World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:14, No:11, 2020

A Deep-Learning Based Prediction of Pancreatic Adenocarcinoma with Electronic Health Records from the State of Maine

Xiaodong Li, Peng Gao, Chao-Jung Huang, Shiying Hao, Xuefeng B. Ling,Yongxia Han, Yaqi Zhang, Le Zheng, Chengyin Ye, Modi Liu, Minjie Xia, Changlin Fu, Bo Jin, Karl G. Sylvester, Eric Widen

Open Science Index, Medical and Health Sciences Vol:14, No:11, 2020 publications.waset.org/10011557.pdf

***Abstract*—**Predicting the risk of Pancreatic Adenocarcinoma (PA) in advance can benefit the quality of care and potentially reduce population mortality and morbidity. The aim of this study was to develop and prospectively validate a risk prediction model to identify patients at risk of new incident PA as early as 3 months before the onset of PA in a statewide, general population in Maine. The PA prediction model was developed using Deep Neural Networks, a deep learning algorithm, with a 2-year electronic-health-record (EHR) cohort. Prospective results showed that our model identified 54.35% of all inpatient episodes of PA, and 91.20% of all PA that required subsequent chemoradiotherapy, with a lead-time of up to 3 months and a true alert of 67.62%. The risk assessment tool has attained an improved discriminative ability. It can be immediately deployed to the health system to provide automatic early warnings to adults at risk of PA. It has potential to identify personalized risk factors to facilitate customized PA interventions.

***Keywords*—**Cancer prediction, deep learning, electronic health records, pancreatic adenocarcinoma.

I. Introduction

PA holds the top attention both of men and women since our incapability to track down early-stage illness. Especially, PA was the third most common cause of cancer-related deaths in the United States [1] and the seventh leading cause of global cancer deaths in the worldwide [2]. Every year, lots of people die of "silent killers", called PA which are hard to be identified and treated. Early diagnosis is crucial to its successful treatment before the deterioration of diseases. Medical disease diagnosis using Artificial Neural Networks (ANN) [3] is currently a highly active research field in medicine system and it will be more widely used in biomedical systems for decades to come. Researchers have searched for effective biomarkers [4]-[6] which can aid in early diagnosis and have developed models to support disease risk prediction.

Various epidemiologic and clinical characteristics are related with occurrence of PA, including family history of PA, inherited genetic, variation/influence, anthropometric variables [e.g., Body Mass Index (BMI)] [7], [17]. They are associated with medical comorbidities (e.g., pancreatitis, diabetes) [8], [21]. The age-adjusted cancer-related death rate is increasing for PA, and it is predicted that PA will be the second most significant key-factor of cancer-related deaths by 2030 [9]. Furthermore, classification of high-risky individuals for PA or with early-stage disease was complicated because of the shortage of dependable screening instruments [10]. Cai et al. [11] studied a PA risk stratification prediction method by choosing 138 patients with chronic pancreatitis. A scoring algorithm based logistic regression was used to study the prediction. Further, the method to test the changes of precancerous in the pancreas among high danger individuals by the use of endoscopic ultrasound (EUS), computed tomography (CT) scan, doppler ultrasound (US), magnetic resonance imaging (MRI), or positron emission tomography (PET) has also been validated in several clinical studies [12], [29]. Chari et al. [13] proved that the 3-years cumulative incidence of PA among patients with new onset diabetes is as higher as 8 times than expected. Gold et al. [14] confirmed the potential role of PAM4 in discovering early stage pancreatic cancer, which was uncovered in precursor lesions of PA. Lately, numerous studies had been concentrated early detection of PA through the identification and validation of promising biomarkers [15], [16]. To our knowledge, no established screening method has been presented for sporadic PA. The non-invasive precursor lesions named pancreatic intraepithelial neoplasia (PanIN) progress from PanIN1 to PanIN3 and into PA within an undefined timeline [18]. A scoring method based logistic regression was used to develop the prediction rule. Hsieh et al. [19] predicted PA in the patients with type 2 diabetes using logistic regression and ANN models. Klein et al. [20] developed a relative risk model for men and women of European ancestry based on non-genetic and genetic risk factors for pancreatic cancer [20]. Lucenteforte et al. [22] focused on lifestyle to predict PA.

X.D. Li is with the Computer Science Department, University of Hangzhou Dianzi University, Hangzhou, 310018 China and also with Department of Surgery, Stanford University, Stanford, CA, United States (e-mail: hzxiaodong22@163.com).

P. Gao is with Institute of Pharmacy, Shandong University of Chinese Medicine, Shandong, China (e-mail: 48395504@qq.com).

C.J. Huang is with National Taiwan University-Stanford Joint Program Office of AI in Biotechnology, Ministry of Science and Technology Joint Research Center for Artificial Intelligence Technology and All Vista Healthcare, Taipei, Taiwan (e-mail: cjhuang0717@gmail.com).

S.Y. Hao is with Department of Surgery, Stanford University, Stanford, CA, USA and also with Department of Cardiothoracic Surgery, Stanford University, Stanford, CA, USA (e-mail: shiyingh@stanford.edu).

