

Learning Risk Models for Pancreatic Cancer from Electronic Health Records

by

Jennifer McCleary

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2020

© Massachusetts Institute of Technology 2020. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
January 29, 2020

Certified by
Martin C. Rinard
Professor
Thesis Supervisor

Accepted by
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Learning Risk Models for Pancreatic Cancer from Electronic Health Records

by

Jennifer McCleary

Submitted to the Department of Electrical Engineering and Computer Science
on January 29, 2020, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Pancreatic cancer is the third most lethal cancer in the U.S., causing an estimated 45,750 deaths in 2019. Of all treatments, surgical resection provides the best survival rate for pancreatic cancer. This is not feasible for the majority of pancreatic cancer patients, whose cancer is typically not diagnosed until the tumor is unresectable. Most symptoms of pancreatic cancer are typically subtle, which underscores the need for better risk modeling to predict a patient's chance of pancreatic cancer well before it would usually be diagnosed. We propose a series of novel models that apply standard machine learning techniques to Electronic Health Records (EHRs) to predict risk of pancreatic cancer in advance of cancer diagnosis. On the test dataset, two of our models achieved AUROCs of 0.85 (CI 0.81 - 0.90) and 0.79 (CI 0.77 - 0.82) with a 365-day lead time.

Thesis Supervisor: Martin C. Rinard
Title: Professor

Acknowledgments

I would first like to acknowledge my advisor, Martin Rinard, for his encouragement and guidance. He never failed to ask insightful questions about my work and remind me not to lose sight of the main goal. From the PAC group, I'd also particularly like to thank José Cambroneró, who was my constant point of contact throughout my entire research experience. José provided invaluable technical guidance, launching points for experiment ideas, and guidance with results analysis.

Many thanks to Limor Appelbaum from BIDMC for insight into the clinical basis behind this work and for generously donating her time to review my preliminary results.

I would also like to acknowledge TriNetX for granting us the EHR data that formed the backbone of this research and access to an AWS machine for running our models. In particular, Steve Kundrot, Matvey Palchuk, Jeffrey Warnick, Richard Rast, Drew Halloran, and Victor Yee generously provided their time and assistance to this project.

Finally, I'd like to thank my friends and family for their support. In particular, I am grateful for my roommate, Victoria, with whom I spent many days writing theses together.

Contents

1	Pancreatic Cancer	10
1.1	Pancreatic Cancer Background	10
1.2	Symptoms	10
1.3	Treatments	11
1.3.1	Resectable	11
1.3.2	Locally Advanced	11
1.3.3	Metastatic	12
1.4	Known Risk Factors	12
1.5	Early Detection	12
1.5.1	Screening Methods: Imaging	13
1.5.2	Other Screening Methods	13
1.5.3	Screening for High Risk Patients	14
1.6	Risk Models	14
2	Electronic Health Records	15
2.1	Electronic Health Records Background	15
2.2	Alternative Data Sources	15
2.2.1	Cohort Studies	16
2.2.2	EHRs vs. Cohort Studies	16
2.3	EHRs and Risk Modeling	17
2.3.1	Previous EHR Studies	17
2.3.2	EHRs for Pancreatic Cancer Risk Modeling	17

3	Previous Work	19
3.1	Machine Learning for Healthcare	19
3.1.1	Interpretability	19
3.1.2	Outliers	20
3.1.3	Data Leakage	21
3.1.4	Generalizability	21
3.1.5	Modeling Longitudinal Data	22
3.1.6	Privacy	22
3.2	Machine Learning for Pancreatic Cancer	23
3.2.1	High-Risk Groups	23
3.2.2	Prediction Time Cutoffs	24
3.2.3	Non-EHR Data Sources	24
3.2.4	EHR Data Sources	24
3.3	Our Models	25
4	Methodology	26
4.1	Data Source	26
4.2	Study Population	26
4.2.1	Cancer Patients	27
4.2.2	Control Patients	27
4.2.3	Sampling	28
4.3	Prediction Time Cutoffs	28
4.4	Regularization	29
4.5	Performance Metrics	30
4.5.1	AUROC	30
4.5.2	Precision	30
4.5.3	Confidence Intervals	30
4.6	Cross Validation	31
5	Feature Derivation	32
5.1	Diagnosis and Medication	32

5.1.1	Diagnosis Feature Derivation	32
5.1.2	Diagnosis and Medication Models	33
5.1.3	Windowed Diagnosis and Medication Models	34
5.2	Lab Tests	34
5.2.1	Lab Test Feature Derivation	35
5.2.2	Lab Test Models	36
5.2.3	Windowed Lab Test Models	37
6	Implementation	39
6.1	Logistic Regression	39
6.1.1	Regularization and Solvers	39
6.1.2	Other Parameters	40
6.1.3	Input data types	40
6.2	Other Software	41
6.3	Machine	41
6.4	Data Manipulation Optimizations	41
6.4.1	Data Compression	41
6.4.2	Data Chunking	42
6.4.3	Sparse Datatypes	42
6.5	Data Manipulation Results	43
6.5.1	Data Compression	43
6.5.2	Data Chunking	43
6.5.3	Sparse Datatypes	43
7	Results	47
7.1	Metrics	47
7.1.1	Diagnosis and Medication Models	47
7.1.2	Windowed Diagnosis and Medication Models	50
7.1.3	Lab Tests	51
7.1.4	Windowed Lab Tests	52
7.1.5	L2 vs L1 regularization	53

7.2	Discussion	54
7.2.1	Limitations	54
7.2.2	High Level Takeaways	55
8	Future Work	57
8.1	Learning Windows	57
8.1.1	Pre-processing	57
8.1.2	Multiple Classifiers	58
8.2	Additional Longitudinal Features	59
8.2.1	Challenges of Feature Engineering	59
8.2.2	Languages for Feature Engineering	60
8.2.3	Machine Feature Synthesis	61
8.3	LSTMs	61
9	Conclusion	62
A	Tables	63

List of Figures

5-1	Diabetes diagnosis relative to prediction time cutoff	33
5-2	Levels of glucose in blood relative to cancer diagnosis, with fitted line	35
5-3	Levels of glucose in blood relative to cancer diagnosis, with fitted lines per window	36
5-4	Levels of glucose in blood relative to cancer diagnosis, with fitted lines per window	37
6-1	Memory usage vs. dataframe size	44
6-2	Time vs. dataframe size	45
6-3	Time vs. dataframe size	46
7-1	Training Data: Area-Under-ROC (AUROC)	48
7-2	Training Data: Average Precision Score (APS)	48
7-3	Test Data: Area-Under-ROC (AUROC)	49
7-4	Test Data: Average Precision Score (APS)	49
7-5	AUROC on Diag and Med	50
7-6	AUROC on Diag, Med, Diag-Win, and Med-Win	50
7-7	AUROC on Lab Test Models	51
7-8	AUROC on Lab Test Models with Windowing	52
7-9	AUROC on Lab Test Models, L2 vs. L1 regularization	53

List of Tables

4.1	Counts(percentage) of different demographic indicators across the study population. Percentages are calculated with respect to either the case population or control population.	29
A.1	AUROC and APS scores for all models on training and test data . . .	64
A.2	95% Confidence Intervals for AUROC scores for all models on training and test data	65
A.3	95% Confidence Intervals for APS scores for all models on training and test data	66

Chapter 1

Pancreatic Cancer

1.1 Pancreatic Cancer Background

Pancreatic ductal adenocarcinoma (PDAC), more commonly known as pancreatic cancer, is a cancer that begins in the cells lining the pancreatic duct. PDAC is one of the most lethal cancers in the United States. The American Cancer Society estimates that there will be 56,770 new cases of pancreatic cancer and 45,750 deaths from pancreatic cancer in the United States in 2019 [68].

Among all cancers, pancreatic cancer has one of the worst five-year survival rates, at less than 4.8 percent. This is due to the fact that it often is not diagnosed until a very advanced stage, at which point it is difficult to treat [62].

1.2 Symptoms

Pancreatic cancer is usually not detected until a late stage because early symptoms are difficult to detect. As pancreatic tumors grow larger, they make their presence known by compressing surrounding organs, such as the bile ducts and duodenum [49]. Once the tumors are large enough to cause noticeable symptoms, the cancer has often reached an advanced stage and is virtually incurable.

Earlier symptoms of pancreatic cancer are often vague, including jaundice (a yellowing of the skin and eyes caused by excess bile); digestive problems caused by the

tumor pressing against the stomach; back pain; unexplained weight loss; abdominal bloating; new-onset diabetes; and depression. Pancreatic cancer may also cause biochemical processes such as consumption of LDL (low-density lipoprotein, also known as "bad" cholesterol) by the tumor [50] and protein breakdown [54]. Studies suggest that these symptoms may exist years before a patient is formally diagnosed with pancreatic cancer, which gives us hope that machine learning models may be able to detect risk factors for pancreatic cancer from previous medical data [74].

1.3 Treatments

Treatment options include surgery, chemotherapy, radiation, and more experimental treatments. The plausibility of each option depends on how advanced the cancer is.

1.3.1 Resectable

A resectable tumor is one which can be removed by surgery. Around 15-20% of patients with pancreatic cancer have resectable pancreatic tumors; among these patients, the 5-year survival rate increases to 20% [88]. Patients with resectable tumors often also undergo chemoradiation or chemotherapy [29], although clinical research has not yet rigorously established which is most helpful [63].

1.3.2 Locally Advanced

Locally advanced pancreatic tumors are those that are not resectable, but not yet metastasized to distant organs. These tumors generally are not resectable because they encase vascular structures, such as the superior mesenteric artery [49]. Patients with this form of pancreatic cancer are usually treated with fluorouracil-based chemoradiation. Chemoradiation tends to provide a palliative benefit, providing pain relief in 50-85% of patients [58]. However, it has a low chance of controlling the spread of the pancreatic tumor. In a study by the Gastrointestinal Tumor Study Group, patients with locally advanced pancreatic tumors who received fluorouracil-

based chemoradiation had a median survival of only 42-44 weeks [59]. Patients often experienced new metastatic diseases to the liver or other organs within a few months of treatment.

1.3.3 Metastatic

Pancreatic cancer is typically not diagnosed until it becomes metastatic, but at this stage chemotherapy is not curative [49]. The standard treatment for metastatic pancreatic cancer is currently a drug called gemcitabine. While gemcitabine provides slightly higher survival rates than other drug treatments, it is not at all close to a cure, with a median survival rate of 5-6 months [13].

1.4 Known Risk Factors

The most commonly found risk factors for pancreatic cancer are age (much more likely in old age) and cigarette smoking [6]. Family history is another risk factor. Genetic syndromes such as hereditary pancreatitis, hereditary non-polyposis colorectal cancer, familial breast cancer, familial atypical multiple mole melanoma, and Peutz-Jeghers syndrome are correlated with increased risk of pancreatic cancer [33]. Some studies suggest that the pancreas is susceptible to carcinogen exposure and DNA damage, which may cause genetic mutation that leads to the development of a tumor [81].

Despite many studies investigating risk factors of pancreatic cancer, the causal factors are very poorly understood. Knowing these clinical risk factors alone is not enough to help detect pancreatic cancer at early stages.

1.5 Early Detection

When pancreatic cancer is detected earlier, five-year survival rates increase dramatically [3]. We now present some existing methods for early detection of pancreatic cancer.

1.5.1 Screening Methods: Imaging

Pancreatic cancer is most commonly screened for using imaging methods such as transcutaneous ultrasound (TCUS), computed tomography (CT) scan, magnetic resonance imaging (MRI), and endoscopic ultrasound (EUS) [46]. Many of these imaging methods are unable to detect pancreatic cancer until the tumors are large and unresectable. EUS is the most sensitive method, and has been shown to be especially superior at detecting tumors less than 2 cm in length [72].

If an imaging method detects a mass in the pancreas, a tissue diagnosis is needed to confirm whether or not the mass is a tumor. Tissues can be evaluated using methods such as CT-guided biopsy or EUS-guided fine needle aspiration [31].

1.5.2 Other Screening Methods

Recent advances have shown that molecular methods may be able to detect early signs of pancreatic cancer. Sequence of precursor lesions have been able to find the genes that are mutated by different lesions in the pancreas [7]. To look for mutant genes shed by precursor lesions, doctors can observe pancreatic secretions using an endoscope. One study showed that the GNAS genetic mutations were detected in pancreatic secretion samples from two-thirds of patients with an IPMN (a type of precursor lesion to pancreatic cancer) [40]. GNAS mutations were also found in a patient with no pancreatic lesion who later developed a lesion.

The screening method described above is invasive, so is less practical than other methods such as blood tests. One study shows that the KRAS gene mutation was discovered in the plasma of 85% of patients with metastatic pancreatic cancer; however, it was only found in 45% of patients with resectable pancreatic cancer [45]. It remains to be shown whether pancreatic cancer can be detected early through non-invasive methods like plasma analysis.

While molecular methods are promising, for the moment they are still not as precise as imaging methods. Early screening for pancreatic cancer thus typically makes use of imaging methods such as EUS (see section 1.5.1).

1.5.3 Screening for High Risk Patients

Screening the general public to look for tumors is infeasible due to the fact that the incidence of pancreatic cancer is fairly low. One study estimated that for pancreatic cancer screening to be effective, screening needed to have a 16% chance or greater of detecting the disease [71].

Some studies have made use of screening methods by only screening patients who are deemed to be at high risk for pancreatic cancer. These studies were primarily focused on patients with family histories of the disease. As mentioned in section 1.4, family history of one of the main known risk factors for pancreatic cancer.

For example, a study by Canto et. al followed 38 patients with no symptoms of pancreatic cancer who had at least two first-degree relatives who had had pancreatic cancer [16]. The patients were regularly screened using EUS (see section 1.5.1). Over a period of three years, six of the patients were found to have pancreatic masses. One of the patients went on to develop pancreatic cancer. This patient was treated and was alive and cancer-free at the time the study was published, 5 years after surgery. This study provided hope that screening high-risk populations could yield higher rates of early detection.

