Medical Ontology Embeddings and an Example Application: Identifying Cachexia

Mukund Sridhar & Robert Norberg

Bio-Data Club Aug 17, 2023

MOFFITT CANCER CENTER®

Key Contributors

Robert Norberg

Advanced Analytics,
Health Data Services

Poornima Ramaraj
Neil Mason

Advanced Analytics

Rodrigo Carvajal
Kavita Ghia
Vonetta Williams
CDSC

Chandan Challa
Data Engineering

What is the business problem we are solving?

➢ Predict targeted medical conditions along from Clinical Notes & related Biomarkers

Revenue Cycle Management

Improved clinical coding

efficiencies via Clinical note abstractions

Accurate revenue capture

through NLP solutions for clinical coding

Sponsor: Lynn Ansley

Quality & IHM

Improved quality risk profiling

with ML predictions for medical conditions

Sponsor: Dr. Tim Hembree

Why Amazon Comprehend?

- Leverage solutions within Moffitt's AWS ecosystem
- Pre-trained NLP to read clinical notes generate multiple ICD 10 codes & confidence levels

Why use ICD10 Embeddings?

- Ability to leverage pre-built Vector ICD10 embeddings
- Clinicians and RCM teams are familiar with ICD10 language

Why Cachexia?

- Lack of agreed definition & diagnostic criteria
- One of the top weighted conditions for quality risk scores

Outline

- Introduction
- ICD-10 Code Embeddings
    - Understanding ICD-10 codes
    - What are embeddings?
    - Benefits of embeddings for ICD-10 codes
    - Training embeddings for ICD-10 codes
    - Interpreting ICD-10 code embeddings
    - Limitations
- An application of ICD-10 code embeddings – identifying cachexia
    - Known cachexia
    - Using biomarkers (patient weight)
    - Using co-occurring and prior diagnoses (represented as embeddings)
    - Using patient notes
- Conclusion
- Q & A

Introduction

Introduction

Although generative AI dominates the news, discriminative machine learning algorithms are still the workhorses of applied machine learning in industry.

Embeddings are a foundational concept in even state-of-the-art AI that can also be used in more traditional machine learning methods.

We use embeddings to represent a medical ontology (ICD-10 diagnosis codes), then use these embeddings to identify cachexic patients that did not have cachexia coded.

There are several benefits of mapping data onto an ontology like ICD-10 as an intermediate step in AI/ML:

- Data Integration: Ontologies unify disparate data sources by providing a common representation. This allows AI/ML models process information from diverse data sources in a structured and coherent manner.
- Transfer Learning: Representing disparate data sources using a unifying ontology facilitates transfer learning, allowing AI/ML models to leverage domain knowledge beyond any specific application. This is particularly useful in domains with limited training data.
- Explainability: Users can trace how conclusions were reached by following the logic in the AI/ML model back into the ontological representation of the input data, providing transparency to AI/ML decision-making.
- Search and Recommendation: Ontologies lend themselves to search and recommendation systems by understanding user queries and preferences in the context of relationships. This leads to more precise search results and better recommendations.