B. Ling is with the Department of Surgery, Stanford University, Stanford, CA, USA and also with Clinical and Translational Research Program, Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Palo Alto, CA, USA (corresponding author, e-mail: bxling@stanford.edu).

Y.X. Han and Y.Q. Zhang are with the School of Electrical Power Engineering, South China University of Technology, Guangzhou, China (e-mail: epyxhan@scut.edu.cn, 253648628@qq.com).

Open Science Index, Medical and Health Sciences Vol:14, No:11, 2020 publications.waset.org/10011557.pdf

Masahiro et al. [23] presented a PA risk prediction model in the general population in Japan with AUC of 0.63 (95% confidence interval, 0.60-0.66) or 0.61 (0.58-0.64) ,which was on the basis of data including directly determined or imputed SNP genotypes for 664 pancreatic cancer case and 664 age- and sex-matched control subjects. While an ANN model performed considerably well to predict PA on the basis of commonly available data in the EHR, inclusion of personal high-risk features for PA (e.g., pancreatic cysts, family history etc.) could potentially improve the performance of the model. Pannala et al. [24] proved that diabetes appeared to be associated with early stage PA. It is judged that symptoms manifest about 6 months after PA gets unresectable. Therefore, identifying those at high risk yet asymptomatic is very significant to test PA while it is still resectable. Hence, it can be shown that diabetes associated with PA may be a paraneoplastic phenomenon caused by the cancer [25], [26]. Permuth-Wey et al. [27] studied the quantified familial risks of PA, and through a meta-analysis, obtained more accurate estimates of familial risk. In another study, an ANN model was created to test PA based on a data set of symptoms [28]. A total sample of 120 patients (i.e., 90 training samples and 30 test samples) with 11 possible symptoms and 3 outputs were chosen for this model [29]. In another method, Wang et al. [30] predicted familial PA risk through a Mendelian algorithm (i.e., PancPRO) that was built by extending the Bayesian modeling framework.

In this study, we aimed to develop an EHR-based risk assessment model to forecast patients' PA risk as early as 3 months before the onset of PA in a statewide. By using the EHR data from the population in the State of Maine, U.S., and the deep learning algorithm, we believed that the deep model could uncover the underlying clinical and pathophysiological patterns/interactions of impactful predictors, and eventually obtain a higher accuracy.

II. Method

A. Dataset

The study cohort was formed by patients with age of 35 years and older that visited Maine health care facilities, including 35 hospitals, 34 federally qualified health centers, from January 1, 2014 to March 31, 2018. This retrospective dataset was a subset of the health information exchange (HIE) network and was authorized by the HealthInfoNet organization after the de-identification process. The personal information was removed during the analysis and publication procedure. This study was exempted from ethics review by the Stanford University institutional review board. The inclusion and exclusion criteria were summarized in Fig. 1. A total of 265,225 individuals were recruited in this study.

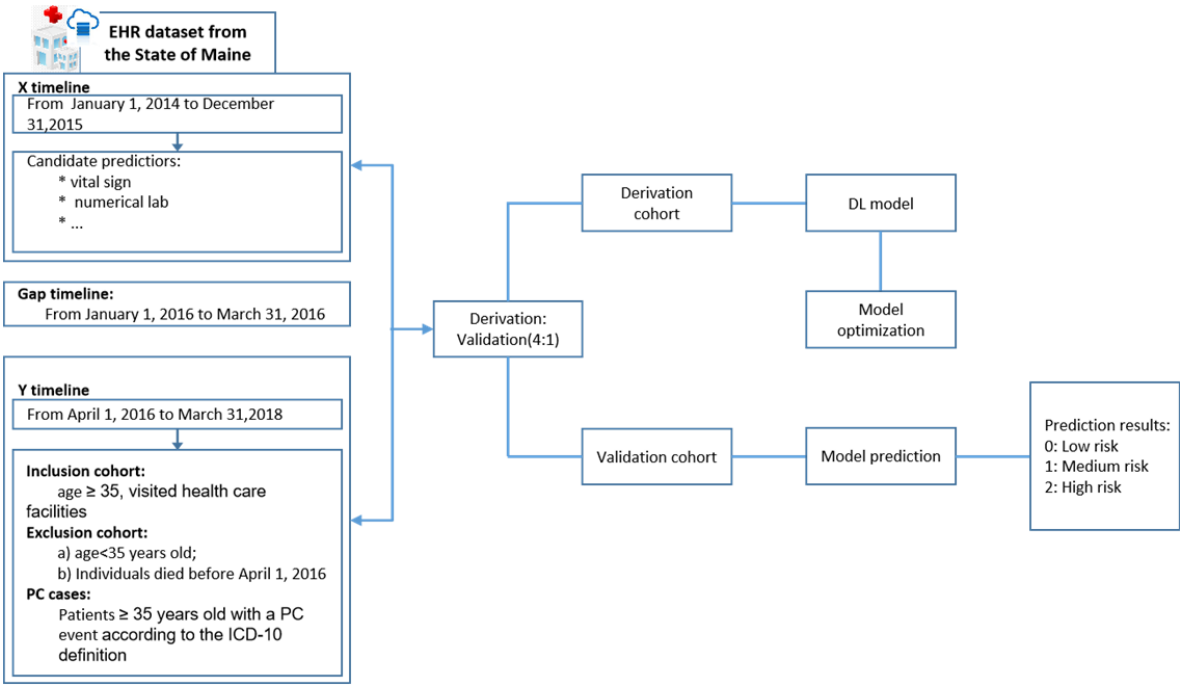

Fig. 1 Study design. The study cohort was derived from EHRs of patients with age of 35 years and older that visited Maine health care facilities and divided into the derivation and validation cohorts for model development and evaluation