1.6 Risk Models

The abysmally low 5-year survival rate of pancreatic cancer underscores the need for early detection. Screening methods such as EUS are somewhat effective at detecting early-stage pancreatic tumors, but it is infeasible to screen the general population. Current screening methods typically focus on patients with family histories of pancreatic cancer to make the screening cost-effective. This thesis proposes several new risk models that use Electronic Health Record data to identify high-risk patients in the general population, beyond just patients with known clinical risk factors such as genetics and smoking. Our hope is that these models will help identify people who are at high risk of pancreatic cancer, who can then receive early screening that would otherwise have missed them until their tumors were metastatic.

Chapter 2

Electronic Health Records

2.1 Electronic Health Records Background

Electronic Health Records (EHRs), which are digital versions of a patient's paper chart, have becoming increasingly available over the past decade. In 2008, only 9.4% of non-Federal acute care hospitals had basic EHRs; by 2015, 83.8% had adopted EHRs [19]. The vast quantity of information stored in EHRs provides a rich source for clinical risk models to build upon.

EHRs contain a wealth of patient data over time, including demographics, diagnoses, medications, lab test results, vital signs data, immunizations, radiology reports, and notes written by doctors [57]. We are able to more easily access patients' medical histories using EHRs.

2.2 Alternative Data Sources

Before the advent of EHRs, and even after, many risk prediction algorithms were developed on alternative data sources, in particular cohort studies.

2.2.1 Cohort Studies

A prospective cohort study first identifies a study population that is free of the disease to be researched. The study then follows this population over a number of years, collecting data from patients at regular intervals [76].

One of the most well-known cohort studies is the Framingham Heart Study, which began in Framingham, Massachusetts in 1948 with 5209 adult participants [86]. The study is still ongoing, with a fourth generation of participants now involved. This landmark cohort study has led to the production of over 3000 peer-reviewed scientific papers and a much better understanding of risk factors for heart disease, as well as other afflictions.

Cohort studies are advantageous for risk modeling because they have regular follow-up with patients and defined metrics that will be collected during each follow-up. They also provide a means for assessing causality. However, they can be expensive to set up, and require long periods of time to pass before results become available.

2.2.2 EHRs vs. Cohort Studies

EHR data is generally collected whenever a patient comes into contact with a hospital. This data is semi-structured: it contains both structured data, such as tables with diagnoses and lab test data, and unstructured data, such as images and clinical notes. EHR data also comes at irregular intervals and does not contain the same metrics each time, depending on what occurred during each visit. By definition, EHR studies are retrospective because they consider data from patients' pasts. This stands in stark contrast to cohort studies, which are prospective, and in which patients come in at regular intervals and have the same metrics observed during each visit.

EHR-based risk prediction has advantages: it allows a study to consider far more patients than a cohort study does, for a far smaller cost. EHR data may be more frequent than cohort study data, depending on the patient. Since EHRs are based on real clinical visits, rather than study settings, risk models based on EHRs may be more directly applicable outside of the study than risk models based on cohort

studies.

The way in which EHR data is collected also leads to various challenges for risk model development. Since people tend to go to the hospital when they are sick, EHR data may not be representative of how healthy the general population is. Clinicians only collect select metrics at each visit, so data is not consistent. Data will almost certainly be messier than that of cohort studies, which is standardized per visit [57].

2.3 EHRs and Risk Modeling

Despite their challenges, EHRs have been used for a variety of clinical risk prediction models, showing their promise for further research purposes. For example, models have been built on EHR data to predict gastric, lung, ovarian, and pancreatic cancer [10].

2.3.1 Previous EHR Studies

According to a study of 107 EHR-based studies, most studies considered large sample sizes [10]. The median size was 26,100 observations; some studies had over 100,000 observations. However, the number of predictors (here, predictors are equivalent to features) used was relatively small: the median was 27 predictors.

The study also investigated how EHR-based studies handled challenges of EHRs, such as repeated measurements over time and missing data. Most studies did not model repeated measurements over time, or simply used summary metrics. Many studies did not consider missing data; the ones that did mainly used multiple imputation to fill the gaps. None considered the presence vs. absence of metrics as a feature.

2.3.2 EHRs for Pancreatic Cancer Risk Modeling

In this thesis, we chose to develop our models on EHR data. Because of the wealth of EHR data at our disposal, this allowed us to consider many more patients than

with a cohort study. This is especially helpful for rare diseases, like pancreatic cancer. Using EHR data also makes our models more applicable to a clinical setting, as they could be used to analyze EHRs present in hospitals in the future.

Chapter 3

Previous Work

The advent of EHRs have led to an explosion in the amount of medical data available for analysis. We'll first consider some general concerns about using machine learning for healthcare analysis, then consider some previous studies related to risk modeling for pancreatic cancer specifically.

3.1 Machine Learning for Healthcare

Machine learning can be applied to a variety of problems in healthcare. For example, it has been used to improve patient risk score systems, streamline hospital operations, and predict the onset of disease [14]. We'll now explore some concerns researchers need to address when using machine learning for healthcare purposes.

3.1.1 Interpretability

Interpretability is an important concern for machine learning solutions in healthcare. Machine learning algorithms often function as "black boxes" that output results with no explanation of the reasoning behind the answer. In the case of healthcare, these results may be used to inform diagnoses and care pathways. Clinicians and others involved in these decisions would most likely want to know the reasoning that led to certain results.

Interpretable Traditional Models

Various studies have attempted to create interpretable machine learning models for healthcare by utilizing more traditional models. Some use decision trees, which produce sets of rules [43]; k-nearest neighbors [26] or distance metric learning [78], which find similar patients and allow for case-based reasoning; and logistic regression, which produces interpretable weights for features [25].

For our research, we are interested in the interpretability of logistic regression models. Logistic regression is a generalized linear model that uses a weighted sum to output a probability between 0 and 1 [32]; the feature weights can be interpreted as a multiplicative factor of the odds [60]. Therefore, we can interpret a higher weight for a feature as meaning that the feature contributed more to the output. We can detect which features were most useful for the final classification by looking at the features that had the highest weights in the trained model. Logistic regression is commonly used in healthcare analysis [83, 70].

Interpretable Deep Learning Models

Recurrent neural networks have been successfully used to predict disease progression from longitudinal EHR data [17]. However, they are much more difficult to interpret than traditional models such as decision trees. Choi et. al attempted to fix this problem by using an attention mechanism that emulated the behavior of physicians during a patient encounter [18]. Their model, called RETAIN, uses a standard attention mechanism as well as a second layer that generates interpretable outputs. The interpretable outputs identify the visits that contribute to predictions, and the medical variables within each visit that are most relevant.

3.1.2 Outliers

Data for machine learning can come from many sources, including structured EHRs, clinical notes, physiological waveforms, radiologic images, and more [85]. Wherever the data comes from, researchers will have to deal with "messiness" inherent in the

data, including outliers, incorrect lab test results, errors in clinical notes, and more. We want EHR data to be "plausible" - that is, the data values should lie in a believable range [39].

Estiri et al. used a clustering approach to detect outliers in electronic health records data [23]. They hypothesized that there should be very few outliers in their datasets. Using K-means clustering, they separated the EHR data values into clusters and removed clusters that had sparse populations.

A survey of 80 articles about outlier detection in healthcare data found that most articles used statistical methods to detect outliers [27]. Statistical methods generally fit a model for normal behavior to the data and then use a statistical inference method to determine if data points are outliers to this model or not. Other common methods for detecting outliers include clustering, in which similar data is grouped into clusters to identify outliers, and classification, which trains a binary classifier to identify data as either normal or an outlier.

3.1.3 Data Leakage

Data leakage can cause models to return results that are too good to be true. For example, consider a model trying to predict diagnoses of a certain disease. If the model is trained on medication data, a certain medication prescribed may already encode the outcome (if the medication is highly correlated with a certain disease diagnosis) [85]. Therefore, researchers must be careful not to include data that already specifies the outcome.

3.1.4 Generalizability

Models may often be trained on data from a single hospital. Because different institutions may collect and store data in different ways, models may generalize poorly to different institutions than the one in which they were developed [84]. Models can counter this problem by developing generalized methods that can be trained on institution-specific data to produce institution-specific methods, or by training models

using data from many different institutions.

Researchers often test the generalizability of their models by testing models on different populations and ensuring that metrics such as AUROC (Area Under the Receiver Operating Characteristic Curve) do not change significantly between the different datasets. For example, Kartoun et al. developed a post-discharge mortality prediction model for cirrhosis patients. Their model was trained on a population of 314,292 patients from two different hospitals, and was validated on a population of 18 million patients from independent EHR data sources [42].

3.1.5 Modeling Longitudinal Data

EHRs and other datasets often contain longitudinal data, such as lab test results over long periods of time. Because data points are not collected at regular intervals for EHRs, researchers must find ways to deal with missing data points when modeling longitudinal data in EHRs.

Missing values are typically filled with imputation [48]. This includes filling values with the most recent previous value, or filling with the mean or median of the other available values. Lipton et. al consider missingness itself as a feature, as the presence or absence of a value can indicate an important clinical decision [52].

3.1.6 Privacy

When working with medical data, it is important to protect the identity of patients involved. Various methods have been developed to increase privacy strength in machine learning applications for healthcare.

De-Identification

EHR data is usually anonymized: data that could possibly identify the patient is removed [5]. However, it is often possible for attackers to re-identify patients from the remaining indicators in de-identified data [22]. Updated de-identification methods include k-anonymity [66], in which asterisks replace various data values and some data

values are replaced with general values (e.g., birth year instead of exact birthday), and L-diversity [53], in which methods from k-anonymity are used to ensure that any given record maps onto at least k different records in the original data, where k is a customizable parameter.

Privacy Advances in ML

Aside from de-identifying data, some researchers have taken further steps to ensure patient privacy in their machine learning models. Graepel et al. developed a model that could be trained on encrypted data using a homomorphic encryption scheme [28]. Elmisery et al. showed that data privacy could be better ensured by storing patient data in distributed databases where each database had a different horizontal or vertical partition of the data [21]. They developed a clustering algorithm that would partition the data among different healthcare data providers.

3.2 Machine Learning for Pancreatic Cancer

Now we focus on previous machine learning work that deals with pancreatic cancer specifically.

3.2.1 High-Risk Groups

There are several known risk factors for pancreatic cancer, as discussed in section 1.4. Some risk models for pancreatic cancer focus specifically on these high-risk groups. Boursi et al. focus on patients with new-onset diabetes [9], while Hsieh et al. focus on patients with any diagnoses of diabetes 2 [34].

Some models do not focus specifically on high-risk patients, but derive their features from symptoms known to be associated with pancreatic cancer. For example, Muhammad et al. chose 18 features from literature and clinician-judged plausibility [61]. These features included age, smoking habits, drinking frequency, family history, and previous cancer diagnoses.

Our novel models do not use hand-crafted feature sets. Instead, we considered most of the features available in our EHRs, excluding some that occur very rarely.

3.2.2 Prediction Time Cutoffs

Because predicting pancreatic cancer a short time prior to a cancer diagnosis would not be useful in a clinical setting, our models considered data in sufficient advance of a cancer diagnosis. Some existing work consider prediction time cutoffs [24], but others make no mention of using cutoffs.

3.2.3 Non-EHR Data Sources

Many existing pancreatic cancer risk modelings were developed on non-EHR datasets. Klein et al. considers data from the PanScan Consortium, which consists of more than a dozen prospective epidemiologic cohort studies [4]. Other models have considered data sources from survey data to case-control studies [24].

Our studies focused on using EHR datasets, due to the fact that they contain richer data than cohort datasets such as the PanScan Consortium and are more directly applicable to clinical settings.

3.2.4 EHR Data Sources

Few existing studies have used EHRs to develop risk prediction models for pancreatic cancer. Stapley et al. used EHRs from the General Practice Research Database to perform a case-control study for prediction of pancreatic cancer [77]. This study differs from ours because it does not consider longitudinal data, such as changes in lab tests over time. It also uses a case-control ratio of 1:5, as opposed to our ratio of 1:100.

3.3 Our Models

Our work drew upon various points of inspiration from the works cited above to develop novel models. We performed a case-control study on data available in EHRs. Rather than hand-crafting feature sets based on clinical knowledge, we considered most of the features available in the EHR dataset. We made use of more features by considering longitudinal data, such as lab tests, as opposed to just binary features such as diagnoses.

We also drew upon various practices from other works to reduce concerns with using machine learning for healthcare. To prevent data leakage, we only used data prior to a diagnosis, and instituted an additional prediction cutoff to ensure that the model is helpful at identifying risk of cancer well in advance of when it would be diagnosed with current clinical practices. We tried to ensure generalizability by training models on EHRs collected from many different healthcare organizations. We used logistic regression rather than deep learning methods to increase the interpretability of our models.

Chapter 4

Methodology

4.1 Data Source

We used data in the form of Electronic Health Records (EHRs) compiled by TriNetX [1]. TriNetX is a global health research network that provides real-time access to de-identified EHR data from almost 70 different health care organizations (HCOs). The dataset we used contained data from 26 different HCOs and contained information on diagnoses, medications, lab tests, vital signs, genomics, and demographics.

All patient data provided by TriNetX was de-identified. Patients were labelled with 40-character hashes, and personally identifying data such as exact birthdays and weight was not provided. The data was hosted on TriNetX's servers, and we accessed the data using machines provided by TriNetX. The machine's disks were encrypted and could be wiped remotely if necessary.

4.2 Study Population

We now discuss how patients were selected for our study.

4.2.1 Cancer Patients

We focus on a 60-80 year old subpopulation as pancreatic cancer incidence in this group is known to be higher than in younger subpopulations. We defined the age cutoff as the age of the cancer patient at the time of their cancer diagnosis. In addition, using the full dataset was impractical due to computational requirements.

We used ICD10 codes C25.0, C25.1, C25.2, C25.3, C25.7, C25.8, and C25.9 to identify potential pancreatic cancer patients. To reduce the chance of false positives, we cross-checked the cancer patients against the tumor registry in the TriNetX data, which included records from the HCOs' cancer registries and additional patients that TriNetX has identified by using natural language processing to analyze patient records. Only about 25% of the patients in the EHR dataset with ICD codes indicating pancreatic cancer also had records in the tumor registry. This supports the findings of Taxiarchis Botsis et al, who conducted a study on pancreatic cancer patients from The Columbia University Medical Center's EHR datasets. Botsis et al found that only half of the patients who had ICD-9 codes corresponding to pancreatic cancer also had a diagnosis present in pathology reports [11].

