ISYE6748 – PRACTICUM

Midterm Report

Nikhil **Anand** @nanand33
William **Lam** @wlam39
Shrea **Shyam** @sshyam3

The Goal

Create a set of models which power a web interface that allows an interrogator to survey the likelihoods of disease incidence based on a condition or combinations of conditions.

Data Exploration – Medical Coding Systems

We prioritized the ICD-10 system

It offers a **finer resolution** of patient maladies than ICD-9 and seeks to mitigate issues with improper or ambiguous application/transcription

~13,000 ICD-9 codes versus ~68,000 ICD-10 † codes (Source: American Medical Association)

Offers a **greater room for expansion**

E.g. UXX is "Provisional assignment of new diseases of uncertain etiology or emergency use" and includes COVID-19

According to the CDC, "The content [ICD-9] is no longer clinically accurate and has limited data about patients' medical conditions and hospital inpatient procedures, the number of available codes is limited, and the coding structure is too restrictive."

† In our explorations with 2023 ICD10 data, we counted 73,639 codes

Data Exploration – ICD-10 Layers

We determined four **'layers'** of broad → specific disease classifications as illustrated.

- A00–B99 Certain infectious and parasitic diseases
- C00–D49 Neoplasms
- D50–D89 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- E00–E89 Endocrine, nutritional and metabolic diseases
- F01–F99 Mental, Behavioral and Neurodevelopmental disorders
- G00–G99 Diseases of the nervous system
- H00–H59 Diseases of the eye and adnexa
- H60–H95 Diseases of the ear and mastoid process
- I00–I99 Diseases of the circulatory system
- J00–J99 Diseases of the respiratory system
- K00–K95 Diseases of the digestive system
- L00–L99 Diseases of the skin and subcutaneous tissue
- M00–M99 Diseases of the musculoskeletal system and connective tissue
- N00–N99 Diseases of the genitourinary system
- O00–O9A Pregnancy, childbirth and the puerperium
- P00–P96 Certain conditions originating in the perinatal period
- Q00–Q99 Congenital malformations, deformations and chromosomal abnormalities
- R00–R99 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- S00–T88 Injury, poisoning and certain other consequences of external causes
- U00–U85 Codes for special purposes
- V00–Y99 External causes of morbidity
- Z00–Z99 Factors influencing health status and contact with health services

Layer 1

ICD-10-CM Range K00–K95

Diseases of the digestive system

- K00–K14 Diseases of oral cavity and salivary gla...
- K20–K31 Diseases of esophagus, stomach and duode...
- K35–K38 Diseases of appendix
- K40–K46 Hernia
- K50–K52 Noninfective enteritis and colitis
- K55–K64 Other diseases of intestines
- K65–K68 Diseases of peritoneum and retroperitone...
- K70–K77 Diseases of liver
- K80–K87 Disorders of gallbladder, biliary tract ...
- K90–K95 Other diseases of the digestive system

Layer 2

K35 Acute appendicitis
K36 Other appendicitis
K37 Unspecified appendicitis
K38 Other diseases of appendix

Layer 3

ICD-10-CM Diagnosis Codes K38–*

- K38 Other diseases of appendix
  - K38.0 Hyperplasia of appendix
  - K38.1 Appendicular concretions
  - K38.2 Diverticulum of appendix
  - K38.3 Fistula of appendix
  - K38.8 Other specified diseases of appendix
  - K38.9 Disease of appendix, unspecified

Layer 4

Data Exploration – ICD-10 Layers

With this system each disease ICD-10 Code is classified into

22	Layer 1 Classes
283	Layer 2 Classes
1,914	Layer 3 Classes
73,639	Layer 4 Classes

Data Exploration – Classes

In our *initial* explorations we decided that our maximum resolution is a Layer 3 class (e.g. K42 - “Umbilical Hernia”) and *not* a Layer 4 class (e.g. K42.1 - “Umbilical hernia with gangrene”)

👉 But wait! We say that we prefer ICD-10 over ICD-9 for “greater resolution” of disease conditions.

Why are we preferring ICD-10 over ICD-9? Why are we ‘throwing away’ Layer 4 data?