B. Prediction Task

Fig. 2 illustrates the prediction task. The task is to predict the onset of PA diseases prior to diagnosis time. We use EHR information accumulated in the history window to predict diagnoses in the prediction window. We also add a gap period between the end of the history window and the start of the prediction window. The study is to prohibit the model from counterfeit data generated right before the diagnosis time. Specifically, our model selected a 12-month historical window, a 3-month gap and a 24-month predictive window in

this study.

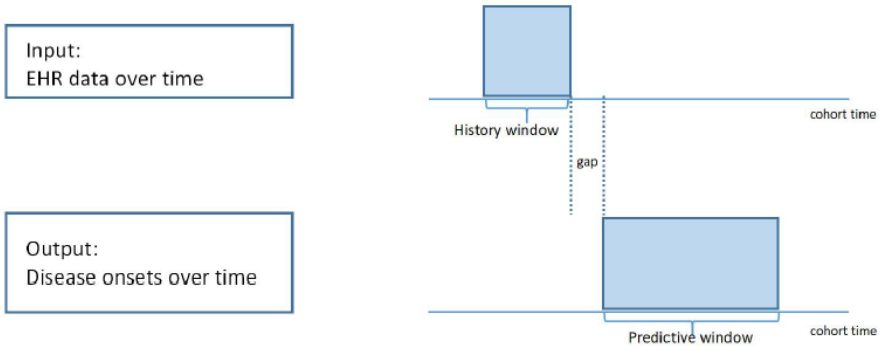

Fig. 2 Overview of prediction framework

C. Definition of PA and Predictive Variables

A PA record was defined according to the codes of C25.0-C25.9 from the International Classification of Disease, 10th Revision code (ICD-10).

To predict the risk of new-incident PA in the following two years, we compiled EHR datasets of the patients from a 12-month historical window. Patients who suffered PA during the targeted time frame were chart-reviewed by internal physician curators such that only the first PA records were utilized in a 24-month prediction window in this study. As a result, a total of 4361 PAs were identified, and a binary outcome label (1 or 0) was assigned to the cases and controls as the predictive dependent variables.

Accordingly, the candidate predictors were extracted from the EHR dataset during the time period of January 1, 2014 to December 31, 2015, which were mainly demographic characteristics, clinical utilization features, disease diagnosis from ICD-10 codes. We sampled the case or control group to randomly divide the same group subjects to 4:1 (training:test) subgroups. Subsequently, these 4/5 or 1/5 of the case and control subjects were combined to form the training or test cohort respectively. Therefore, the training or testing dataset was stratified and divided, rather than constructed by pure random.

D. Model Construction and Interpretation

1 Data Preprocessing

a. Feature Extraction

Every patient in the dataset was represented by a sequence of events, with each event providing the patient information that was recorded within a year period; The available data within these a year windows, along with additional summary statistics and augmentations, formed a feature set that was used as input to our predictive models.

We extract numerical lab and vital sign values from the EHR data. According to target dictionary, we selected common vital signs (i.e., weight, BMI, blood pressure, temperature, pulse and respiration rate) as well as lab test names. To keep the test simple, we will reference both vital signs and lab test values as lab values in the following sections.

Overall there are more than 300 types of lab test extract from the EHR. The most frequent item, weight, is found in about 65% of encounters, while the prevalence other diseases of pancreas to around 1% at the 233rd most frequent item.

In each encounter, the same item may have multiple values if the patient was diagnosed multiple times in a day. We computed the median, min and max within each encounter. We also standardized the sample values by subtracting mean and dividing by the standard deviation of the training sample distribution.

We did not use any imputation of missing numerical values, because common imputation of missing values does not always provide consistent augmentation to predictive models built on EHR. Instead, we associated each numerical feature with one or more discrete 'presence' features to help our algorithms to differentiate between the absence of a numerical value and an actual value of zero. Moreover, these existence features enciphered whether particular numerical values were considered to be medium, very low or very high. For some data points, the explicit numerical values were not recorded (usually when the values were considered medium), and the provision of this encoding of the numerical data allowed our algorithms to deal with these measurements even in their absence. Discrete features, such as diagnostics or procedural codes, were also quantified as binary presence features.

All numerical features were standardized to the [0,1] range after covering the extreme values at the 1st and 99th percentile. This prohibits the normalization from being possessed by potentially large data entry errors, while preserving most of the vital feature value.