Some hospitals often treat patients who are referred from outside practices. However, this means that their medical histories in the EHRs from that hospital are very limited. To reduce the number of patients with missing data, we only considered patients who had one or more visits at least 6 months before their cancer diagnosis.

After filtering by age, tumor registry, and prior visits, we were left with 2430 cancer patients.

4.2.2 Control Patients

For control patients, we considered patients that had never had a diagnosis of pancreatic cancer. To ensure that patients had enough data, we required that they had at least one hospital encounter in the last 5 years (as of December 2019) and at least one visit 6 months prior to that.

4.2.3 Sampling

We chose to use a case-control approach in which each cancer patient was matched with some number of control patients. The cancer patient and control patient had to be the same sex; the control patient had to have a visit during the same year as the cancer patient’s diagnosis; and the cancer and control patient had to be the same age during that year.

We allowed a tolerance of plus or minus three years for the visit year and age to increase the number of cancer patients who were matched. For example, we ran sampling at a ratio of 1 cancer patient to 100 control patients. Of the 2430 cancer patients, if we required the year and age to be the same, only 2350 patients were matched with control patients; with a tolerance of three years, 2420 patients were matched.

Because of memory constraints on our machine (see section 6.3 for machine specification), we had to reduce the sampling size for some of the models. We used a sampling ratio of 1:100 for the basic diagnosis and medication models (see section 5.1.2); 1:50 for the windowed diagnosis and medication models (see section 5.1.3); and 1:25 for the lab test models (see section 5.2.2). The windowed diagnosis and medication models had three times as many features as their non-windowed counterparts. Because of memory restraints, we weren’t able to run these larger models with 1:100 sampling. Similarly, the lab test models had many more features than any of the diagnosis and medication models, so we had to reduce the sampling to 1:25.

Table 4.1 shows demographics information for cancer patients and control patients who were selected with 1:100 sampling. Percentages for 1:50 sampling and 1:25 sampling were similar, so they are not included.

4.3 Prediction Time Cutoffs

Identifying pancreatic cancer too close to the time of cancer diagnosis is clinically useless because the tumor would likely be unresectable. Therefore, for our models we picked prediction time cutoffs of 180, 270, and 365 days before diagnosis. For cancer

Table 4.1: Counts(percentage) of different demographic indicators across the study population. Percentages are calculated with respect to either the case population or control population.

Stat	Value	Cases	Controls
Age	<60	0 (0.0)	0.0 (0.0)
Age	60-65	666 (27.5)	66700.0 (27.6)
Age	65-70	688 (28.4)	68796.0 (28.4)
Age	70-75	586 (24.2)	59104.0 (24.4)
Age	75-80	480 (19.8)	47400.0 (19.6)
Age	>80	0 (0.0)	0.0 (0.0)
Sex	F	1176 (48.6)	117600.0 (48.6)
Sex	M	1244 (51.4)	124400.0 (51.4)
Race	American Indian or Alaska Native	5 (0.2)	725.0 (0.3)
Race	Asian	41 (1.7)	5712.0 (2.4)
Race	Black or African American	240 (9.9)	21016.0 (8.7)
Race	Native Hawaiian or Other Pacific Islander	1 (0.0)	232.0 (0.1)
Race	White	2026 (83.7)	197520.0 (81.6)
Race	Unknown	107 (4.4)	16795.0 (6.9)

patients, the cutoffs were relative to the cancer diagnosis date; for control patients, the cutoffs were relative to the date of their last hospital encounter. For each cutoff, we removed data that occurred after the cutoff, and we also removed patients that had no data before the cutoff.

4.4 Regularization

All models implemented in this paper used logistic regression with regularization. The main types of regularization used with logistic regression are L1 (also known as Lasso) and L2 (also known as ridge regression). L1 regularization uses a penalty term that attempts to minimize the sum of the absolute values of the features, while L2 regularization uses a penalty term that attempts to minimize the sum of the squares of the features [64]. L1 regularization causes many parameter weights to zero out, so it simultaneously learns weights and reduces the size of the feature vector [69]. This makes L1 regularization useful in feature selection settings, when researchers want to reduce the number of features. However, there is some evidence to suggest that this feature selection can come at the cost of accuracy [80].

We chose to use L2 when possible in our models, and L1 when we had a large number of features that we wanted to prune.

4.5 Performance Metrics

To evaluate our models, we used two main metrics: Area Under the Receiver Operating Characteristic Curve (AUROC) [38], and Average Precision Score (APS) [8].

4.5.1 AUROC

The Receiver Operating Characteristic curve (ROC) plots the true-positive rate against the false-positive rate, showing the trade-off between these two metrics. The higher the area under this curve, the better the model is at discriminating between the two classes. An AUROC of 0.5 is equivalent to a binary classifier randomly guessing between two options, while an AUROC of 1.0 represents a perfect classifier.

4.5.2 Precision

Ideally, we would want a classifier to have both high precision (accurate results) and high recall (identifies positive results). The precision-recall curve represents the tradeoff between these two metrics, and the Average Precision Score (APS) is an approximation of the area under this curve. As with the AUROC, the higher the APS, the better the classifier.

One property to note about APS is that while recall is monotonically increasing with respect to the classification threshold, precision is not necessary monotonically decreasing with respect to the threshold. This means that if the classification threshold increases, precision may either increase or decrease.

4.5.3 Confidence Intervals

For all of the models, we report the AUROC and APS values as well as 95% confidence intervals. We compute confidence intervals using the empirical bootstrap method [12].

This method involves randomly sampling the patients with replacement to calculate the metric, then averaging this metric over 1000 iterations.

4.6 Cross Validation

We ran all of our models using 10-fold cross-validation. Various studies have found that 10-fold cross-validation can successfully estimate classification error on real-world datasets [47, 44], and it is a popular method of estimating classification error and selecting models [30]. To compute the metrics (AUROC and APS), we averaged the scores over all 10 folds.

Chapter 5

Feature Derivation

5.1 Diagnosis and Medication

During the course of a hospital encounter, patients may receive various diagnoses for conditions such as diabetes mellitus and pancreatitis. In EHRs, these diagnoses are denoted with alphanumeric strings called ICD codes. Some diagnoses, such as depression and new-onset diabetes, are known to be correlated with pancreatic cancer. Similarly, patients may receive various medications during a medical encounter. By considering all diagnoses present in the dataset and all medication codes present in the dataset, the goal is to find diagnoses and medications correlated with pancreatic cancer.

5.1.1 Diagnosis Feature Derivation

There are two main aspects with which we can consider a single ICD code: whether or not a patient ever had a diagnosis of that ICD code, and when the patient received that diagnosis. As an example of what this would mean for a specific ICD code, consider figure 5-1, which shows when a sample of patients received their first diagnosis of diabetes prior to the prediction time cutoff (see section 4.3). Figure 5-1a represents cancer patients, while figure 5-1b represents control patients. The figure shows that most cancer patients did not receive their first diagnosis of diabetes until relatively

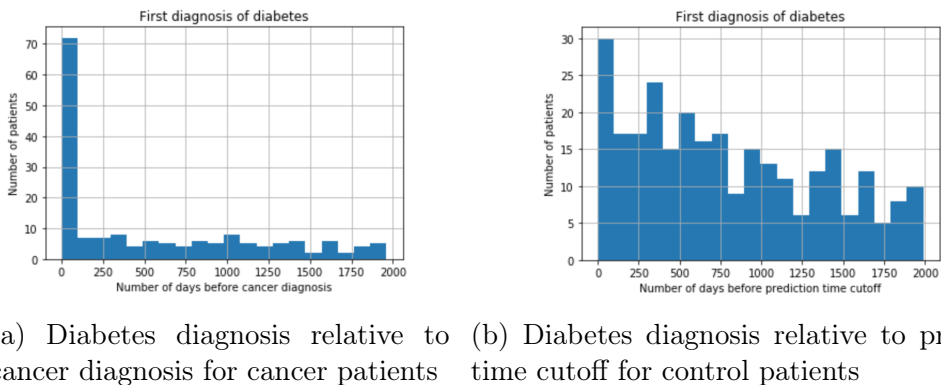


Figure 5-1: Diabetes diagnosis relative to prediction time cutoff

close to their cancer diagnosis, while control patients received their first diagnosis of diabetes at a wide variety of times before their prediction time cutoff. Having a diagnosis of diabetes may be a risk factor for cancer, but this graph suggests that when a patient has their first diagnosis of diabetes may be an important factor as well. This reflects clinical knowledge that new-onset diabetes is a risk factor for pancreatic cancer [49]. This graph does not include all patients and is not a suggestion that we focus on new-onset diabetes specifically, but is an example of why we may consider as features both 1) the fact that a patient has ever had a certain diagnosis and 2) the relative date that patient had their first instance of that diagnosis.

5.1.2 Diagnosis and Medication Models

For our diagnosis model, we considered all ICD codes that appeared at least 100 times in the dataset. We did this to prevent our model from overfitting on diagnoses that occur very rarely. This left us with 27162 unique ICD codes. For each of these diagnoses, we used a binary feature that was 1 if the patient had ever received that diagnosis, and 0 otherwise. We called this model *Diag*.

For our medication model, we used medication codes to derive the same binary features that were used in the diagnosis model. As with the diagnosis model, we only considered medication codes that appeared at least 100 times in the dataset. This left us with 2059 unique medication codes. This model is called *Med*.

Both the medication and diagnosis models were trained with L2-regularized logis-

tic regression with a regularization weight of 1.0 (see section 4.4).

5.1.3 Windowed Diagnosis and Medication Models

We mentioned in section 5.1.1 that instead of considering each diagnosis as a single binary feature, we can also take into account the time at which a patient received a diagnosis.

In order to take time into account, we defined windows of time. The specific windows we used were [more than 2 years before the prediction time cutoff], [2 years to 1 year before the prediction time cutoff], and [less than 1 year before the prediction time cutoff]. We picked these particular windows to capture data over a large enough range of time to see trends, but close enough to the cancer diagnosis to reflect changes that were occurring due to the presence of undiagnosed pancreatic cancer. These windows are defined in relation to the prediction time cutoff date, not the cancer diagnosis date. We derived 3 binary features from each ICD code representing whether or not the patient had ever had a diagnosis of that ICD code within that window or earlier. For example, if a patient received an "E210" diagnosis 1.5 years before the prediction cutoff date, the features would be 0, 1, and 1, respectively.

We used the procedure above to derive binary features for the diagnosis and medication data, creating two separate models. Both were trained with L2-regularized logistic regression with a regularization weight of 1.0 (see 4.4). We will refer to these models as *Diag-Win* and *Med-Win*.

5.2 Lab Tests

Our EHR dataset also contains lab test data. For example, a patient might be administered a blood panel test while in the hospital. Their records would then contain values for platelet count and hemoglobin density in blood. The lab tests are categorized using LOINC codes [35].

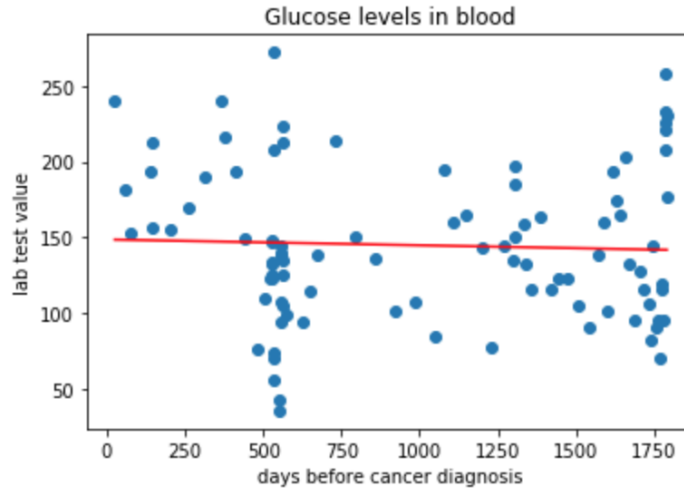


Figure 5-2: Levels of glucose in blood relative to cancer diagnosis, with fitted line

5.2.1 Lab Test Feature Derivation

As with diagnosis codes, we could simply consider whether or not a patient has ever had a certain lab test performed on them. However, we have significantly more data than in the diagnosis case, as we have various values returned by the lab tests that we can consider as features for our models. There are various ways of combining the many data points for a single lab test into a feature. One of the simplest is aggregation features, such as the mean or range of the data points over time. As with diagnoses, we might also want to consider the time at which the lab tests were performed. One way to combine time and lab test values into a single feature is to perform a linear regression of lab test values as a function of time relative to cancer diagnosis. We could then extract features from the fitted linear regression, such as the slope, intercept, and R^2 value.

As an example of what the linear regression analysis might look like for a single LOINC code, consider figure 5-2, which shows glucose levels in blood for a single cancer patient over time with a fitted line. The line is a linear regression over the glucose levels with respect to time before cancer diagnosis.

As with our diagnosis models, we may also want to consider windows of time for each lab test. As an example of how windows would work with features from linear regression, consider figure 5-3, which shows glucose levels in blood for the same cancer

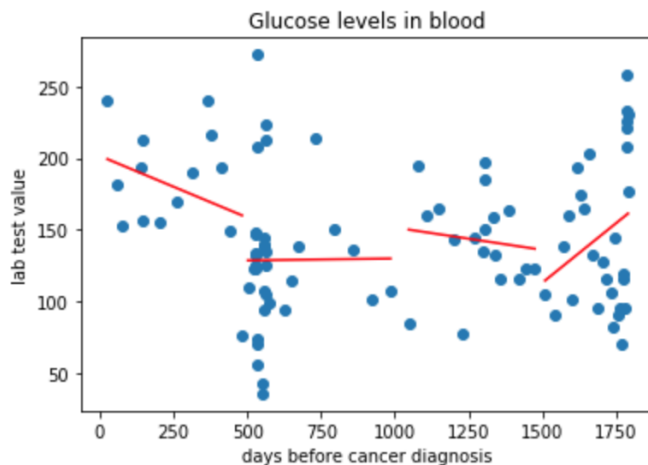


Figure 5-3: Levels of glucose in blood relative to cancer diagnosis, with fitted lines per window

patient as in figure 5-2, but with the linear regression performed over four windows. This graph represents a particular patient with one specific LOINC code, and is not necessarily indicative of what trends we expect to see in a pancreatic cancer patient. Rather, it is a demonstration of how we derive features for our model.