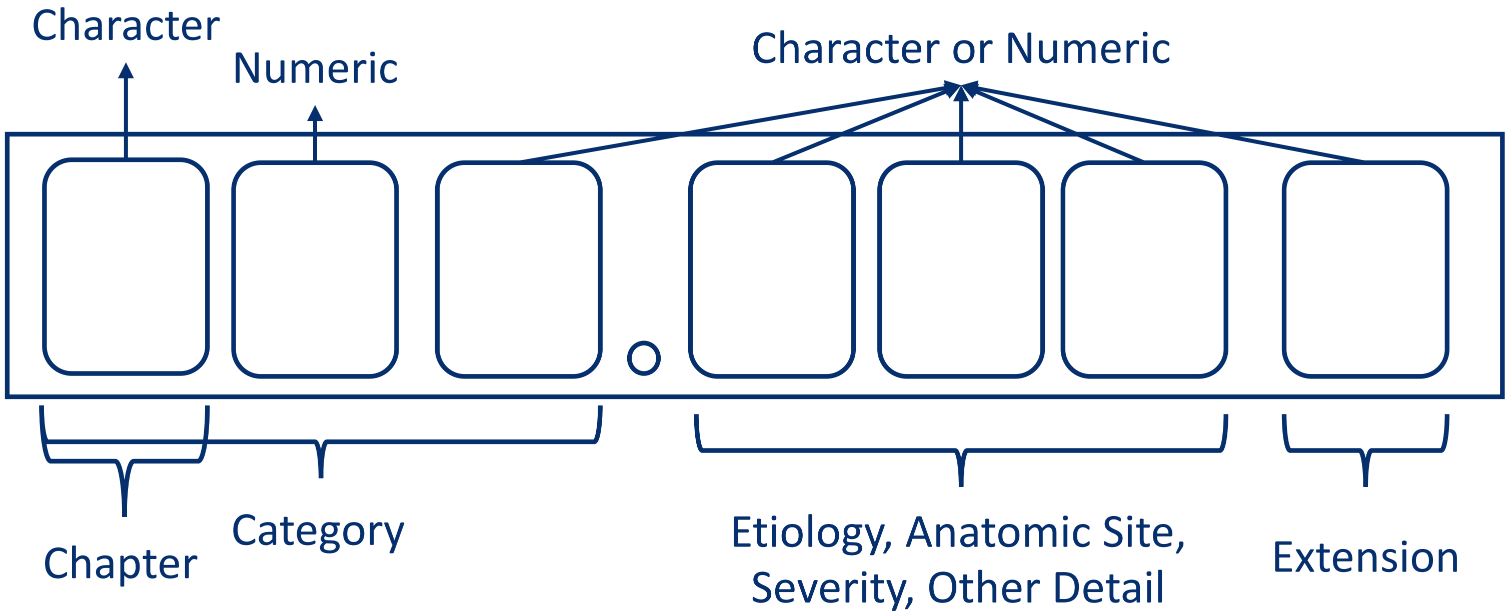
ICD-10 Code Embeddings

Understanding ICD-10 Codes

ICD-10 is short for International Classification of Diseases, Tenth Revision.

ICD-10 codes are a standardized ontology used in healthcare to classify and categorize medical conditions, assigning unique alphanumeric codes to facilitate accurate medical recordkeeping, billing, and research analysis. They provide a universal language for conveying essential medical information across healthcare institutions and systems.

ICD-10 codes have a hierarchical structure, as illustrated below.



Understanding ICD-10 Codes

- T63 - Toxic effect of contact with venomous animals and plants
- T63.5 - Toxic effect of contact with venomous fish
- T63.51 - Toxic effect of contact with stingray
- T63.513 - Toxic effect of contact with stingray, assault
- T63.513A - Toxic effect of contact with stingray, assault, initial encounter

- W56 - Contact with nonvenomous marine animal
- W56.0 – Contact with dolphin
- W56.01 – Bitten by dolphin
- W56.02 – Struck by dolphin
- W56.09 – Other contact with dolphin

LIVE
BREAKING NEWS
LOCAL STINGRAY IN GREAT MOOD
18:30
THE WATER IS NICE AND WARM AND HIS FLAPPY PARTS FEEL VERY FLAPPY INDEED

- C34 - Malignant neoplasm of bronchia and lung
- C34.9 - Malignant neoplasm: Bronchus or lung, unspecified
- C34.90 - Malignant neoplasm of unspecified part of unspecified bronchus or lung
- C34.91 - Malignant neoplasm of unspecified part of right bronchus or lung

What Are Embeddings?

Embeddings are numerical representations of words, codes, or items that a computer can understand more easily.