2. Models for Predicting PA

Our predictive system operates over the EHR. At the time point, input features (as described in 'Feature representation') were provided to a statistical model, the output of which is a probability of stage of PA occurring in the next two years. If this probability exceeds a chosen operating threshold, we make a positive prediction that can then trigger a dangerous signal. This is a common framework within which existing algorithms also fit, and we describe the baseline methods in 'Competitive baseline methods' below. The contribution of this work is in the design of the deep model that is used and its training procedure, and the demonstration of its effectiveness

Open Science Index, Medical and Health Sciences Vol:14, No:11, 2020 publications.waset.org/10011557.pdf

—on a large-scale EHR dataset and across many different regimes— in making better predictions of future PA.

Fig. 3 gives a schematic view of our model, which makes predictions by first transforming the input features using an embedding module. This embedding is fed into a multi-layer neural network, the output of which at the time point is fed into a prediction module that provides the probability of future PA at the time horizon for which the model will be trained. To provide useful predictions, we train an ensemble of predictors to estimate the confidence of the model, and the resulting ensemble predictions are then calibrated using isotonic regression to reflect the frequency of observed outcomes [33].

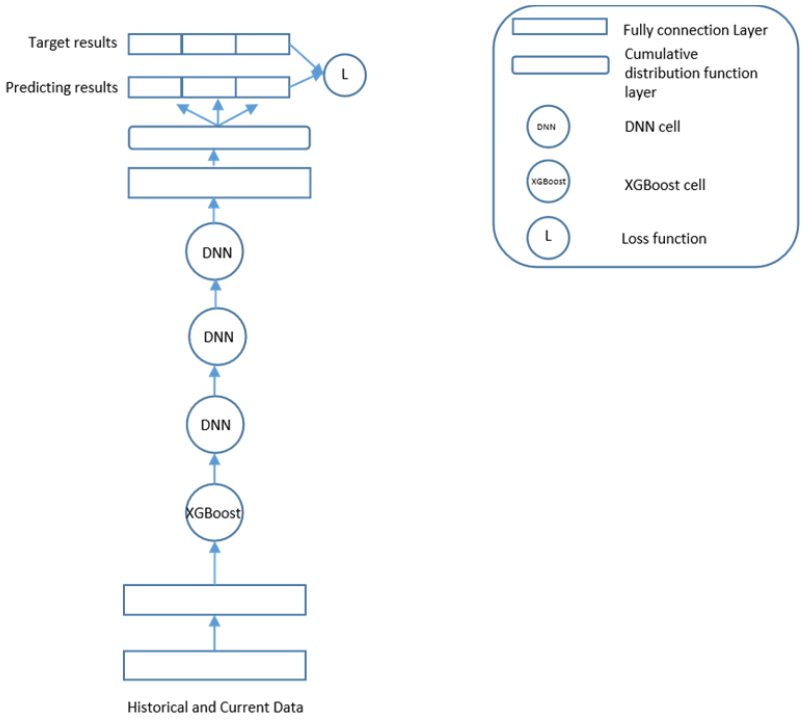

Fig. 3 A schematic view of our model

a. XGBoost Cell and DNN Cell

Firstly, the XGBoost layers transform the high-dimensional and sparse input features into a lower-dimensional continuous representation that makes subsequent prediction simply. Then, we use a deep neural network perceptron with residual connections and rectified-linear activations and use L_1 regularization on the embedding parameters to prevent overfitting and to ensure that our algorithm still focuses on the most-salient features. We compared simpler linear transformations, which did not perform as well as the multi-layer version we used.

b. Prediction Targets and Training Samples

The output of the DNN is fed to a final linear prediction layer that makes predictions over future prediction windows (2-year windows 3-month ahead). We use a cumulative distribution function layer across time windows to keep monotonicity, because the presence of PA within a shorter time window implies a presence of PA within a longer time window. Each of the resulting two outputs provides a binary prediction for PA at a specific time window and is compared to the ground-truth label using the cross-entropy loss function (Bernoulli log-likelihood).

Our overall loss function is the weighted sum of the cross-entropy loss from the PA predictions and the squared loss for each of the laboratory-test predictions. We inquired into the use of oversampling and overweighting of the positive labels to account for class imbalance. For oversampling, each mini-batch contains a larger percentage of positive samples than average in the entire dataset. For overweighting, the prediction for positive labels contributes proportionally more to the total loss.

c. Hyperparameters

We framed our proposed model on basis of the validation set performance and subsequently performed an analysis of the design choices. All variables are initialized via normalized initialization [31] and trained using the Adam optimization scheme [32]. We use exponential learning-rate decay during training process. The best validation results were implemented using an initial learning rate of 0.001, with a batch size of 128. The best-performing DNN architecture used a cell size of 128 units per layer and 3 layers.

d. Competitive Baseline Methods

Established models for future PA prediction make use of L_1-regularized gradient-boosted trees, trained on a clinically relevant set of features that are known to be important either for routine clinical practice or the modelling of pancreatic function. A curated set of clinically relevant features was

Open Science Index, Medical and Health Sciences Vol:14, No:11, 2020 publications.waset.org/10011557.pdf

chosen using existing PA literature and the consensus opinion of 5 clinicians: three senior attending physicians with over twenty-five years expertise, two pancreatic cancer experts. This set was further extended to include 30 of the most-salient features discovered by our deep learning model combined with XGBoost(XGBoost+DL) that were not in the original table, to give further predictive signal to the baseline. The final curated dataset contained 233 base features of demographics, admission information, vital sign measurements, select laboratory tests and medications, and diagnoses of chronic conditions that are directly associated with an increased risk of PA. The top 173 feature set is listed in Table I.