Another important consideration is that the choice of windows matters. Figure 5-4 shows the same graph from 5-3, but with a different set of windows. Comparing the graphs, we can see that the choice of windows can cause the trend of the linear regression to flip entirely. To choose our windows, we must decide how many windows to use and where to place them, which leads to an exponential explosion in choices. See section 8.1 in the future work chapter for a sketch of an algorithm to learn windows.

5.2.2 Lab Test Models

Our lab test models considered all LOINC codes that appeared at least 100 times in the dataset. This left us with 3123 unique LOINC codes. We considered four different features over lab tests: mean, slope of linear regression, intercept of linear regression, and R^2 value of linear regression. For an explanation of how the linear regression was computed, see section 5.2.1. We developed three models with these features: 1) using only the mean as a feature, 2) using the slope, intercept, and R^2 value of linear

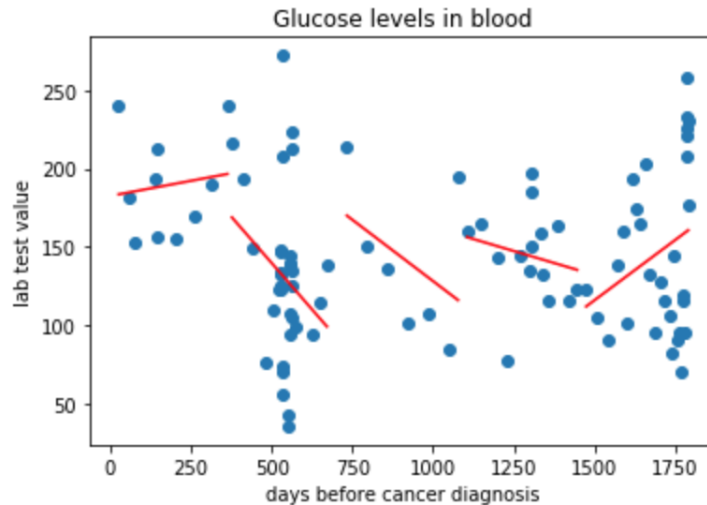


Figure 5-4: Levels of glucose in blood relative to cancer diagnosis, with fitted lines per window

regression as features, and 3) using the mean as well as the slope, intercept, and R^2 value of linear regression as features.

We normalized all of these numeric features by subtracting the mean value and dividing the standard deviation (that is, we used z-scores). For patients who did not have any values for a certain lab test, we used the mean of other patients' values for that lab test. For each lab test, we also included a binary feature that was 1 if the patient had any data for that lab test and 0 otherwise. Note that the mean and standard deviation were always calculated with reference to the training data, not the test data. This was done to prevent data leakage (see section 3.1.3).

The three models described above were trained with L1-regularized logistic regression with a regularization weight of 1.0. We chose to use L1 regularization for these models because they had a large number of features, and we wanted to prune the feature vector to reduce overfitting (see 4.4). We will refer to these models as *Mean*, *LinReg*, and *Mean-LinReg*.

5.2.3 Windowed Lab Test Models

We mentioned in section 5.2.1 that we can also consider lab test features within windows, as we did for our diagnosis and medication data. For our windowed lab

test model, we considered the same three sets of features as in section 5.2.2, but each feature was computed separately for each time window. The windows were the same as in the windowed diagnosis and medication models: [more than 2 years before the prediction time cutoff], [2 years to 1 year before the prediction time cutoff], and [less than 1 year before the prediction time cutoff]. See section 5.1.3 for rationale on choosing these windows. The models were again trained with L1-regularized logistic regression with a regularization weight of 1.0. As with the non-windowed lab test models, we chose to use L1 regularization to prune the feature vector. We will refer to these models as *Mean-Win*, *LinReg-Win*, and *Mean-LinReg-Win*.

Chapter 6

Implementation

All code written for this project can be downloaded from [55].

6.1 Logistic Regression

For all of our models, we used the logistic regression implementation from Scikit-learn [67].

6.1.1 Regularization and Solvers

All diagnosis and medication models were run with L2 regularization, while lab test models were run with L1 regularization. We used the default regularization penalty of 1.0. The solvers used varied depending on the model.

For the diagnosis and medication models, we used the 'sag' solver because it supports L2 regularization and is typically able to converge faster than other solvers for large datasets such as ours. This solver uses Stochastic Average Gradient (SAG) descent. SAG combines features from both the typical gradient descent method, in which a full gradient (FG) is computed for each data point on each time step, and the stochastic gradient descent (SGD) method, which computes a gradient for a single, randomly selected data point on each time step [75]. SGD has a lower iteration cost than the FG method, making it better for larger datasets, but it does not have the

convergence guarantees of the FG method. Like SGD, SAG only computes a single gradient on each timestep. However, like FG, it also incorporates a gradient for each data point on each timestep. It does so by using the last gradient value computed for that data point. By combining these two methods, SAG achieves faster convergence than SGD with the same minimal computation as SGD.

The 'sag' solver is recommended for use on datasets with a large number of samples and large number of features. Therefore, it works well for our diagnosis and medication models.

The 'sag' solver does not support L1 regularization, so we used the 'liblinear' solver for our lab test models. The liblinear solver uses a coordinate descent (CD) algorithm. The CD algorithm is iterative; it minimizes the objective function with respect to one dimension of the weight vector at each timestep, while holding the rest of the weight vector constant [87].

6.1.2 Other Parameters

We increased the maximum number of iterations for the Scikit-learn logistic regression model to 1000. The default is 100, but the model sometimes could not converge with only 100 iterations. For all other parameters, we used the defaults.

6.1.3 Input data types

Scikit-learn recommends that the input to the logistic regression model be in the form of CSR (compressed sparse row format) matrices containing 64-bit floats for optimal performance. Following this advice, we used a sparse CSR matrix with 64-bit floats. We used sparse datatypes to reduce memory usage because our datasets had many 0 values (especially for binary features). For example, the dataframe for *Diag* that was passed into our logistic regression model was only 0.12% ones.

6.2 Other Software

All data pre-processing and models were written in Python, using open-source libraries. As mentioned in the previous section, we used Scikit-learn for modeling. For data preprocessing, we mostly used Pandas [56]. As Pandas is rather memory-intensive, we used NumPy instead where necessary but also convenient [65]. We used Matplotlib and Seaborn for data visualization [82, 36].

6.3 Machine

All models were run on an AWS EC2 instance with 60 GB of RAM and 300 GB of storage.

6.4 Data Manipulation Optimizations

We had to make various optimizations to allow the models to run in a reasonable amount of time. This was especially important because our datasets were quite large; we used 10-fold cross-validation, which is expensive because each model needs to be fitted and validated 10 times; and we had many different models that we needed to run.

6.4.1 Data Compression

Our machine had 300 GB of storage, but the fully uncompressed EHR datasets with all patients exceeded 400 GB. Our first data optimization step consisted of dictionary-encoding patient hashes to reduce the size of these datasets. That is, we created a mapping of patient hashes to integers and replaced each reference of a patient hash in our dataset with its corresponding integer.

The EHR datasets provided by TriNetX use patient hashes of length 40 to protect patient identity. These patient hashes are used to link different records from the same patient among the different files, such as diagnosis data and medication data.

To reduce the data size, we encoded these 40-character hashes as 32-bit integers. After replacing all of the hashes with integers, the size of the total dataset decreased to 134 GB (about 32% of the original size) and could fit comfortably on our machine.

6.4.2 Data Chunking

We used the Pandas pivot table option to create one-hot encodings, e.g. for the diagnosis model, to indicate whether or not a patient had ever had a particular diagnosis. However, creating a pivot table on the entire dataset at once was infeasible. Pandas pivot tables require large amounts of memory, so trying to create a pivot table from the entire dataset caused the script to crash with a memory error.

To solve this memory issue, we split the dataset into chunks of 5000 patients each and ran the pivot table separately on each dataset. We then concatenated the ensuing pivot tables to create a pivot table for the full dataset. All pivot tables were reindexed to include all diagnosis codes, regardless of whether or not any patient present in that subset had the diagnosis code, to ensure that all pivot tables would have the same columns.

6.4.3 Sparse Datatypes

Pandas provides sparse datatypes that are useful for efficiently storing sparse data. These datatypes "compress" dataframes in such a way that any values in the dataframe matching a specified value (which could be NaNs, 0s, or other user-specified values) are omitted from the dataframe. For dataframes that are "sparse" in this sense, using Pandas sparse datatypes can significantly reduce the memory usage of the dataframe.

Because our models used dataframes with a number of rows on the order of a hundred thousand, we had to be very conscious of our memory usage. These dataframes were sparse: for example, our diagnosis model used a one-hot encoding on the dataset to indicate whether or not a patient had ever had a particular diagnosis. This means that every value in the dataframe was either 1 or 0, so encoding the dataset with 0 as the sparse value could significantly reduce the memory footprint of the dataset. As

mentioned in section 6.1.3, the dataframe for *Diag* that was passed into our logistic regression model was only 0.12% ones.

6.5 Data Manipulation Results

In this section, we present results demonstrating the efficacy of our data manipulation optimizations from section 6.4.

6.5.1 Data Compression

As mentioned in section 6.4.1, after using our data compression methods, the size of the total set decreased to 32% of its original size. This allowed us to fit our data on our machine.

6.5.2 Data Chunking

Data chunking allowed us to run our models without raising out of memory errors.

6.5.3 Sparse Datatypes

To exhibit how memory usage and timing differ for sparse and non-sparse datatypes, we constructed smaller demonstration dataframes. We did not run these experiments on the full dataset because non-sparse datatypes would raise out of memory errors on the full dataset.

All of our demonstration dataframes were created with 20 columns, and the number of rows was varied from 100 to 5000 to show the effect of these optimizations when varying scale. Three-quarters of the dataframe values were zeroes, while the other one-quarter were ones.

Figure 6-1 shows how memory usage increased with respect to the number of rows in our demonstration dataframes. The blue line shows memory usage for the non-sparse dataframe, while the orange line shows memory usage for the dataframe with sparse datatypes. The graph shows that memory usage is significantly lower for the

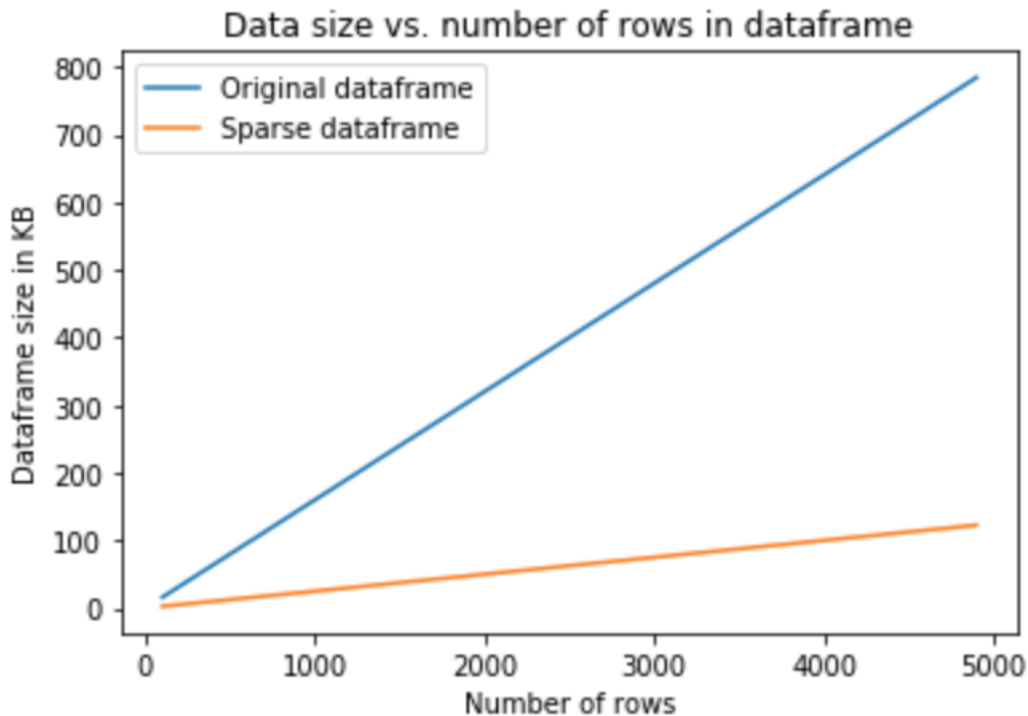


Figure 6-1: Memory usage vs. dataframe size

sparse dataframe than for the non-sparse dataframe, and this difference increases as the number of rows grows. These dataframes are much smaller than the ones we use in our models, but we assume that the difference in memory usage for the smaller demonstration dataframes also applies to our larger full dataset.

Using less memory is important because it reduces the chance that the model will fail due to memory issues. It also reduces the amount of time needed to perform operations on the dataframe. One common operation we performed in our data pre-processing was selecting a subset of rows from a dataframe based on patient ID. Figure 6-2 shows the amount of time needed to perform a row selection operation on our demonstration dataframes versus the number of rows. We used the same setup that was used to measure memory usage in 6-1, but also added a column containing randomly generated "patient IDs" represented by integers between 10000 and 20000. We then randomly sampled one-fifth of the patients to form our patient subset. The dataframe row selection operation consisted of selecting all rows from the dataframe where the patient ID was in the patient subset list. Figure 6-2 shows that the pandas

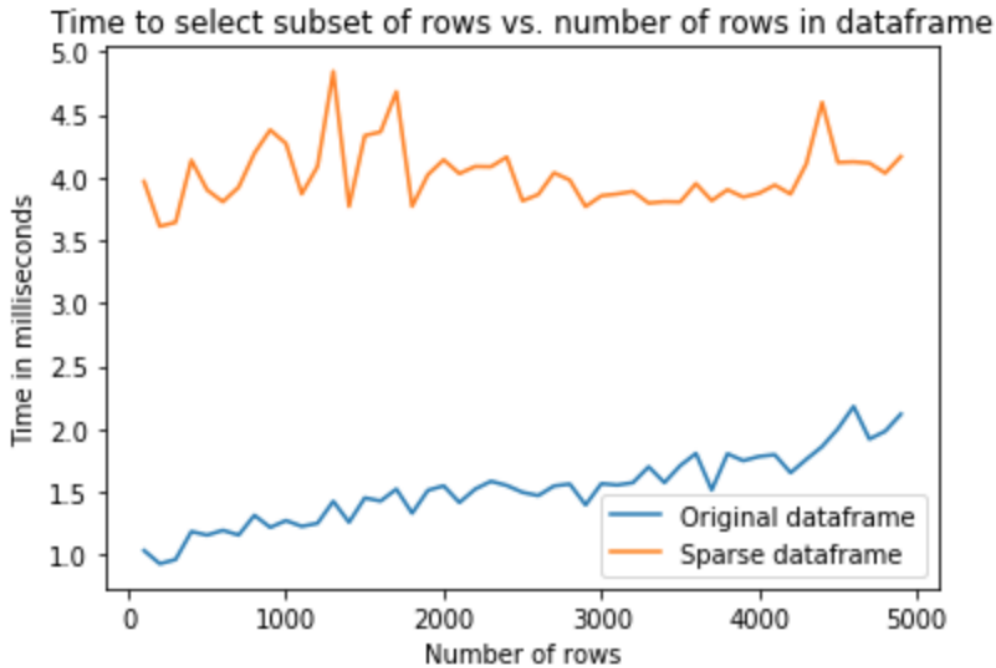


Figure 6-2: Time vs. dataframe size

row selection operation takes significantly less time with sparse datatypes than with non-sparse datatypes.