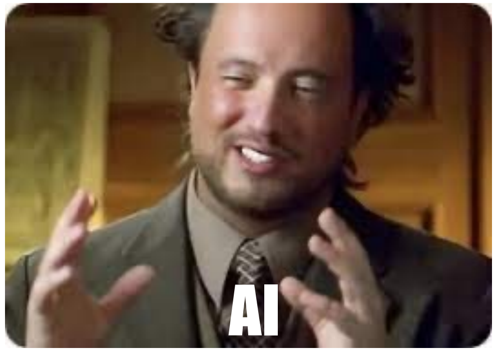
While it is intuitive to us that “toxic effect of contact with stingray” and “struck by dolphin” have some similarities, that is not obvious to a computer. Embeddings capture the semantic relationships and contextual meanings of words by mapping them to points in a multi-dimensional space that computers can better work with.

A single word (or ICD-10 code in our case) is represented as a point in the multi-dimensional embedding space. The coordinates for a point in this space constitute a numeric vector, which computers can work with quite efficiently.

Google’s Word2Vec popularized embeddings in 2013 and is one of the most well-known methods for creating embeddings of English words. Famously in Word2Vec, “king” – “man” + “woman” = “queen”. This is what is meant when it is said that embeddings capture semantic relationships.

Words that often appear in similar contexts have coordinates that are near each other in the embedding space. For example, “king” and “queen” are near each other in the embedding space because either word could complete these sentences:

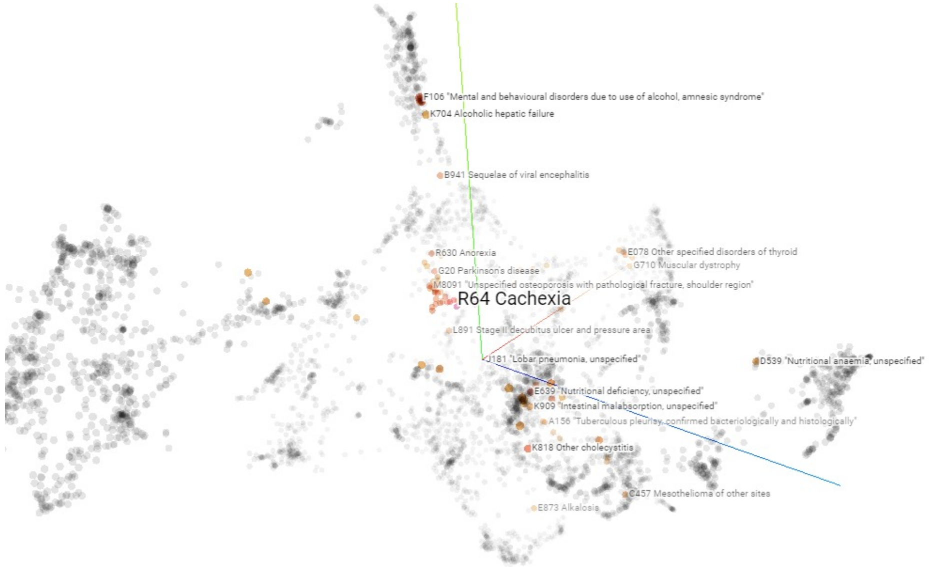
- He played an ace, then I played a _____.
- All hail the _____!
- Casey is the Scrabble ____!



Benefits of Embeddings for ICD-10 Codes

There exist over 69,000 ICD-10 diagnosis codes. Representing patients or claims using ICD-10 codes requires one row per patient or claim and one column for each ICD-10 code. Each element in this matrix contains a 1 if the code is associated with the patient or claim, and a 0 otherwise. This is a very wide, very sparse matrix. Our embeddings project each diagnosis code into a continuous 50-dimensional space, so we can represent the same information with an N x 50 matrix instead of an N x 69,000 matrix.

- An N x 50 matrix is much more tractable for machine learning algorithms than an N x 69,000 matrix.
- Codes that are similar have similar embedding representations, even if they are far away from each other in the ICD-10 code book.