TABLE I
THE TOP 173 FEATURE IMPORTANCE VALUES CALCULATED BY XGBOOST+DL

Feature Description	Gain	Feature Description	Gain	Feature Description	Gain
Age Group >=85	0.11263	Encounter for screening for malignant neoplasms	0.00260	Polyene Antifungal	0.00070
Tobacco use	0.10638	Glaucoma	0.00259	Other interstitial pulmonary diseases	0.00070
BMI	0.10231	Age-related cataract	0.00244	Serotonin-3 Receptor Antagonist	0.00068
Overweight and obesity	0.08176	Asthma	0.00235	Thiazide-like Diuretic	0.00067
Other diseases of pancreas	0.05101	Vitamin K Antagonist	0.00232	Other cardiac arrhythmias	0.00067
Male, Age Group 65-74	0.03421	alpha-Adrenergic Agonist	0.00221	Phosphodiesterase 5 Inhibitor	0.00065
Other disorders of kidney and ureter, not elsewhere classified	0.02613	Hypertensive chronic kidney disease	0.00219	Other hypothyroidism	0.00065
Female, Age Group 50-64	0.01856	Other and unspecified polyneuropathies	0.00219	Vitamin B12	0.00065
Disorders of lipoprotein metabolism and other lipidemias	0.01557	Sleep disorders	0.00217	BMI 34.0-34.9, adult	0.00064
Opioid Agonist	0.01288	Diverticular disease of intestine	0.00198	Malignant neoplasm of prostate	0.00063
Age Group 50-64	0.01159	Personal history of other diseases and conditions	0.00195	Gastrointestinal System, Excision procedures	0.00061
Penicillin-class Antibacterial	0.01087	Problems related to lifestyle	0.00189	Calcium Channel Blocker	0.00061
Proton Pump Inhibitor	0.01058	Gout	0.00185	Patient had Other Government insurance	0.00061
Age Group 65-74	0.01038	Disorders of vitreous body	0.00177	Insulin Analog	0.00059
Essential primary hypertension	0.00996	Female	0.00176	Purpura and other hemorrhagic conditions	0.00058
Angiotensin 2 Receptor Blocker	0.00866	Malignant neoplasm of esophagus	0.00170	Aortic aneurysm and dissection	0.00056
Long term current drug therapy	0.00812	Other diseases of liver	0.00167	Non-narcotic Antitussive	0.00054
Nonsteroidal Anti-inflammatory Drug	0.00711	gamma-Aminobutyric Acid-ergic Agonist	0.00151	Loop Diuretic	0.00054
Live Attenuated Herpes Zoster Virus Vaccine	0.00687	Biguanide	0.00147	Paroxysmal tachycardia	0.00053
Presence of cardiac and vascular implants and grafts	0.00667	Potassium Salt	0.00143	Heart failure	0.00052
Serotonin and Norepinephrine Reuptake Inhibitor	0.00666	Patient had Medicaid insurance	0.00138	Estrogen	0.00051
Male	0.00644	Herpes Simplex Virus Nucleoside Analog DNA Polymerase Inhibitor	0.00127	Disorders of mineral metabolism	0.00049
Beta-Adrenergic Blocker	0.00609	Xanthine Oxidase Inhibitor	0.00126	Creatinine [Mass/volume] in Blood	0.00048
Type 2 diabetes mellitus	0.00609	Personal risk factors, not elsewhere classified	0.00125	Complications and ill-defined descriptions of heart disease	0.00046
Encounter for other postprocedural aftercare	0.00604	Persons encntr hlth serv for spec proc & trtmt, not crd out	0.00124	Cardiomyopathy	0.00045
Herpesvirus Nucleoside Analog DNA Polymerase Inhibitor	0.00602	Lower Joints, Replacement procedures	0.00124	Cholinergic Muscarinic Antagonist	0.00045
Histamine-2 Receptor Antagonist	0.00601	Gastro-esophageal reflux disease	0.00123	Phenothiazine	0.00044
Patient had Medicare insurance	0.00588	Presence of other functional implants	0.00122	Malignant neuroendocrine tumors	0.00043
Osteoporosis without current pathological fracture	0.00575	Central Nervous System Stimulant	0.00121	Encounter for screening for infec/parastc diseases	0.00042
Anti-epileptic Agent	0.00554	Antiarrhythmic	0.00118	Spondylosis	0.00042
Quinolone Antimicrobial	0.00553	Peroxisome Proliferator Receptor alpha Agonist	0.00118	Chronic kidney disease CKD	0.00042
Corticosteroid	0.00536	Allergy status to drug/meds/biol subst	0.00117	Central alpha-2 Adrenergic Agonist	0.00042
Patient had Blue Cross insurance	0.00535	Polymyxin-class Antibacterial	0.00113	l-Thyroxine	0.00042
Serotonin Reuptake Inhibitor	0.00521	Disorders resulting from impaired renal tubular function	0.00111	Thoracic, thoracolum, and lumbosacral intvrt disc disorders	0.00042
Tricyclic Antidepressant	0.00485	Anticholinergic	0.00111	Nitroimidazole Antimicrobial	0.00041
Beta2-Adrenergic Agonist	0.00435	Malignant neoplasm of breast	0.00110	Progestin	0.00038
Encntr for general exam w/o complaint, susp or reprtd dx	0.00428	Sulfonylurea	0.00109	Abnormal results of liver function studies	0.00035
Thiazide Diuretic	0.00415	Dopamine-2 Receptor Antagonist	0.00109	Arteriolar Vasodilator	0.00034
Aminoketone	0.00407	Male erectile dysfunction	0.00108	Other chronic obstructive pulmonary disease	0.00034

