Predicting Pancreatic Cancer from EMR Data

Siwei Zhang, Jake A. Ukleja, Tyler J. Massaro, Joseph E. Lucas, James L. Abbruzzese, and Lisa L. Satterwhite

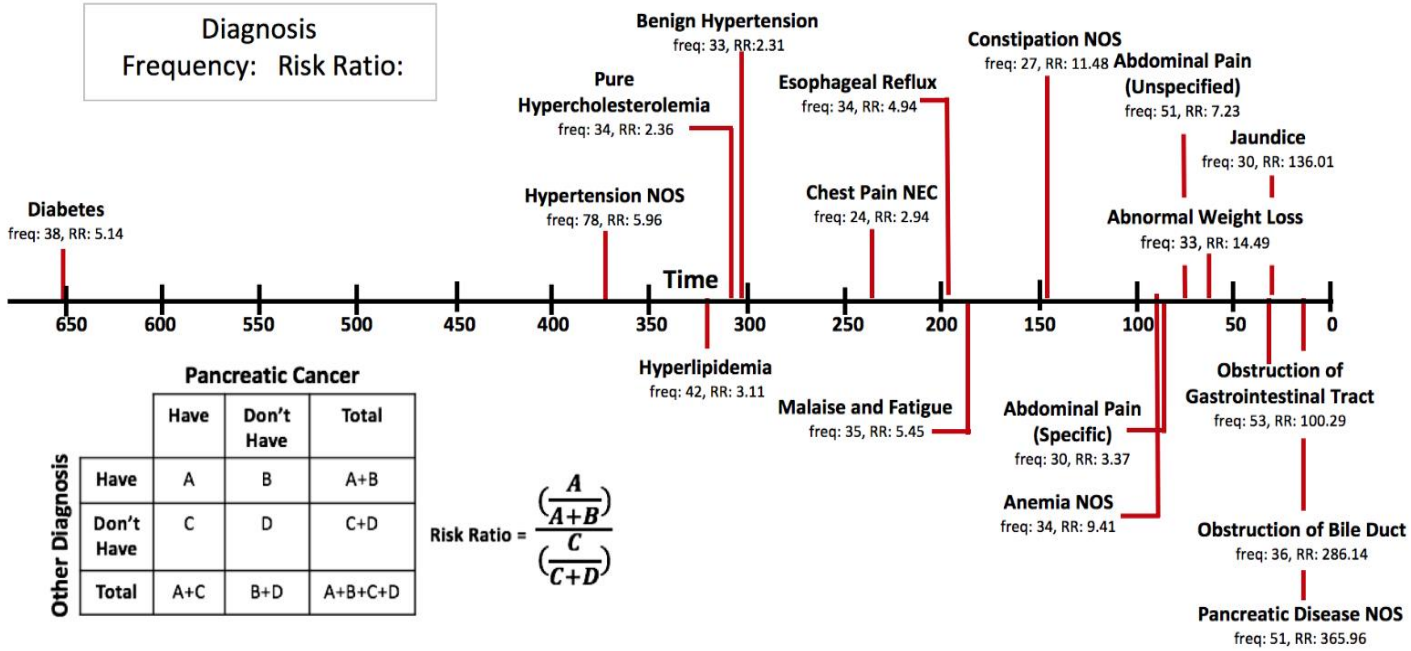
Introduction

Pancreatic ductal adenocarcinoma (PDAC) is the 4th leading cause of cancer deaths in the US and the only cancer predicted to rise in the next decade. PDAC is most often found in stage III and IV with a survival measured in months. Our goal is to identify asymptomatic early stage PDAC from the Duke EMR data using a supervised topic model to and to follow high risk patients prospectively.

Duke Electronic Medical Record (EMR)

- Spans 2004-2013
- 210,140 patients primarily from Durham County
- 11,550 unique ICD9 diagnosis codes
- 15,293 patients with diabetes
- 11,234 patients over 50 years with diabetes
- 5,712 over 50 with diabetes >3 years "new-onset diabetes"

Timeline to PDAC

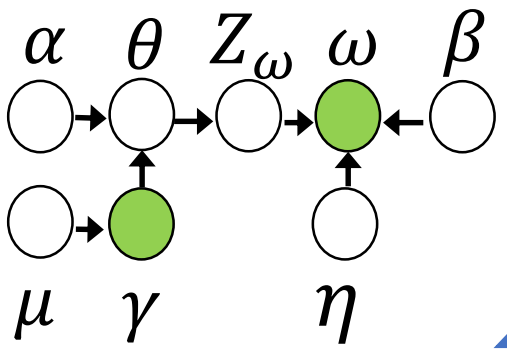


- The timeline shows median times of diagnosis codes prior to PDAC and their respective frequency.
- Diagnoses found closer to a PDAC diagnosis pose greater risk.
- While individually, each code may not warrant concern, in conjunction they may be predictive of risk.