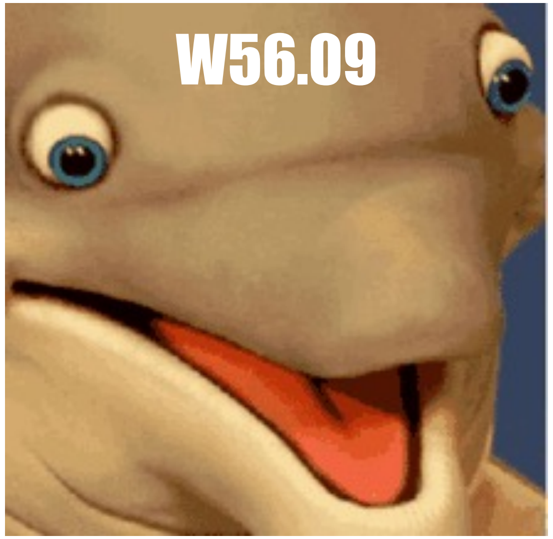
You can explore these embeddings here.

Training of Embeddings for ICD-10 Codes

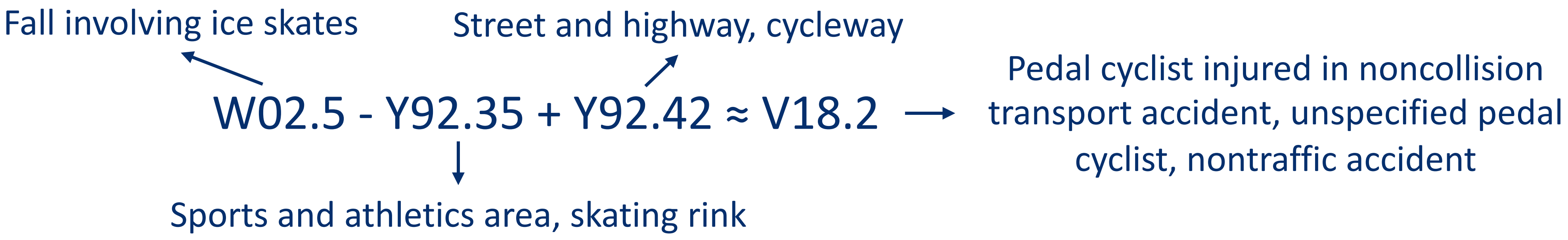
Word2Vec was trained by showing a simple neural network sentences with one word hidden, tasking the neural network with guessing the hidden word. This is how synonymous words come to have similar coordinates. This also does not require a “labelled” training data set; it simply requires sentences of text.

This is adapted by thinking of a single claim or episode as a sentence, and each recorded ICD-10 code in that claim/episode as a word. Instead of the English language, the vocabulary is all ICD-10 codes. Since there is no “grammar” to ICD-10 codes on a claim, it is sufficient to simply train the neural network using pairs of co-occurring ICD-10 codes (one is the input and the other is the hidden output the neural network is tasked with guessing).

Imagine a claim that includes a diagnosis of “bitten by dolphin”. What other diagnoses might also appear on that claim? What diagnoses would almost certainly not appear on that claim?