Open Science Index, Medical and Health Sciences Vol:14, No:11, 2020 publications.waset.org/10011557.pdf

Feature Description	Gain	Feature Description	Gain	Feature Description	Gain
Patient had Commercial insurance	0.00384	Serotonin-1b and Serotonin-1d Receptor Agonist	0.00107	Encounter for adjustment and management of implanted device	0.00033
Angiotensin Converting Enzyme Inhibitor	0.00382	Beta Lactamase Inhibitor	0.00105	Nonergot Dopamine Agonist	0.00033
Muscle Relaxant	0.00367	Dihydrofolate Reductase Inhibitor Antibacterial	0.00100	Thiazolidinedione	0.00032
Nitrate Vasodilator	0.00349	Atrial fibrillation and flutter	0.00098	Nonrheumatic aortic valve disorders	0.00032
Dihydropyridine Calcium Channel Blocker	0.00348	Other pulmonary heart diseases	0.00096	Abnormal laboratory tests SODIUM	0.00032
Amide Local Anesthetic	0.00347	Azole Antifungal	0.00095	Malignant neoplasm of bronchus and lung	0.00030
Benzodiazepine	0.00335	Low Molecular Weight Heparin	0.00089	Encounter for contraceptive management	0.00025
Encounter for immunization	0.00332	Acquired absence of organs, not elsewhere classified	0.00085	5-alpha Reductase Inhibitor	0.00025
Personal history of malignant neoplasm	0.00331	Osmotic Laxative	0.00083	Atypical Antipsychotic	0.00024
Antihistamine	0.00323	Irritable bowel syndrome	0.00081	Benign prostatic hyperplasia	0.00022
Personal history of certain other diseases	0.00323	Calcineurin Inhibitor Immunosuppressant	0.00079	Dependence on enabling machines and devices, NEC	0.00020
Macrolide Antimicrobial	0.00320	Pain, not elsewhere classified	0.00077	Chronic sinusitis	0.00019
Factor Xa Inhibitor	0.00309	Catecholamine	0.00076	Type 1 diabetes mellitus	0.00019
Encounter for other aftercare and medical care	0.00308	Chronic ischemic heart disease	0.00076	Platelet Inhibitor	0.00018
alpha-Adrenergic Blocker	0.00299	Encounter for screening for other diseases and disorders	0.00074	Other postprocedural states	0.00018
Peroxisome Proliferator Receptor gamma Agonist	0.00296	HMG-CoA Reductase Inhibitor	0.00074	Persons encntr health serv for oth cnsl and med advice, NEC	0.00013
Encntr for oth sp exam w/o complaint, suspected or reprtd dx	0.00279	Encntr for f/u exam aft trtmt for cond oth than malig neoplm	0.00074	Other and unspecified osteoarthritis	0.00013
Antiemetic	0.00269	Other specified health status	0.00073	Cardiac Glycoside	0.00006
		Other and unspecified hearing loss	0.00073	Vitamin D deficiency	0.00006

e. Evaluation

The data were split into training, validation and test sets in such a way that information from a given patient was present only in one split. The training split was used to train the proposed models. The validation set was used to iteratively improve the models by selecting the best model architectures and hyperparameters. Deep learning models with softmax or sigmoid output trained with cross-entropy loss are prone to miscalibration, and recalibration ensures that consistent probabilistic interpretations of the model predictions can be made [34]. The best models were evaluated on the independent test set that was retained during model development.

The main metrics used in model selection and the final report are: the area under the receiver operating curve. The PA episode sensitivity corresponds to the percentage of all PA episodes that were correctly predicted ahead of 3-month time within the corresponding time windows of up to 2-year.

III. Results

A. Model Performance

The AUC ROC of the XGBoost+DL model was 0.809(95% CI: 0.764–0.853) in the independent prospective cohort, indicating that the model was acceptable (Fig. 4). The AUC ROC curves of other models built with some popular algorithms, such as XGBoost (AUC 0.79, 95% CI: 0.747–0.832) and KNN (AUC 0.734, 95% CI: 0.685–0.734), on the same datasets were also shown in Fig. 4, which indicated that the XGBoost+Deep Learning model performed better than XGBoost and KNN.