Figure 6-3 shows the amount of time needed to fit a logistic regression model to our demonstration dataframes. The timing difference between sparse and non-sparse datatypes here is not as significant as in 6-2. This is most likely because scikit-learn coerces the Pandas dataframe type passed in to a numpy array before performing logistic regression [67]. This representation takes up much less memory than Pandas dataframes, so the difference in memory usage becomes less significant, and thus the timing difference also becomes less significant. Again, we assume that the timing difference on our demonstration dataframes will also apply to our full dataset.

Considering the timing and memory results presented above, we chose to use sparse datatypes for all of our models. This allowed us to run our models faster and without raising out of memory errors.

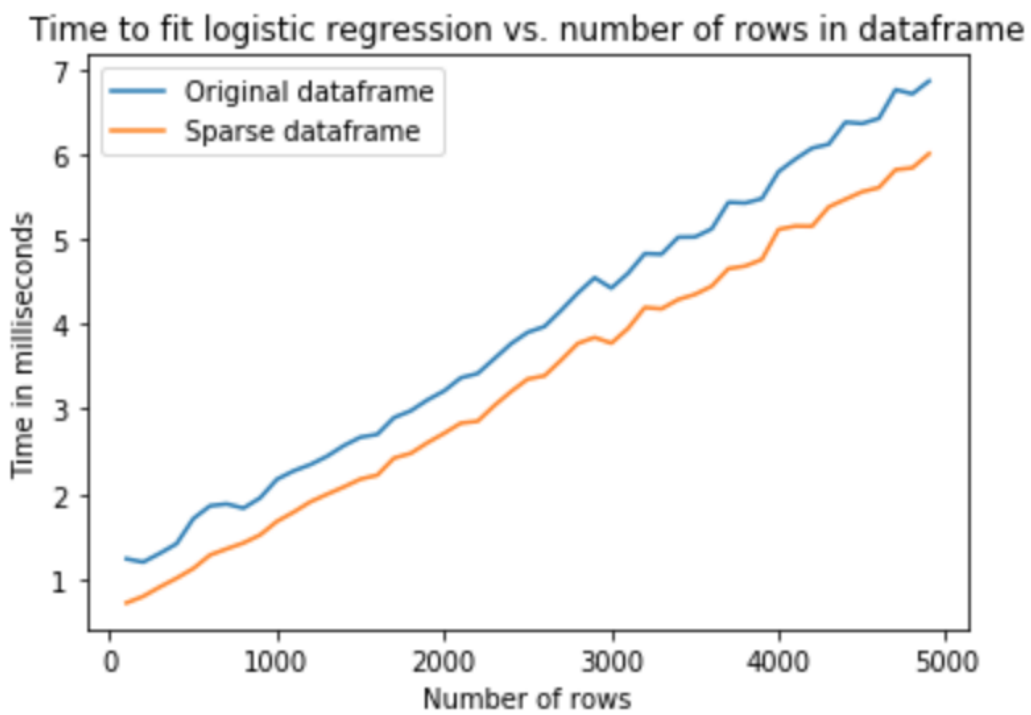


Figure 6-3: Time vs. dataframe size

Chapter 7

Results

7.1 Metrics

Figures 7-1 and 7-2 show the AUROC and APS scores on all 10 models on training data. Figures 7-3 and 7-4 show the AUROC and APS scores on all 10 models on test data. All figures show 95% confidence intervals. All metrics were computed using 10-fold cross-validation (see section 4.6). The exact values for all models can be found in tables A.1 through A.2 in Appendix B. We'll now delve into comparisons between relevant subsets of models.

7.1.1 Diagnosis and Medication Models

In this section we'll consider *Diag* and *Med*, which both use simple binary features.

Figure 7-5 shows the AUROC for *Diag* and *Med*. On the training data, *Diag* obtained AUROCs of 0.99, 0.99 and 0.99. The AUROC for *Med* was 0.86, 0.86, and 0.86.

On the test data, *Diag* had AUROCs of 0.74, 0.73 and 0.73, and *Med* had AUROCs of 0.75, 0.76, and 0.75.

Comparing these numbers, we can see that *Diag* outperforms *Med* on the training data, but their performance is roughly the same on the test data. *Med* suffers a smaller drop from training AUROC to test AUROC than *Diag* does, which may

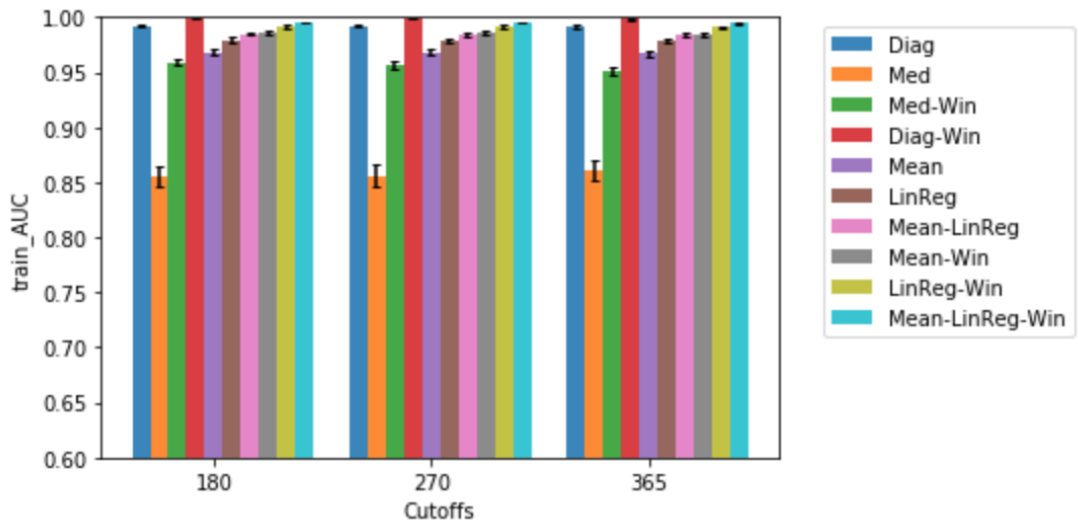


Figure 7-1: Training Data: Area-Under-ROC (AUROC)

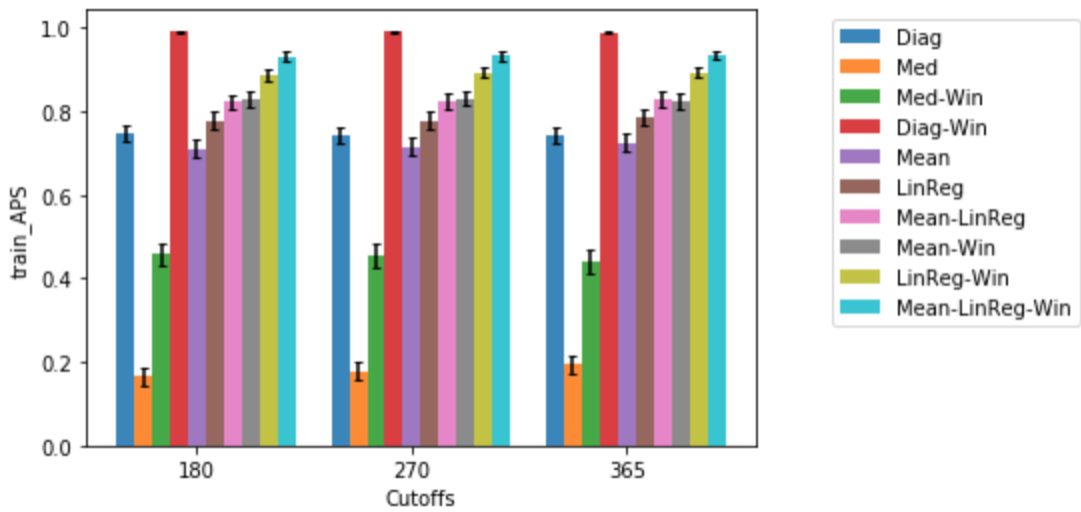


Figure 7-2: Training Data: Average Precision Score (APS)

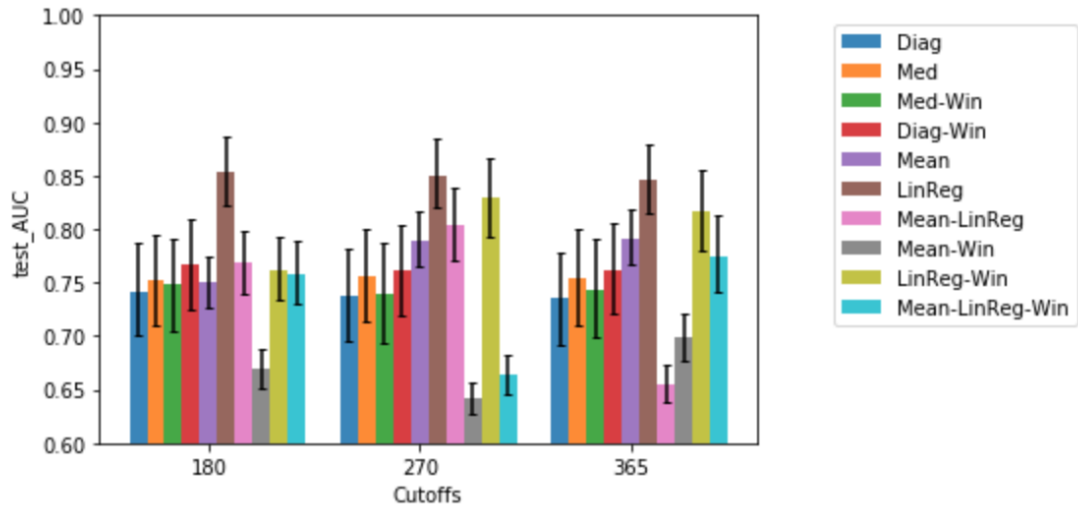


Figure 7-3: Test Data: Area-Under-ROC (AUROC)

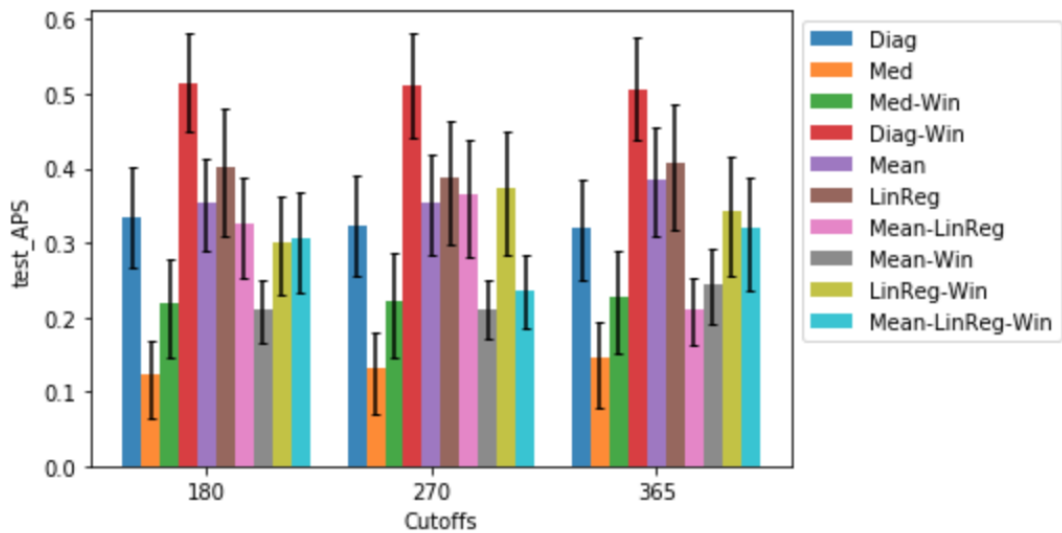
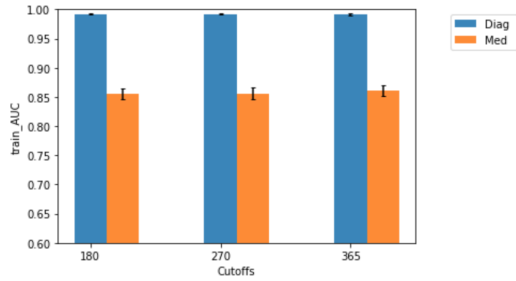
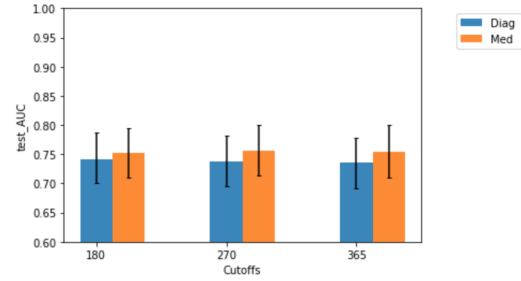


Figure 7-4: Test Data: Average Precision Score (APS)