Interpreting ICD-10 Code Embeddings

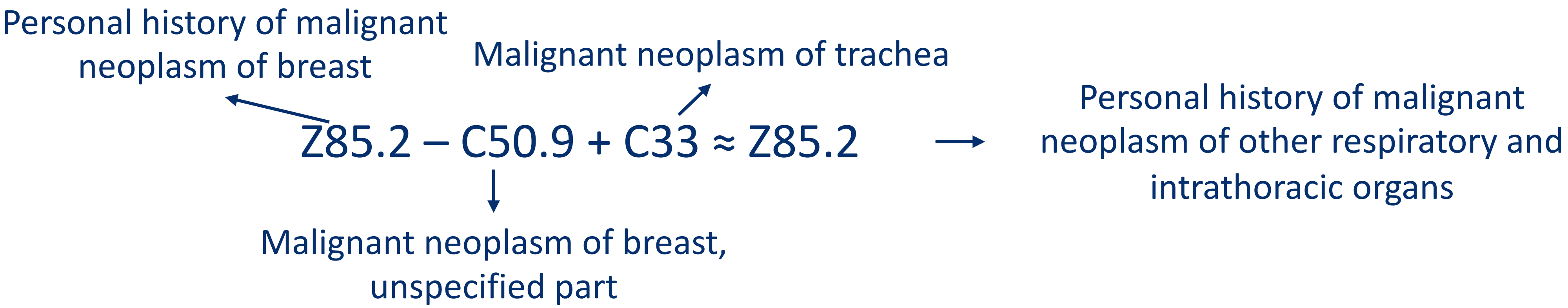


Fall involving ice skates

Street and highway, cycleway

W02.5 - Y92.35 + Y92.42 ≈ V18.2 → Pedal cyclist injured in noncollision transport accident, unspecified pedal cyclist, nontraffic accident

Sports and athletics area, skating rink



Personal history of malignant neoplasm of breast

Malignant neoplasm of trachea

Z85.2 – C50.9 + C33 ≈ Z85.2 → Personal history of malignant neoplasm of other respiratory and intrathoracic organs

Malignant neoplasm of breast, unspecified part

Interpreting ICD-10 Code Embeddings

The diagnosis codes nearest to cachexia (R64) in the 50-dimensional embedding space are:

Diagnosis	Description	Distance from Cachexia
E43	Unspecified severe protein-energy malnutrition	0.175
E44.0	Moderate protein-energy malnutrition	0.291
E63.9	Nutritional deficiency, unspecified	0.367
K18	Other cholecystitis	0.394
R54	Senility	0.394
E44.1	Mild protein-calorie malnutrition	0.395
M80.91	Unspecified osteoporosis with pathological fracture	0.402
K56.4	Other impaction of intestine	0.403
E87.0	Hyperosmolality and hypernatraemia	0.404
F10.6	Mental and behavioral disorders due to use of alcohol, amnesic syndrome	0.405

Limitations

There are some downsides to using embeddings.

- When explaining why an ML model made a specific prediction, a variable that is 1 when W56.0 appears and 0 otherwise is much easier to interpret than “embedding dimension 12”.
    - There is ongoing research in identifying interpretable subspaces within embeddings.
- The embeddings we are currently using were trained on a data set not specific to cancer treatment, so some ICD-10 codes that appear frequently in our data did not appear frequently enough in the training set to be included in the embedding vocabulary.

An Application of ICD-10 Code Embeddings: Identifying Cachexia

Motivation

- Identify past cachexic patients for research or clinical trials, and some cachexic patients may not have been coded as such.
- Identify current patients with cachexia so we can accurately code this condition, improving
    - Consistency of patient care
    - Data quality for future research
    - Quality metrics
    - Potentially even reimbursement

Known Cachexia

1,226 distinct patients with a cachexia diagnosis (R64) recorded in our billing data (*MCAP_PROD.DATACORE.V_BILLING_DIAGNOSIS*) between Apr 1, 2016 (six months after the date Moffitt switched from ICD-9 to ICD-10) and Apr 1, 2023.

For each patient, we found the date of their first cachexia diagnosis and consider this their “index date”. All data considered for each patient will be up to and including this index date.