Supervised Latent Dirichlet Allocation Approach

Theory of sLDA

γ: Label set
θ: Topic proportion in each patient
Z_ω: Topic assignment for each diagnosis code
β: Codes contribution for each topic
ω: Observed diagnosis code
α: Dirichlet topic prior parameter
η: Dirichlet word prior parameter
μ: Label prior for topic



Label pancreatic cancer patients with 0; label all general patients with 1.

Group partition

189 PDAC patients

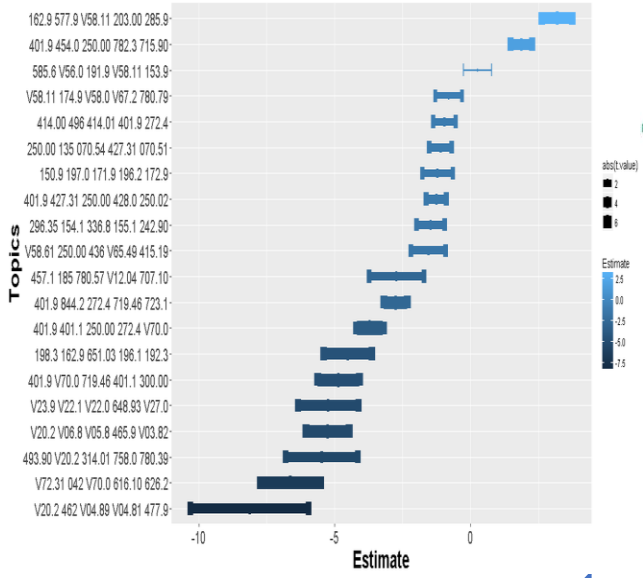
~2000 general patients

Applied sLDA on 105 groups of 210,140 general patients

Documents →	Patients
Words →	Diagnosis Codes
Topics →	Latent Diseases

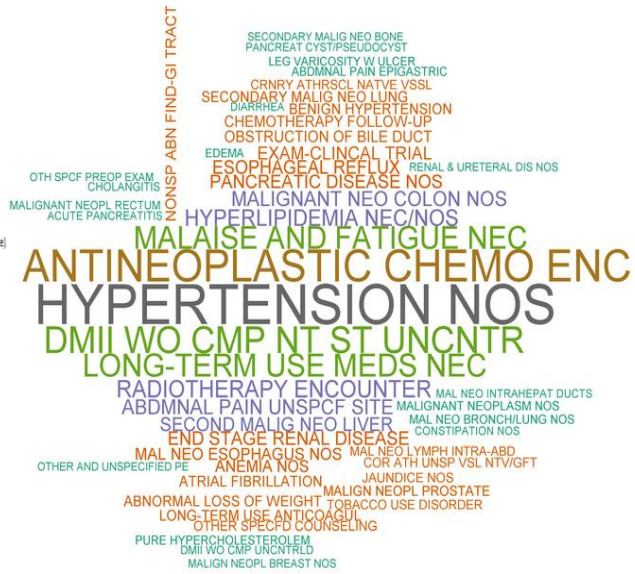
Randomly partition subset of general population into groups of about 2000 patients. Apply the topic model to each of the partitioned groups separately.

Regression Coefficient



Get regression coefficients from the topic model to make a predictive list. Positive topics are those with regression coefficients above 0.