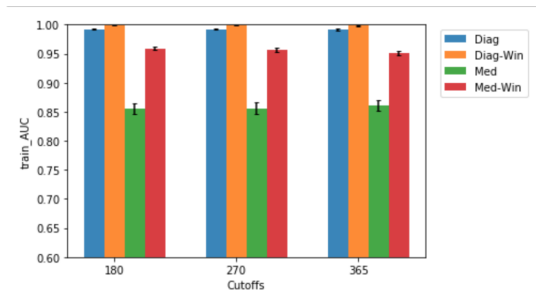


(a) Training Data: Area-Under-ROC

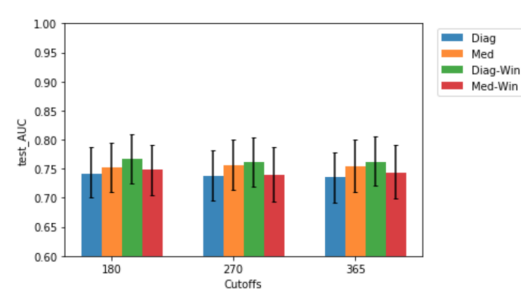


(b) Test Data: Area-Under-ROC

Figure 7-5: AUROC on Diag and Med



(a) Training Data: Area-Under-ROC



(b) Test Data: Area-Under-ROC

Figure 7-6: AUROC on Diag, Med, Diag-Win, and Med-Win

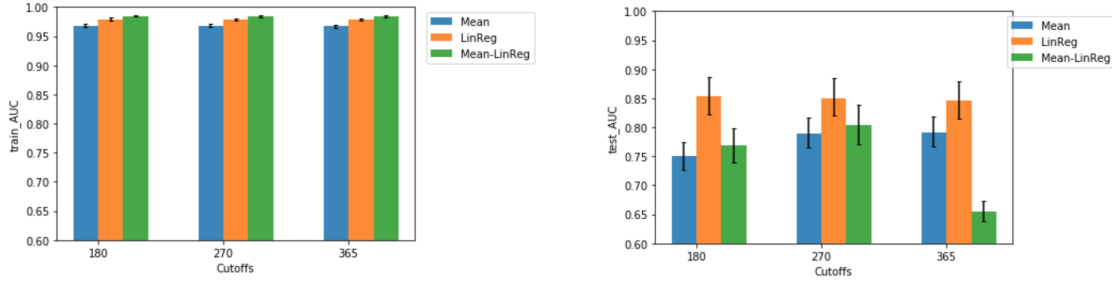
indicate that *Diag* is overfitting.

7.1.2 Windowed Diagnosis and Medication Models

We now consider how *Diag* and *Win* compare to *Diag-Win* and *Med-Win*.

Figure 7-6 shows the AUROC for these models. On the training data, *Diag-Win* had AUROCs of 0.99, 0.99, and 0.99; *Med-Win* had AUROCs of 0.96, 0.96, and 0.95. On the test data, *Diag-Win* had AUROCs of 0.77, 0.76, and 0.76; *Med-Win* had AUROCs of 0.75, 0.74, and 0.74.

Comparing these with the AUROCs for *Diag* and *Med* (7.1.1), we can see that both windowed models outperform their non-windowed counterparts on the training data. However, on the test data, the differences become negligible, and *Med-Win* in fact performs worse than *Med*. *Med-Win*'s drop in performance from training data to test data could indicate overfitting.



(a) Training Data: Area-Under-ROC

(b) Test Data: Area-Under-ROC

Figure 7-7: AUROC on Lab Test Models

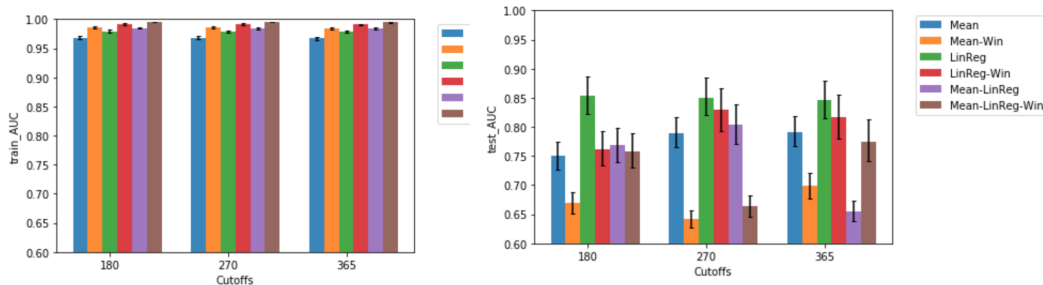
7.1.3 Lab Tests

We now consider the non-windowed lab test models: *Mean*, *LinReg*, and *Mean-LinReg*. See figure 7-7 for barplots.

On the training data, *Mean* had AUROCs of 0.97, 0.97, and 0.97; *LinReg* had AUROCs of 0.98, 0.98, and 0.98; and *Mean-LinReg* had AUROCs of 0.98, 0.98 and 0.98. On the test data, *Mean* had AUROCs of 0.75, 0.79, and 0.79; *LinReg* had AUROCs of 0.85, 0.85, and 0.85; and *Mean-LinReg* had AUROCs of 0.77, 0.80, and 0.66.

From the above, we can see that the lab test models performed about the same on the training data set, but *LinReg* stood out as the best performer on the test set, with AUROCs of about 0.85 for every cutoff. *Mean* and *Mean-LinReg* performed similarly to each other, but worse than *LinReg*.

Mean-LinReg includes all of the features from *Mean* and all of the features from *LinReg*, so the fact that it performs worse than *LinReg* alone is somewhat surprising. This could be because the features in *LinReg* (the slope, R^2 value, and intercept) already contain information similar to that provided by the mean. Including the mean on top of these features could then cause overfitting and worse performance on the test set.



(a) Training Data: Area-Under-ROC

(b) Test Data: Area-Under-ROC

Figure 7-8: AUROC on Lab Test Models with Windowing

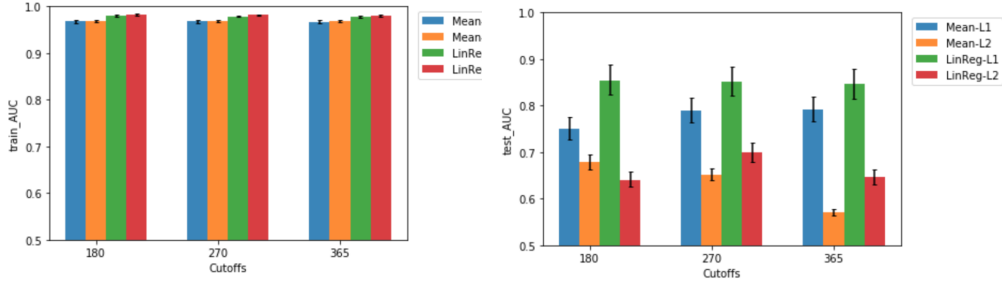
7.1.4 Windowed Lab Tests

We'll now consider how the lab tests compare with their windowed counterparts. See figure 7-8 for barplots.

On the training data, *Mean-Win* had AUROCs of 0.99, 0.99, and 0.98; *LinReg-Win* had AUROCs of 0.99, 0.99, and 0.99; and *Mean-LinReg-Win* had AUROCs of 0.99, 0.99, and 0.99. On the test data, *Mean-Win* had AUROCs of 0.67, 0.64, and 0.70; *LinReg-Win* had AUROCs of 0.76, 0.83, and 0.82; and *Mean-LinReg-Win* had AUROCs of 0.76, 0.66, and 0.78.

Comparing these values to those from 7.1.3, we can see that the windowed models perform about equivalently to the non-windowed models on the training set, with AUROCs close to 1. However, the windowed models consistently perform worse than the non-windowed models on the test set. For *Mean-Win* the results are much worse than for *Mean* (0.67 vs. 0.75 for a 180-day cutoff), while for *LinReg-Win* the results are close for some cutoffs (0.83 vs. 0.85 for a 270-day cutoff). The only case where a windowed model outperforms a non-windowed model is *Mean-LinReg-Win* with a 365-day cutoff (0.78 vs. 0.66 for *Mean-LinReg*).

The fact that windowed models performed worse than non-windowed models was disappointing. As mentioned in 5.2.1, we hoped that windowing would make the results of linear regression more meaningful, as they would show more nuanced trends and how these trends shifted over time. As we also mentioned in 5.2.1, changing the windows can lead to significantly different trends in the linear regression. It's possible



(a) Training Data: Area-Under-ROC

(b) Test Data: Area-Under-ROC

Figure 7-9: AUROC on Lab Test Models, L2 vs. L1 regularization

that for some choice of windows, the model would perform much better, but for the particular choice of windows we used for our model, the trends were not informative.

7.1.5 L2 vs L1 regularization

All lab test models were run with L1 regularization because we wanted to prune the size of the initially large feature set (see section 4.4). We had 3123 unique LOINC codes, and for each LOINC code we might generate up to 4 features (e.g., mean, slope, intercept, and R^2 value for *Mean-LinReg*). In addition, the windowed models had three times as many features as the non-windowed models, because each feature was computed for each of the three windows. This put us at almost 40,000 features initially for *Mean-LinReg-Win*.

Figure 7-9 shows AUROCs for *Mean* and *LinReg* with either L1 or L2 regularization. While both types of regularization produced similar results on the training data, L1 regularization significantly outperformed L2 regularization for both models. For example, *Mean-L1* had AUROCs of 0.75, 0.79, and 0.79 on the test set, while *Mean-L2* had AUROCs of 0.68, 0.65, and 0.57. These results aligned with our expectations that L1-regularized models would generalize better to the test set than L2-regularized models would.

7.2 Discussion

We have implemented a set of models that use readily available data from EHRs to predict pancreatic cancer risk up to 365 days in advance of pancreatic cancer diagnosis. These models leveraged binary data from diagnoses and medications, as well as longitudinal data from lab tests. In particular, *LinReg* achieved an AUROC of 0.85 on the test dataset with a 365-day lead time, while *Mean* achieved an AUROC of 0.79 with a 365-day lead time. *Diag* and *Med* also achieved AUROCs of about 0.75 with the same prediction cutoff.

Unlike previous studies that focused on hand-crafted feature sets [2], we chose to use all features available to us in the EHR datasets, with some filters to remove features that rarely occurred. We found that this approach yielded higher AUROC scores than using select features when applied to our EHR dataset.

Many previous studies have focused on cleaner sets of data, such as the cohort studies found in the PanScan Consortium [4]. We chose to use EHR datasets instead because they contain data on a wider variety of features and are more directly applicable to clinical settings. Despite the many known challenges of working with EHRs, our models have shown that it is feasible to apply standard machine learning techniques to learn risk models from EHRs.

7.2.1 Limitations

Our pancreatic cancer risk models have some limitations which must be addressed before the models can be applied in practice. Firstly, while our EHR data does cover multiple hospitals, these hospitals are all restricted to the U.S. Therefore, our models may not generalize to non-U.S. populations. Secondly, our EHR model is retrospective: that is, the outcomes have already occurred, and we are using past data to learn the risk of the outcome. In order to put the model in practice clinically, we would need to validate the model with a prospective study, where we test the model on current patients who have not yet developed pancreatic cancer and see if the model correctly predicts which ones will develop pancreatic cancer in the future.

Thirdly, our model contains no information on what stage the patient will be in if we detect their cancer 180, 270, or 365 days before a doctor would typically diagnose them. We hope that the cancer would be detected early enough to be resectable, but our experiments have not addressed this point directly. We would need to test this model in practice in order to know if the cancer would be detected early enough to be treatable. And finally, as we mentioned in section 5.2.1, the choice of windows for the model matters. We manually picked a set of windows for our model, but it might not have been the choice of windows that would optimize the amount of information learned from trends in lab test data. The model could be even more effective than it currently is if it had a mechanism to learn windows.

7.2.2 High Level Takeaways

Our best risk model, which was based on longitudinal lab test data, achieved an AUROC of 0.85 with a 365-day lead time. We have shown that using longitudinal lab test data can provide a significant improvement over models solely based on diagnostic and medication data.

However, despite model improvements, we highlight that further improvement in discrimination is necessary for use as a general screening tool. Currently, at a specificity of 99%, our best model has a sensitivity of 10.2%. This means that to detect pancreatic cancer early in 20 patients, assuming that the incidence rate of pancreatic cancer in the U.S. is 13 out of 100,000 [37], we would need to screen at least 1.5 million patients. On our test data, our model obtained a positive predictive value (PPV) of 47.3%. However, this reflects our case/control matching ratio (1 cancer to 25 control patients, see section 4.2.3), as PPV is influenced by disease prevalence. In contrast to PPV, sensitivity and specificity are intrinsic model characteristics and do not depend on the prevalence of cancer in the population. Assuming that the incidence rate of pancreatic cancer in the U.S. is 13 out of 100,000 [37], our model's implied PPV (based on the obtained specificity/sensitivity) is 0.13%. The incidence rate of pancreatic cancer among the U.S. population over the age of 65 is about 70 out of 100,000 [2]; this subpopulation is more reflective of our dataset, since we only

considered patients between the ages of 60 and 80. Assuming this higher incidence rate, our model's implied PPV on the U.S. population over the age of 65 is 0.70%. As pointed out by Rulyak et al. and Baecker et al., a general screening tool would likely require high specificity (approximately 99%) and high sensitivity (approximately 84%) for wide deployment [71, 2]. In this regard, improving model discrimination remains a goal for future work.

However, our results do show evidence that use of this model as a pre-screening tool to identify high-risk patients holds promise. In particular, our best performing model (*LinReg*) points to an implied PPV of 0.70% on the U.S. population over the age of 65, compared to a PPV of approximately 2.8% obtained by high-risk screening guidelines based on familial and genetic factors [15]. While our model is less predictive than one based on familial and genetic factors, our model could still be applied passively as a first filter. Note that even with a PPV of 2.8%, it is considered worthwhile to screen high-risk patients with endoscopic ultrasonography (EUS), MRI/magnetic resonance cholangiopancreatography, and endoscopic retrograde cholangiopancreatography [15]. The willingness of the medical establishment to engage in these expensive and complicated procedures, even in the presence of a PPV as small as the PPV of 2.8% based on familial and genetic factors, highlights the substantial risk posed by pancreatic cancer and the significant effort that the medical establishment is willing to expend in the hope of obtaining an early diagnosis for this otherwise likely fatal disease.