Using the XGBoost+DNN on the EHR-based data, our prediction model found that PA patients were more likely to be in age groups of ≥ 65 years, to have diagnosed overweight and obesity, to have other diseases of pancreas. In general, a total of 233 features were significant in the predictive model. The performance of model decision interpretation is demonstrated using three representative individuals from the prospective cohort.

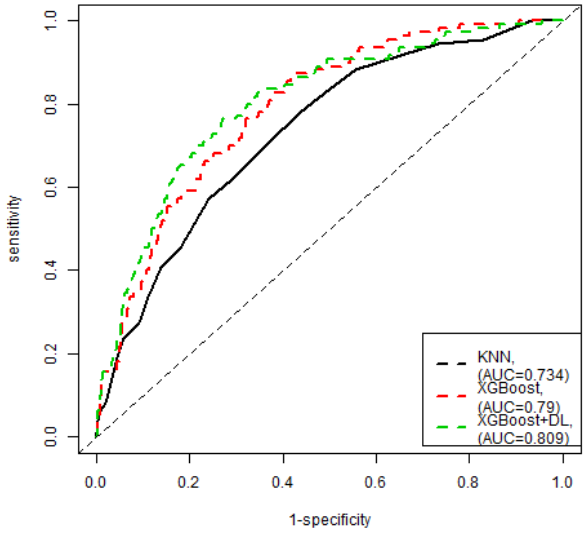

Fig. 4 ROC curves of three different algorithms applied on the prospective cohort. The AUC of XGBoost+DL model is 0.809, the AUC of the XGBoost model is 0.790, and the AUC of the KNN model is 0.734. The XGBoost+DL model has highest AUC and best performance compared to the KNN model and the XGBoost model

Open Science Index, Medical and Health Sciences Vol:14, No:11, 2020 publications.waset.org/10011557.pdf

Open Science Index, Medical and Health Sciences Vol:14, No:11, 2020 publications.waset.org/10011557.pdf

IV. Discussion

In this study, we constructed an EHR-based PA risk predictive model that adopted a XGBoost+DNN algorithm to automatically integrate useful clinical information of disease diagnoses, medication consumption, clinical utilization, lab-test results and predicted an older individual's risk of PA in the future two years. In the validation phase, this model attained an AUC of 0.809, and stratified individuals into three distinct risk categories of PA (high, intermediate, and low). About 54.35% of the individuals that had a confirmed PA event with a 2-year EHR cohort were classified into the increased risk categories. More importantly, our model successfully captured 91.20% of all PA that required subsequent chemoradiotherapy, with a lead-time of up to 3 months and a true alert of 67.62%, indicating the model's better performance for the long-term PA prediction.

V. Conclusion

In conclusion, we have constructed and validated a powerful risk assessment tool to predict adults' risk of PA in the future two years, by using the EHR data from the population in Maine. We intend that this constructed PA risk assessment tool could be immediately deployed to provide early three months warnings to adults with increased PA risk and identifying their personalized risk factors to facilitate customized PA interventions.

Funding

X.D. is supported by Natural Science Foundation of Zhejiang Province of China (Grant No. LY19F020042).

Author's Statement

The authors certify that this manuscript is not under review by any other journal. All authors have contributed to and read the manuscript and approved the final copy.

Declaration of Competing Interest

The authors certify that they do not have any financial or other conflicts of interest in relation to this manuscript.

Acknowledgments

The authors would like to thank and express their gratitude to the hospitals, medical practices, physicians, and nurses participating in Maine's HIE. They also thank the biostatistics colleagues at the Department of Health Research and at the All Vista Healthcare Center.

References

[1] Bray F, Ferlay J, Soerjomataram I, et al. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries: Global Cancer Statistics 2018 (J). CA A Cancer Journal for Clinicians, 2018, 68.

[2] Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods (J). International Journal of Cancer, 2019, 144.

[3] Hidalgo M. Pancreatic cancer (J). New England Journal of Medicine, 2010, 362(17): 1605-1617.

[4] Botsis T, Anagnostou VK, Hartvigsen G, Hripcsak G, Weng C. Developing a multivariable prognostic model for pancreatic endocrine tumors using the clinical data warehouse resources of a single institution. Appl Clin Inform 2010;1(1):12.

[5] Verma M. Pancreatic cancer biomarkers and their implication in cancer diagnosis and epidemiology. Cancers 2010;2(4):1830-7.

[6] Chakraborty S, Baine MJ, Sasson AR, Batra SK. Current status of molecular markers for early detection of sporadic pancreatic cancer. BBA-Rev Cancer 2011;1815(1):44-64.

[7] Arslan A A, Helzlsouer K J, Kooperberg C, et al. Anthropometric measures, body mass index, and pancreatic cancer: a pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan) (J). Archives of internal medicine, 2010, 170(9): 791-802.

[8] Ben Q, Xu M, Ning X, et al. Diabetes mellitus and risk of pancreatic cancer: a meta-analysis of cohort studies (J). European journal of cancer, 2011, 47(13): 1928-1937.

[9] Boursi B, Finkelman B, Giantonio B J, et al. A clinical prediction model to assess risk for pancreatic cancer among patients with new-onset diabetes (J). Gastroenterology, 2017, 152(4): 840-850. e3.

