cancer in 0–6 months	Cancer in 6–12 months	Cancer in 12–24 months	Cancer in 24–36 months
1	Unspecified jaundice	Other diseases of biliary tract	Medical observation and evaluation for suspected diseases and conditions	Medical observation and evaluation for suspected diseases and conditions
2	Medical observation and evaluation for suspected diseases and conditions	Unspecified jaundice	Other diseases of biliary tract	Other diseases of pancreas
3	Other diseases of biliary tract	Medical observation and evaluation for suspected diseases and conditions	Other diseases of pancreas	Other diseases of biliary tract
4	Abdominal and pelvic pain	Other diseases of pancreas	Abdominal and pelvic pain	Non insulin-dependent diabetes mellitus
5	Malignant neoplasm of other and unspecified parts of biliary tract	Malignant neoplasm of other and unspecified parts of biliary tract	Non insulin-dependent diabetes mellitus	Unspecified jaundice
6	Other diseases of pancreas	Abdominal and pelvic pain	Malignant neoplasm of other and unspecified parts of biliary tract	Abdominal and pelvic pain
7	Secondary malignant neoplasm of respiratory and digestive organs	Secondary malignant neoplasm of respiratory and digestive organs	Unspecified jaundice	Malignant neoplasm of other and unspecified parts of biliary tract
8	Symptoms and signs concerning food and fluid intake	Non insulin-dependent diabetes mellitus	Other functional intestinal disorders	Gastritis and duodenitis
9	Non insulin-dependent diabetes mellitus	Malignant neopleasm without specification of site	Diseases of pancreas	Insulin-dependent diabetes mellitus
10	Other anaemias	Other anaemias	Secondary malignant neoplasm of respiratory and digestive organs	Other anaemias

Results - Datasets

- Need independent model training for optimal performance

Train	Test	AUROC	RR
DK	DK	0.879	104.6
DK	US-VA	0.710	57.4
US-VA	US-VA	0.775	80.4



Outline

- Research Background
- Methods
- Experiments
- Results
- **Discussion**

Discussion

- Contribution
  - Propose a deep-learning based framework for early detection of pancreatic cancer
  - Utilize longitudinal clinical records for general use
  - Verify the applicability of different models
- Weakness
  - Lack of originality
  - Unclear methods description
- Future Works
  - Incorporate multimodal clinical data
  - Embed the system in wearable devices
  - Apply in real-world clinical scenarios

Quiz

- When and why do we need the positional encodings of date and age?
- What could be the reason that the authors use AUPRC for model selection while use AUROC and RR for model evaluation?

Thanks