Word Cloud

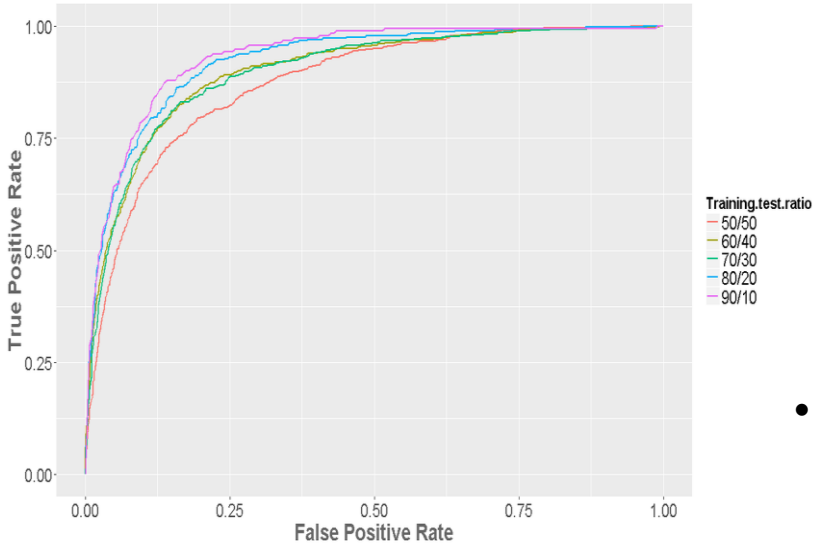


105 groups of regression coefficients, and over 200 positive topics, are the sum of loadings in all positive topics for each diagnosis code. The top 50 diagnosis codes are shown.

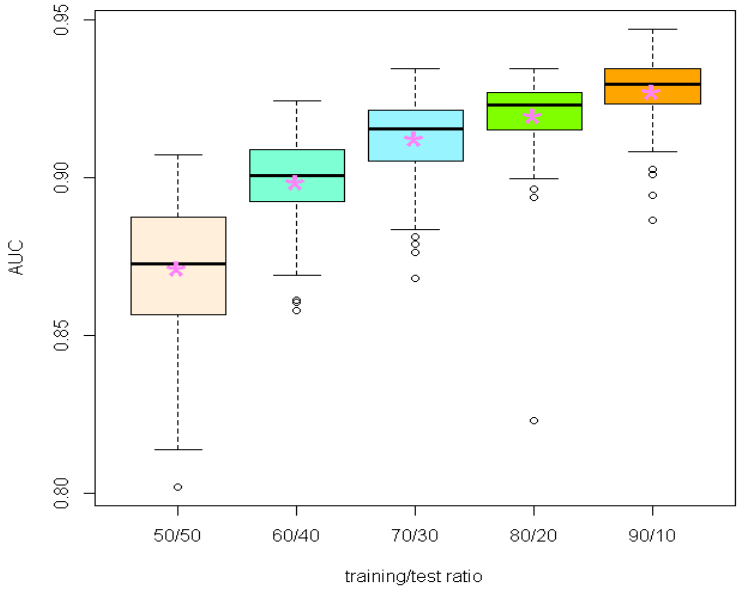
Predictive Performance and Future Work

Predictive Performance

ROC Curve



AUC Plot



Results

	A	B	C
1	patient_key	label	prediction score
2	146027	1	10.99257963
3	1315202	1	10.96494122
4	1406838	1	10.93188439
5	540994	1	10.42059657
6	3418	1	10.02609077
7	443045	1	9.656186709
8	474481	1	9.523979409
9	454748	1	8.913626535
10	1529042	1	8.263634274
11	4545354	1	8.145281642
12	1450625	1	8.061722577
13	563141	1	7.803349857
14	1280891	1	7.522927601

	A	B	C
26	450718	1	5.620354222
27	693	1	5.471068475
28	4379603	1	5.432106771
29	389589	1	5.413876683
30	587104	1	5.398610591
31	1217686	1	5.179579547
32	661457	1	5.042474738
33	166714	1	5.000611156
34	136313	1	4.771846935
35	373526	1	4.640293158
36	446294	1	4.594021857
37	384829	1	4.582021219
38	329895	1	4.533883107
39	4572434	1	4.006724837
40	475482	0	3.898726057
41	499058	1	3.84290464
42	468771	1	3.514677548

- Patients with a label of 0, without a pancreatic cancer 157.* diagnosis codes for PDAC, and a prediction score greater than 0, which predicts PDAC (yellow) are candidate high risk.
- 500 high risk patients were found in common across multiple trials of different seeds in the general patient population.

Future Work

- Apply the topic model to see the predictive performance in other cancers or neurodegenerative diseases that also develop silently.
- Provide clinicians a list of high risk patients for a prospective study.
- Validate externally with EMR data from other health systems.