Data Exploration – Classes

At the moment, we would like to test our data refinement, feature engineering + selection, and model selection + evaluation with smaller matrices to check our hunches

At Layer 3, ICD-10 still offers a *relatively* finer resolution with **1,914** classes (e.g K17, A23, G17) than ICD9 at this layer with **1,042** classes (e.g. 049, 389, V82)

We simply prefer a more modern/recent disease classification system *if* the data allows for our preference (which it does)

There simply isn't much ICD-9 primary/secondary/tertiary diagnosis data to go by in the supplied datasets

The Goal – Examples with ICD-10 Codes, Classes, & Layers

"Create a set of models which power a web interface that allows an interrogator to survey the likelihoods of disease incidence based on a condition or combinations of conditions."

What is the likelihood of **broad diseases of the digestive system** (K00-K95) given a **disease of the genitourinary system** (N00-N99)?

Layer 1 ↔ Layer 1

What is the likelihood of a **hernia** (K40-K46) given a **pregnancy** (O00-O9A)?

Layer 2 ↔ Layer 1

Do **sleep disorders** (G47) affect the **respiratory system** (J00-J99)?

Layer 3 ↔ Layer 1

And so on.

Data Exploration – Raw Medical and Rx Datasets

8,092,330 Medical Encounters

- 100,000 patients as identified by unique “Medical Life ID”
- 2.41% (195,498) Medical Life IDs were `-1` indicating missing data

2,570,133 Pharmacy/Rx Claims

- 54,001 patients as identified by unique “Medical Life ID”
- 41.14% (1,057,426) Medical Life IDs were `-1` indicating missing data

Union of these datasets on “Medical Life ID” field was dissatisfying

- Yielded 233 *additional* IDs not present in Medical encounter datasets
- Ostensibly members with Rx claims but no Medical encounters

Data Exploration – ICD10 Codes

We examined ICD-10 Codes across **Primary**, **Secondary**, and **Tertiary** assignment tiers and weighted all codes equally across these tiers for our proposed models.

Rationale

- You can keep visiting the doctor for a simple stomach ache (**Primary** assignment R52) that might *eventually* be deemed a “malignant neoplasm of the colon” (**Primary** assignment C18)
- Codes are assigned with no particular weight across various tiers as the physician documents the progression of a diagnosis
- We are interested in looking for diseases or conditions that cluster together

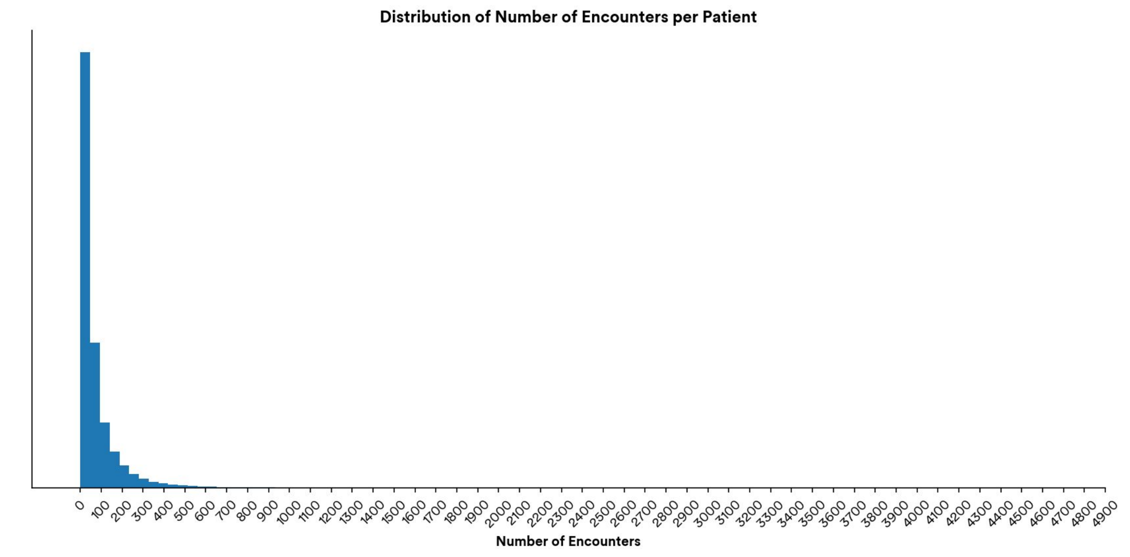
Note: Every observation in the Medical dataset contained at least one Primary, Secondary, or Tertiary ICD-10 assignment.