[10] Boursi S B, Finkelman B, Giantonio B J, et al. A clinical prediction model to assess risk for pancreatic cancer among patients with pre-diabetes (J). 2018.

[11] Cai Q C, Chen Y, Xiao Y, et al. A prediction rule for estimating pancreatic cancer risk in chronic pancreatitis patients with focal pancreatic mass lesions with prior negative EUS-FNA cytology (J). Scandinavian journal of gastroenterology, 2011, 46(4): 464-470.

[12] Canto M I, Goggins M, Hruban R H, et al. Screening for early pancreatic neoplasia in high-risk individuals: a prospective controlled study (J). Clinical gastroenterology and hepatology, 2006, 4(6): 766-781.

[13] Chari S T, Leibson C L, Rabe K G, et al. Probability of pancreatic cancer following diabetes: a population-based study (J). Gastroenterology, 2005, 129(2): 504-511.

[14] Gold D V, Goggins M, Modrak D E, et al. Detection of early-stage pancreatic adenocarcinoma (J). Cancer Epidemiology and Prevention Biomarkers, 2010, 19(11): 2786-2794.

[15] Radon T P, Massat N J, Jones R, et al. Identification of a Three-Biomarker Panel in Urine for Early Detection of Pancreatic Adenocarcinoma (J). Clinical Cancer Research An Official Journal of the American Association for Cancer Research, 2015, 21(15):3512.

[16] Grønborg M, Bunkenborg J, Kristiansen T Z, et al. Comprehensive proteomic analysis of human pancreatic juice (J). Journal of proteome research, 2004, 3(5): 1042-1055.

[17] Hart P A, Kamada P, Rabe K G, et al. Weight loss precedes cancer specific symptoms in pancreatic cancer associated diabetes mellitus (J). Pancreas, 2011, 40(5): 768.

[18] Hruban R H, Goggins M, Parsons J, et al. Progression model for pancreatic cancer (J). Clinical cancer research, 2000, 6(8): 2969-2972.

[19] Hsieh M H, Sun L M, Lin C L, et al. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models (J). Cancer management and research, 2018, 10: 6317.

[20] Klein A P, Lindström S, Mendelsohn J B, et al. An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population (J). PloS one, 2013, 8(9): e72311.

[21] Lowenfels A B, Maisonneuve P, Cavallini G, et al. Pancreatitis and the risk of pancreatic cancer[J]. New England Journal of Medicine, 1993, 328(20): 1433-1437.

[22] Lucenteforte E, La Vecchia C, Silverman D, et al. Alcohol consumption and pancreatic cancer: a pooled analysis in the International Pancreatic Cancer Case–Control Consortium (PanC4) (J). Annals of oncology, 2012, 23(2): 374-382.

[23] Masahiro N, Yingsong L, Hidemi I, et al. Prediction model for pancreatic cancer risk in the general Japanese population (J). PLoS ONE, 2018, 13(9):e0203386.

[24] Pannala R, Basu A, Petersen G M, et al. New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer (J). The lancet oncology, 2009, 10(1): 88-95.

[25] Pelaez-Luna M, Takahashi N, Fletcher J G, et al. Resectability of presymptomatic pancreatic cancer and its relationship to onset of diabetes: a retrospective review of CT scans and fasting glucose values prior to diagnosis (J). American Journal of Gastroenterology, 2007, 102(10): 2157-2163.

[26] Sah R P, Nagpal S J S, Mukhopadhyay D, et al. New insights into pancreatic cancer-induced paraneoplastic diabetes (J). Nature reviews Gastroenterology & hepatology, 2013, 10(7): 423.

[27] Permuth-Wey J, Egan K M. Family history is a significant risk factor for pancreatic cancer: results from a systematic review and meta-analysis (J). Familial cancer, 2009, 8(2): 109-117.
[28] Sanoob M U, Madhu A, Ajesh K, et al. Artificial neural network for diagnosis of pancreatic cancer[J]. International Journal on Cybernetics & Informatics, 2016, 5(2): 40-49.
[29] Verna E C, Hwang C, Stevens P D, et al. Pancreatic cancer screening in a prospective cohort of high-risk patients: a comprehensive strategy of imaging and genetics (J). Clinical cancer research, 2010, 16(20): 5028-5037.
[30] Wang W, Chen S, Brune K A, et al. PancPRO: risk assessment for individuals with a family history of pancreatic cancer (J). Journal of clinical oncology: official journal of the American Society of Clinical Oncology, 2007, 25(11): 1417.
[31] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks (J). Journal of Machine Learning Research, 2010, 9:249-256.
[32] Kingma D, Ba J. Adam: A Method for Stochastic Optimization (J). Computer ence, 2014.
[33] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates (C). The eighth ACM SIGKDD international conference. ACM, 2002.
[34] Mares T, Janouchova E , Kucerova A . Artificial neural networks in calibration of nonlinear mechanical models (J). Advances in Engineering Software, 2016, 95:68-81.

Open Science Index, Medical and Health Sciences Vol:14, No:11, 2020 publications.waset.org/10011557.pdf