Chapter 8

Future Work

8.1 Learning Windows

As mentioned in section 5.2.1, the choice of windows in our windowed models makes a significant difference in the features extracted from longitudinal data. Because there are many different possible choices of windows (parameters include the size of each window, number of windows, where to place each window, etc.), it is difficult to manually pick windows and expensive to empirically validate window choices to pick the best.

We now propose some methods for learning windows, which are left for future work.

8.1.1 Pre-processing

Instead of having our model learn the windows itself, we could use statistical methods to determine windows before running our model. One promising method would be detecting turning points in the windows by looking for significant changes in the mean, variance, or slope of a line.

8.1.2 Multiple Classifiers

Instead of pre-generating windows, we could take a multiple-classifier approach in which we train many classifiers on differently-windowed datasets, each derived from the main dataset by applying different sets of windows. We could then combine the results of these weak classifiers to form a final prediction for each patient. We will now outline a sketch of the model.

For this model, we will first randomly generate many sets of windows. We will do so by first randomly sampling a number of windows between 1 and 10, and then by continually sampling the difference in days (between 30 and 365) between each consecutive window. For example, let's say we randomly select the number of windows as 4. This means we will define 4 windows, so we need to select three numbers to represent the size of each window. We'll then sample three differences in days: let's say we randomly select 60, 180, and 360. This means that the windows will be [60 days or less before prediction time cutoff], [60 to $60+180=240$ days before prediction time cutoff], [240 to $240+360=600$ days before prediction time cutoff], [more than 600 days before prediction time cutoff].

We will compute the features described in section 5.2.2 per window. This will leave us with many different datasets, each produced with different windows. We will then train a weak classifier (using logistic regression) on each dataset. Each window for each lab test will be treated as a separate feature in the classifier.

We have two options for aggregating the decisions of the weak classifiers to make a final classification:

1. Perform a majority vote amongst the results from all of the weak classifiers.
2. Use each weak classifier to predict a probability that a patient has cancer. Train another classifier that uses the probabilities from each weak classifier to predict whether or not a patient will later develop pancreatic cancer.

8.2 Additional Longitudinal Features

Our EHR dataset included several forms of longitudinal data: lab tests, vital signs data, and more. There are many possible ways to tackle longitudinal data. We can use summary metrics (mean, median, etc.) on a single lab test over time; take ratios of different lab tests; shift vectors with respect to each other; calculate percentage differences for a series of values; etc.

Because longitudinal models are a fairly unexplored field for pancreatic cancer risk modeling, and even for risk modeling for other diseases from EHRs [10], we chose to focus mainly on summary metrics. Namely, we used the mean of lab test values and features from a linear regression over lab test values. Even with these fairly simple features, we were able to achieve an AUROC of 0.85 (see table A.1 for the *LinReg* model). This outperformed our models based on diagnosis and medication data.

Future work could include extending the lab test models to include more complicated features, such as ratios between different lab tests. Relationships between different lab tests could be early signs of a cancer diagnosis. For example, increasing glucose levels on their own are not normally a sign of concern, but increasing glucose levels combined with weight loss is a potential risk factor for pancreatic cancer [73]. Adding features to our model that take into account multiple lab tests, rather than summarizing a single lab test, could allow us to discover more relationships that are not already clinically known.

8.2.1 Challenges of Feature Engineering

Creating features from raw data is the part of modeling that generally requires the most human interaction, because it involves human intuition as to what features are most useful [41]. Feature selection is very important to the performance of models [20], so either humans must spend a large amount of effort generating good features or machine replacements must be able to engineer features as well as humans.

The search space for features is often huge. For example, consider our problem of generating features for longitudinal lab test data. There are several types of features

we might want to generate: 1) aggregation features such as mean, median, and range; 2) shift features such as shifting a vector; 3) difference features such as percentage differences for one lab test; 4) relational features such as ratios between different lab tests; and more. We can combine these types of features in sequences, producing infinitely many sequences of features. To discover which feature sequences perform the best, we need to empirically validate. This requires that we train a model with the new features and test it, which is expensive and needs to be run many times.

8.2.2 Languages for Feature Engineering

We can partially automate feature engineering by defining a language for feature engineering and allowing a model to explore different compositions of operations to yield new data features. A language would consist of raw features from the data (e.g., a patient had this diagnosis on this date), operators, and a grammar for composing the raw features and operators. Operators include the types of features discussed in 8.2.1, such as a MEAN operation that takes the mean of a vector, a SHIFT operation on one vector that shifts the vector by one position, a PERCENT_DIFFERENCE feature that calculates the percentage difference between each pair of values in a vector, and a DIVIDE operation that operates on two vectors by performing an element-wise division. The language defines which operators can be performed in sequence. For example, we cannot perform a SHIFT operation after a MEAN operation because the MEAN operation collapses the vector to a single point, while a SHIFT operation requires a vector of at least length two.

This language construction would allow a program to automatically generate sequences of features. However, the search space would still be very large, and would require a large amount of machine computation or a clever search technique to fully search.

8.2.3 Machine Feature Synthesis

Kanter et al. developed a deep feature synthesis algorithm that differentiated between functions at the entity level and relational level, and defined features at each level [41]. Their algorithm automatically detects features that can be synthesized in a proper order, following a language of what types of features can be stacked on other types of features. They also provide methods of restricting the search space by specifying a maximum depth and allowing for categorical filters on functions. Finally, Kanter et al. provide a machine learning pipeline that automatically generates features on a relational database using the deep feature synthesis algorithm and autotunes a machine learning "pathway" to use these features [41].

8.3 LSTMs

Even with the automatic feature generation methods discussed in sections 8.2.2 and 8.2.3, exploring the search space generated by these methods is still time-consuming and expensive. We might like to have a model that implicitly generates sets of features.

LSTMs (and neural nets in general) can implicitly generate feature sequences, where the language for sequences consists of affine transformations (such as matrix multiplication), non-linear transformations (such as activation functions), and more [79]. LSTMs are a specialized type of Recurrent Neural Network (RNN); RNNs are a type of neural net that operate over sequences. This makes them particularly useful for analysis of longitudinal data. Additional work would be required to fit our data to LSTMs, especially since our lab test data is collected at irregular intervals and we would need to find a way to align the sequences and account for missing data. However, there are some precedents for using LSTMs to analyze EHRs [51].

Chapter 9

Conclusion

Pancreatic cancer is typically not diagnosed until the pancreatic tumor is unresectable, leading to a dismally low five-year survival rate. Better risk models are needed to identify patients who are at high risk for pancreatic cancer in advance of when a diagnosis would typically be made. In this thesis, we presented several novel machine learning models that analyze Electronic Health Record data to predict a patient's risk of developing pancreatic cancer. Our models are capable of identifying a future cancer patient 365 days before the cancer is actually diagnosed. Our hope is that this would allow hospitals to identify patients at high risk of pancreatic cancer and bring them in for screening.

Appendix A

Tables

Table A.1: AUROC and APS scores for all models on training and test data

Model Name	Cutoff	Train AUC	Test AUC	Train APS	Test APS
Diag	180	0.992	0.742	0.747	0.335
Diag	270	0.992	0.737	0.742	0.324
Diag	365	0.991	0.735	0.744	0.319
Med	180	0.856	0.752	0.166	0.124
Med	270	0.856	0.756	0.178	0.132
Med	365	0.86	0.754	0.194	0.145
Med-Win	180	0.959	0.748	0.458	0.217
Med-Win	270	0.957	0.74	0.457	0.222
Med-Win	365	0.951	0.744	0.44	0.227
Diag-Win	180	0.999	0.767	0.991	0.515
Diag-Win	270	0.998	0.761	0.99	0.512
Diag-Win	365	0.998	0.762	0.989	0.507
Mean	180	0.968	0.751	0.711	0.352
Mean	270	0.968	0.79	0.716	0.354
Mean	365	0.967	0.791	0.725	0.385
LinReg	180	0.979	0.854	0.777	0.401
LinReg	270	0.979	0.851	0.778	0.387
LinReg	365	0.978	0.846	0.786	0.406
Mean-LinReg	180	0.985	0.769	0.823	0.324
Mean-LinReg	270	0.984	0.803	0.822	0.365
Mean-LinReg	365	0.984	0.655	0.828	0.211
Mean-Win	180	0.986	0.669	0.829	0.21
Mean-Win	270	0.985	0.642	0.83	0.212
Mean-Win	365	0.984	0.698	0.826	0.245
LinReg-Win	180	0.992	0.762	0.886	0.302
LinReg-Win	270	0.991	0.83	0.893	0.374
LinReg-Win	365	0.991	0.816	0.893	0.344
Mean-LinReg-Win	180	0.995	0.759	0.931	0.306
Mean-LinReg-Win	270	0.995	0.663	0.932	0.236
Mean-LinReg-Win	365	0.995	0.775	0.935	0.319

Table A.2: 95% Confidence Intervals for AUROC scores for all models on training and test data

Model Name	Cutoff	Train AUC CI	Test AUC CI
Diag	180	0.991 - 0.993	0.7 - 0.787
Diag	270	0.991 - 0.993	0.694 - 0.782
Diag	365	0.99 - 0.992	0.692 - 0.779
Med	180	0.847 - 0.865	0.71 - 0.795
Med	270	0.847 - 0.866	0.714 - 0.8
Med	365	0.851 - 0.87	0.71 - 0.799
Med-Win	180	0.956 - 0.962	0.704 - 0.791
Med-Win	270	0.953 - 0.96	0.694 - 0.788
Med-Win	365	0.947 - 0.954	0.698 - 0.791
Diag-Win	180	0.998 - 0.999	0.725 - 0.809
Diag-Win	270	0.998 - 0.999	0.72 - 0.803
Diag-Win	365	0.998 - 0.999	0.721 - 0.805
Mean	180	0.965 - 0.971	0.727 - 0.775
Mean	270	0.965 - 0.971	0.765 - 0.816
Mean	365	0.964 - 0.97	0.766 - 0.818
LinReg	180	0.977 - 0.982	0.823 - 0.887
LinReg	270	0.977 - 0.981	0.821 - 0.885
LinReg	365	0.976 - 0.98	0.814 - 0.88
Mean-LinReg	180	0.983 - 0.987	0.74 - 0.798
Mean-LinReg	270	0.982 - 0.986	0.771 - 0.838
Mean-LinReg	365	0.982 - 0.986	0.637 - 0.674
Mean-Win	180	0.984 - 0.988	0.651 - 0.688
Mean-Win	270	0.984 - 0.987	0.627 - 0.657
Mean-Win	365	0.982 - 0.986	0.677 - 0.721
LinReg-Win	180	0.99 - 0.993	0.734 - 0.794
LinReg-Win	270	0.99 - 0.993	0.793 - 0.867
LinReg-Win	365	0.989 - 0.992	0.779 - 0.856
Mean-LinReg-Win	180	0.994 - 0.996	0.73 - 0.789
Mean-LinReg-Win	270	0.994 - 0.996	0.645 - 0.682
Mean-LinReg-Win	365	0.994 - 0.996	0.741 - 0.812

Table A.3: 95% Confidence Intervals for APS scores for all models on training and test data

Model Name	Cutoff	Train APS CI	Test APS CI
Diag	180	0.729 - 0.766	0.268 - 0.403
Diag	270	0.723 - 0.761	0.255 - 0.391
Diag	365	0.724 - 0.764	0.251 - 0.385
Med	180	0.145 - 0.185	0.065 - 0.167
Med	270	0.155 - 0.198	0.07 - 0.178
Med	365	0.17 - 0.216	0.078 - 0.195
Med-Win	180	0.43 - 0.485	0.144 - 0.279
Med-Win	270	0.428 - 0.484	0.145 - 0.285
Med-Win	365	0.412 - 0.468	0.15 - 0.29
Diag-Win	180	0.988 - 0.994	0.449 - 0.582
Diag-Win	270	0.988 - 0.993	0.442 - 0.581
Diag-Win	365	0.986 - 0.992	0.437 - 0.577
Mean	180	0.689 - 0.734	0.29 - 0.413
Mean	270	0.694 - 0.738	0.283 - 0.419
Mean	365	0.702 - 0.748	0.308 - 0.454
LinReg	180	0.757 - 0.798	0.31 - 0.481
LinReg	270	0.759 - 0.799	0.297 - 0.463
LinReg	365	0.766 - 0.807	0.316 - 0.485
Mean-LinReg	180	0.805 - 0.841	0.251 - 0.386
Mean-LinReg	270	0.805 - 0.841	0.282 - 0.438
Mean-LinReg	365	0.81 - 0.847	0.164 - 0.251
Mean-Win	180	0.811 - 0.847	0.166 - 0.249
Mean-Win	270	0.813 - 0.848	0.17 - 0.25
Mean-Win	365	0.807 - 0.845	0.191 - 0.292
LinReg-Win	180	0.872 - 0.901	0.229 - 0.362
LinReg-Win	270	0.88 - 0.907	0.282 - 0.451
LinReg-Win	365	0.88 - 0.908	0.256 - 0.415
Mean-LinReg-Win	180	0.92 - 0.942	0.233 - 0.369
Mean-LinReg-Win	270	0.922 - 0.944	0.186 - 0.283
Mean-LinReg-Win	365	0.925 - 0.945	0.236 - 0.388

Bibliography

- [1] TriNetX home page. <https://www.trinetx.com/>. Accessed: 2019-12-09.
- [2] A. Baecker, S. Kim, H. Risch, T. Nuckols, B. Wu, A. Hendifar, S. Pandol, J. Pisegna, C. Jeon. Do changes in health reveal the possibility of undiagnosed pancreatic cancer? Development of a risk-prediction model based on healthcare claims data. *PloS one*, 14(6), 2019.
- [3] A. Kanno, A. Masamune, K. Hanada, et al. Multicenter study of early pancreatic cancer in Japan. *Pancreatology*, 18(1):61–67, 2018.
- [4] A. Klein, S. Lindstrom, J. Mendelsohn, E. Steplowski, A. Arslan, et al. An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. *PloS one*, 8(9), 2013.
- [5] Karim Abouelmehdi, Abderrahim Beni-Hessane, and Hayat Khaloufi. Big health-care data: preserving security and privacy. *Journal of Big Data*, 5(1), 2018.
- [6] KE Anderson, JD Potter, and TM Mack. *Cancer Epidemiology and Prevention*. Oxford University Press: New York, 1996.
- [7] Anne Marie Lennon, Christopher L. Wolfgang, Marcia Irene Canto, Alison P. Klein, Joseph M. Herman, Michael Goggins, Elliot K. Fishman, Ihab Kamel, Matthew J. Weiss, Luis A. Diaz, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein and Ralph H. Hruban. The early detection of pancreatic cancer: What will it take to diagnose and treat curable pancreatic neoplasia? *Cancer Research*, 74(13), 2014.
- [8] Javed A Aslam, Emine Yilmaz, and Virgiliu Pavlu. A geometric interpretation of r-precision and its correlation with average precision. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 573–574. ACM, 2005.
- [9] B. Boursi, B. Finkelman, B. Giantonio, K. Haynes, A. Rustgi, A. Rhim, R. Mantani, Y. Yang. A clinical prediction model to assess risk for pancreatic cancer among patients with new-onset diabetes. *Gastroenterology*, 152(4):840–850, 2017.