Data Exploration – Distribution of Encounters

195,497 encounters had a Medical Life ID of `-1` implying missing data. These were excised from the Medical dataset

Here's a summary of the distribution of encounters

count	99999.000000
mean	78.968320
std	147.033544
min	1.000000
25%	15.000000
50%	37.000000
75%	88.000000
max	4659.000000



Data Exploration – Distribution of Encounters

Whence 100+ encounters per Member Life ID?

The highest number of records for a patient was 4,659. Upon exploration of this patient with Member Life ID `459152`, we found that there are 19 rows with same **Receipt Date**. We then checked the place of service code which was `Outpatient Hospital`, where the patient has end-stage renal disease.

We are assuming that this is the case with the remainder of the outliers.

It is also clear that each observation does not map to a single encounter.

Data Exploration – Feature Selection – Initial Features

We have **84 features** in the **Medical** dataset and **51 features** in the **Rx** dataset

We also, luckily, have a team member who is a highly experienced former medical coder 🍀
We employed her domain expertise to pick these Initial Features

Medical dataset

- Member Life ID
- Gender Code
- Primary Diagnosis Code-ICD10
- Secondary Diagnosis Code-ICD10
- Tertiary Diagnosis Code-ICD10
- Jurisdiction
- Admission Date
- Number of Submitted Inpatient Days
- Number of Services

Rx dataset

- Member Life ID
- Date of Birth

Data Exploration – Feature Selection – Dropped Features

The next few slides will explain why we **did not pick** or **dropped certain fields**.

❌ Surgical Procedure Codes

We are not interested in the *how* a disease was cured; only *that* a disease was diagnosed/present

❌ Institutional and Professional Diagnoses

This pertains to whether the provider is a hospital (institutional) or individual physician (professional). *All this data, without exception* is captured in the Primary/Secondary/Tertiary ICD10 code fields rendering these fields redundant.

Data Exploration – Feature Selection – Dropped Features

❌ **Jurisdiction** (Initial Feature)

Unique values were `Maryland`, `Virginia`, `District of Columbia`, and `Other`

We felt that these would **not** empower our models given that they were mostly from the US East Coast and one ambiguous geography.

Further, given the spread of ICD10 conditions across the dataset, these are too sparse to imply a particular and specific cluster of diseases.

E.g. A patient from the Eastern half of Iowa (US Midwest) has a good chance of developing a rare eye disease like Stargardt's given the Amish populations that have settled there.

Data Exploration – Feature Selection – Dropped Features

❌ **Admission Date** (Initial Feature)

We wanted to use this with the DOB in the **Rx** Claims dataset to create a new field: **Age at Admission.** The earliest admission date is `1959-04-02` and latest is `2019-08-22`.

However, an overwhelming 90%+ of dates are `1999-12-31` indicating missing data

This cannot tenably be an admission date from the year 1999

Further, the **de-duped** Rx dataset only provided 7,976 unique Member IDs, **94% of which are** `-1` indicating missing data. **233 IDs are new/unseen** in the **Medical** Dataset.

We are therefore unable to engineer an "**Age at Admission**" field which we feel would have good predictive powers using the Admission Date field and elected to drop it.

Data Exploration – Feature Selection – Dropped Features

❌ Number of Submitted Inpatient Days

Small number (~0.05%) have negative days.

But 99.57% of submitted days are zero!
This implies that these visits may not have been in a hospital setting.

Dropped this feature since it was a very sparse vector.
Might reintroduce it later

Data Exploration – Feature Selection – Dropped Features

❌ Number of Services

A **small number** (4.82%) of services are **zero or negative**

A **very large number** (82.20%) of them of them are **exactly 1**

The min/max values are -10,000/+10,000 (which is absurd)
Mean number of services is 3.75

We simply did not understand this feature and dropped it.
Might reintroduce it later

Data Transformation

Our **final list of features** from the **Medical** dataset is now

- Member Life ID
- Gender Code
- Primary Diagnosis Code-ICD10
- Secondary Diagnosis Code-ICD10
- Tertiary Diagnosis Code-ICD10

Observations with Member Life IDs that were `-1` represented a small portion (2.41%) of the dataset. We randomized these IDs.

De-duplication based on these features resulted in 1,768,735 observations, **a 4.6x reduction** from the original 8,092,250 observations.

The “Number of Services” field contributed the most to redundancy

Data Transformation – Gender Code

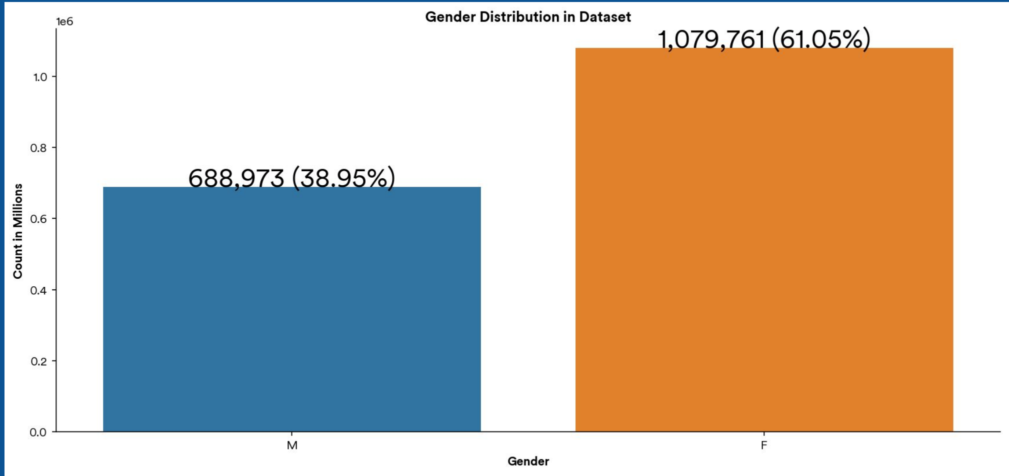
Categorical variable with three values: `M`, `F`, and `U`

`U` presumably represents “Unknown” or “Intersex”

There is **exactly one** observation in the **de-duped** dataset

72 observations (0.0009%) in the original dataset

We removed all observations with `U` given its **weak signal** and converted the field to a binary 0 = M and 1 = F



Data Transformation – ICD10 Codes

We weighted all ICD10 codes equally across Primary, Secondary, and Tertiary assignments.

For reasons described earlier

Created **three data matrices** for **each** of Layers {1, 2, 3} with {22, 283, 1914} classes.

Each class is populated by the **frequency** of each ICD10 code **across all** Primary, Secondary, and Tertiary **assignments**.

Member Life ID	Gender Code	A00-B99	C00-D49	D50-D89	E00-E89	F01-F99	G00-G99	H00-H59	H60-H95	...
236523	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
236523	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
236523	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
236523	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
236523	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
...	...	...	...	...	...	...	...	...	...	...

Sample Data Matrix with **Layer 1** Classes

Current Modeling Approach

👉 At this moment, we are **not** looking to model or factor in **temporal** or **causal** relationships

We are experimenting with **Recommendation Systems** ("recsys") as our **strongest candidate** given a **direct mapping** to our **conception of the problem**.

E.g. "You like *The Godfather*, *Goodfellas*, and *The Sopranos*. You might like *The Irishman* and *Peaky Blinders*"

GENRES american movies crime mafia violent coppola

Maps to

E.g. "You have ICD codes *R50*, *I48*, and *J98*. You might develop *G47* and *E55*"

GENRES fevers diseases of the circulatory system respiratory illnesses vitamin deficiencies

We are building **three models** for each of Layers 1, 2, and 3 and combining the results for display on a web UI (in progress).

Proposed Modeling Approach

We are studying and will evaluate **Image Recognition** as another possible candidate.

Core idea is to

1. Find a suitable image representation of each observation in our refined data matrix
2. Use an image classifier to discover and rank similar observations

We *think* that this approach will address *both* Use Cases I and II

Fin.