- [10] B. Goldstein, A. Navar, M. Pencina, J. Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, 2017.
- [11] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary use of ehr: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010:1, 2010.
- [12] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.
- [13] HA Burris III, Moore MJ, Andersen J, and et al. Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. *Journal of Clinical Oncology*, 15:2403–2413, 1997.
- [14] Alison Callahan and Nigam Shah. *Key Advances in Clinical Informatics: Transforming Health Care Through Health Information Technology*, chapter 19, pages 279–291. Academic Press, 2017.
- [15] Marcia Irene Canto, Femme Harinck, Ralph H Hruban, George Johan Offerhaus, Jan-Werner Poley, Ihab Kamel, Yung Nio, Richard S Schulick, Claudio Bassi, Irma Kluijft, et al. International cancer of the pancreas screening (caps) consortium summit on the management of patients with increased risk for familial pancreatic cancer. *Gut*, 62(3):339–347, 2013.
- [16] MI Canto, M Goggins, CJ Yeo, and et al. Screening for pancreatic neoplasia in high-risk individuals: an eus-based approach. *Clinical Gastroenterology and Hepatology*, 2(7):606–621, 2004.
- [17] E. Choi, M. T. Bahadori, , and J. Sun. Predicting clinical events via recurrent neural networks. In *arXiv preprint*, 2015.
- [18] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *30th Conference on Neural Information Processing Systems*, 2016.
- [19] D. Charles, M. Gabriel, T. Searcy. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2014. *ONC Data Brief*, 23, 2015.
- [20] P Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.
- [21] Ahmed Elmisery and Huaiguo Fu. Privacy preserving distributed learning clustering of healthcare data using cryptography protocols. In *IEEE 34th Annual Computer Software and Applications Conference Workshops*, 2010.

- [22] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PLoS One*, 6, 2011.
- [23] Hossein Estiri, Jeffrey Klann, and Shawn Murphy. A clustering approach for detecting implausible observation values in electronic health records data. *BMC Medical Informatics and Decision Making*, 19, 2019.
- [24] Julie Evans, Alison Chapple, Helen Salisbury, Pippa Corrie, and Sue Ziebland. “it can’t be very important because it comes and goes”—patients’ accounts of intermittent symptoms preceding a pancreatic cancer diagnosis: a qualitative study. *BMJ open*, 4(2):e004215, 2014.
- [25] A. S. Fleisher, B. B. Sowell, C. Taylor, and et al. Clinical predictors of progression to alzheimer disease in amnesic mild cognitive impairment. *Neurology*, 68(19):1588–1595, 2007.
- [26] B. Gallego, S.R. Walter, R.O. Day, A.G. Dunn, V. Sivaraman, N. Shah, C. A. Longhurst, and E. Coiera. Bringing cohort studies to the bedside: framework for a ‘green button’ to support clinical decision-making. *Journal of Comparative Effectiveness Research*, pages 1–7, 2015.
- [27] Juliano Gaspar, Emanuel Catumbela, Bernardo Marques, and Alberto Freitas. A systematic review of outliers detection techniques in medical data - preliminary study. In *HEALTHINF 2011 - Proceedings of the International Conference on Health Informatics*, pages 575–582, 01 2011.
- [28] Thore Graepel, Kristin Lauter, and Michael Naehrig. MI confidential: Machine learning on encrypted data. In *International Conference on Information Security and Cryptology*, pages 1–21, 2012.
- [29] Gastrointestinal Tumor Study Group. Further evidence of effective adjuvant combined radiation and chemotherapy following curative resection of pancreatic cancer. *Cancer*, 59:2006–2010, 1987.
- [30] Shelly Gupta, Dharminder Kumar, and Anand Sharma. Performance analysis of various data mining classification techniques on healthcare data. *International Journal of Computer Science and Information Technology*, 3, 08 2011.
- [31] GC Harewood and MJ Wiersema. Endosonography-guided fine needle aspiration biopsy in the evaluation of pancreatic masses. *Am J Gastroenterol*, 97(6):1386–1391, 2002.
- [32] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 2013.
- [33] Ralph Hruban, Gloria Petersen, Patrick Ha, and Scott Kern. Genetics of pancreatic cancer: from genes to families. *Surgical Oncology Clinics of North America*, 7(1):1–23, 1998.

- [34] Meng Hsuen Hsieh, Li-Min Sun, Cheng-Li Lin, Meng-Ju Hsieh, Chung-Y Hsu, and Chia-Hung Kao. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. *Cancer management and research*, 10:6317, 2018.
- [35] Stanley Huff, Roberto Rocha, Clement McDonald, and et al. Development of the logical observation identifier names and codes (loinc) vocabulary. *Journal of the American Medical Informatics Association*, 5(3):276–292, 1998.
- [36] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.
- [37] National Cancer Institute. Cancer stat facts: Pancreatic cancer. <https://seer.cancer.gov/statfacts/html/pancreas.html>. Accessed: 2020-01-28.
- [38] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 2017.
- [39] MG Kahn, TJ Callahan, J Barnard, AE Bauck, J Brown, B Davidson, and et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS*, 4(1), 2016.
- [40] M Kanda, S Knight, M Topazian, S Syngal, J Farrell, J Lee, and et al. Mutant gnas detected in duodenal collections of secretin-stimulated pancreatic juice indicates the presence or emergence of pancreatic cysts. *Gut*, 62:1024–1033, 2013.
- [41] James Max Kanter and Kalyan Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015.
- [42] Uri Kartoun, Kathleen Corey, Tracey Simon, Hui Zheng, Rahul Aggarwal, Kennedy Ng, and Stanley Shaw. The meld-plus: A generalizable prediction risk score in cirrhosis. *PLoS One*, 12, 2017.
- [43] Hayes M.G. Kho, A.N., L. Rasmussen-Torvik, and et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *JAMIA*, 19(2):212–218, 2012.
- [44] Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, 53(11):3735–3745, 2009.
- [45] I Kinde, N Papadopoulos, KW Kinzler, and B Vogelstein. Fast-seqs: a simple and efficient method for the detection of aneuploidy by massively parallel sequencing. *PLoS One*, 7, 2012.
- [46] Jason Klapman and Mokenge Malafa. Early detection of pancreatic cancer: Why, who, and how to screen. *Cancer Control*, pages 280–287, 2008.

- [47] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.
- [48] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS One*, 2013.
- [49] Donghui Li, Keping Xie, and James Abbruzzese. Pancreatic cancer. *The Lancet*, 363(9414):1049–1057, 2004.
- [50] Xianchao Lin, Bohan Zhan, Shi Wen, Zhishui Li, Heguang Huang Huang, and Jianghua Feng. Metabonomic alterations from pancreatic intraepithelial neoplasia to pancreatic ductal adenocarcinoma facilitate the identification of biomarkers in serum for early diagnosis of pancreatic cancer. *Molecular Biosystems*, 12(9):2883–2892, 2016.
- [51] Zachary Lipton, David Kale, Charles Elkan, and Randall Wetzel. Learning to Diagnose with LSTM Recurrent Neural Networks. In *ICLR*, 2016.
- [52] Zachary Lipton, David Kale, and Randall Wetzel. Modeling Missing Data in Clinical Time Series with RNNs. In *Proceedings of Machine Learning for Healthcare*, volume 56, 2016.
- [53] A Machanavajjhala, J Gehrke, D Kifer, and M Venkitasubramaniam. L-diversity: privacy beyond k-anonymity. In *22nd international conference data engineering (ICDE)*, 2006.
- [54] Jared R Mayers, Chen Wu, Clary B Clish, Peter Kraft, Margaret E Torrence, Brian P Fiske, Chen Yuan, Ying Bao, Mary K Townsend, Shelley S Tworoger, et al. Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nature medicine*, 20(10):1193, 2014.
- [55] Jennifer McCleary. pancreatic-cancer-models. <https://github.com/jennymccleary/pancreatic-cancer-models>, 2020.
- [56] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.
- [57] Nir Menachemi and Taleah H Collum. Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy*, 4:47–55, 2011.
- [58] Bruce Minsky, Basil Hilaris, and Zvi Fuks. The role of radiation therapy in the control of pain from pancreatic carcinoma. *Journal of Pain and Symptom Management*, 3(4):199–205, 1988.

- [59] CG Moertel, S Frytak, RG Hahn, and et al. Therapy of locally unresectable pancreatic carcinoma: a randomized comparison of high dose (6000 rads) radiation alone, moderate dose radiation (4000 rads + 5-fluorouracil), and high dose radiation + 5-fluorouracil. *Cancer*, 48:1705–1710, 1981.
- [60] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, chapter 4. 2019.
- [61] Wazir Muhammad, Gregory R Hart, Bradley Nartowt, James J Farrell, Kimberly Johung, Ying Liang, and Jun Deng. Pancreatic cancer prediction through an artificial neural network. *Frontiers in Artificial Intelligence*, 2:2, 2019.
- [62] N. Howlader, A. Noone, M. Krapcho, N. Neyman, R. Aminou, et al. SEER Cancer Statistics Review. 2011.
- [63] JP Neoptolemos, JA Dunn, DD Stocken, and et al. Adjuvant chemoradiotherapy and chemotherapy in resectable pancreatic cancer: a randomised controlled trial. *The Lancet*, 358(9293):1576–1585, 2001.
- [64] Andrew Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the 21 st International Conference on Machine Learning*, 2004.
- [65] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.
- [66] Samarati P. Protecting respondent’s privacy in microdata release. *IEEE Trans Knowl Data Eng*, 13(6):1010–1027, 2001.
- [67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [68] R. Siegel, K. Miller, and A. Jemal. Cancer Statistics, 2019. *CA: A Cancer Journal for Clinicians*, 2019.
- [69] Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, 3(4), 2015.
- [70] Carol Remmers, Judith Hibbard, David M. Mosen, Morton Wagenfield, Robert E. Hoye, and Ches Jones. Is patient activation associated with future health outcomes and healthcare utilization among patients with diabetes? *Journal of Ambulatory Care Management*, 32(4):320–327, 2009.
- [71] SJ Rulyak, MB Kimmey, DL Veenstra, and et al. Cost effectiveness of pancreatic cancer screening in familial pancreatic cancer kindreds. *Gastrointestinal Endoscopy*, 57(1):23–29, 2003.

- [72] T Rösch, R Lorenz, C Braig, and et al. Endoscopic ultrasound in small pancreatic tumors. *Z Gastroenterol*, 29(3):110–115, 1995.
- [73] Raghuwansh P Sah, Sajan Jiv Singh Nagpal, Debabrata Mukhopadhyay, and Suresh T Chari. New insights into pancreatic cancer-induced paraneoplastic diabetes. *Nature reviews Gastroenterology & hepatology*, 10(7):423, 2013.
- [74] Raghuwansh P. Sah, Ayush Sharma, Sajan Nagpal, and et al. Phases of metabolic and soft tissue changes in months preceding a diagnosis of pancreatic ductal adenocarcinoma. *Gastroenterology*, 156:1742–1752, 2019.
- [75] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.
- [76] Jae Song and Kevin Chung. Observational studies: Cohort and case-control studies. *Plastic Reconstructive Surgery*, 126(6):2234–2242, 2010.
- [77] S Stapley, TJ Peters, RD Neal, PW Rose, FM Walter, and W Hamilton. The risk of pancreatic cancer in symptomatic patients in primary care: a large case–control study using electronic records. *British Journal of Cancer*, 106:1940–1944, 2012.
- [78] J. Sun, F. Wang, J. Hu, and S. Edabollahi. Supervised patient similarity measure of heterogeneous patient records. *ACM SIGKDD Explorations Newsletter*, 14(1):16–24, 2012.
- [79] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. LSTM Neural Networks for Language Modeling. In *13th Annual Conference of the International Speech Communication Association*, 2012.
- [80] Joseph S. Verducci. Prediction and discovery. In *AMS-IMS-SIAM Joint Summer Research Conference, Machine and Statistical Learning*, 2006.
- [81] MY Wang, JL Abbruzzese, H Friess, and et al. Dna adducts in human pancreatic tissues and their potential role in carcinogenesis. *CancerRes*, 58:38–41, 1998.
- [82] Michael Waskom, Olga Botvinnik, Drew O’Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017.
- [83] Ashley Weinberg and Francis Creed. Stress and psychiatric disorder in healthcare professionals and hospital staff. *The Lancet*, 355(9203):533–537, 2000.

- [84] J Wiens, J Guttag, and E Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21:699–706, 2014.
- [85] Jenna Wiens and Erica Shenoy. Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, 66(1):149–153, 2017.
- [86] PW Wilson, RB D’Agostino, D Levy, AM Belanger, H Silbershatz, and WB Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.
- [87] Stephen Wright. Coordinate descent algorithms. *Mathematical Programming*, 151:3–34, 2015.
- [88] Charles Yeo, John Cameron, Keith Lillemoe, James Sitzmann, Ralph Hruban, Steven Goodman, William Dooley, JoAnn Coleman, and Henry Pitt. Pancreaticoduodenectomy for cancer of the head of the pancreas: 201 patients. *Annals of Surgery*, 221(6):721–733, 1995.