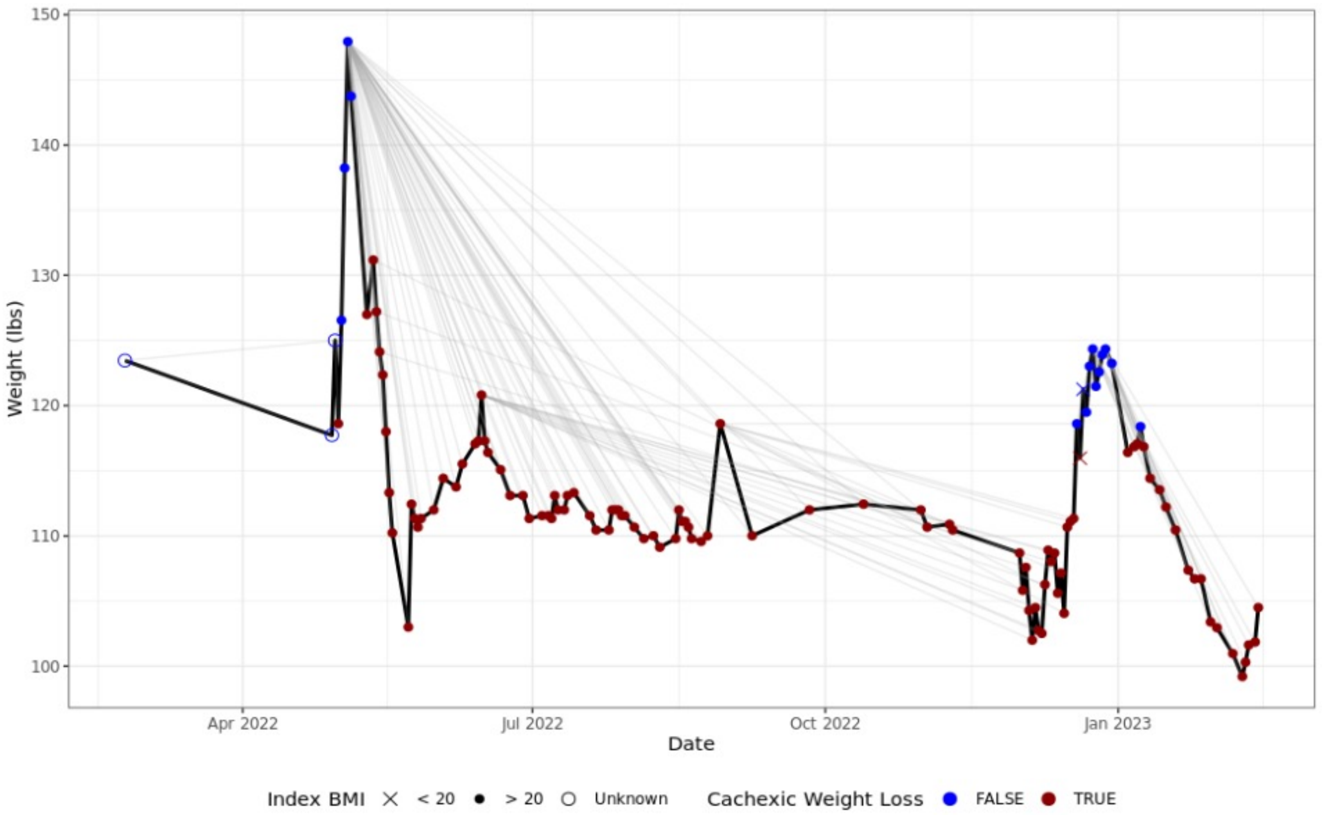
To complete our data set, we randomly selected 1,226 patients without a cachexia diagnosis and a random date on which they received some service at Moffitt to be their “index date”. This brings the **total sample size to 2,452 patients**, half with a cachexia diagnosis and half without.

Using Patient Weight

According to Fearon et. al. cachexia is described as weight loss of more than 5%, or BMI below 20 and weight loss greater than 2%, or sarcopenia and reduced food intake and systemic inflammation.

We used weight and BMI data in *MCAP_PROD.DATACORE.V_EMR_VITALS* to determine if each patient met the BMI/weight loss criteria for cachexia.

Here is an illustration of one patient's cachectic weight loss. The thin grey lines point back from each weight measurement to the prior weight the loss is calculated relative to (the patient's highest weight in the preceding 180 days).



Using Patient Weight

879 patients in our sample (36%) did not have sufficient history in *V_EMR_VITALS* to calculate their weight change.

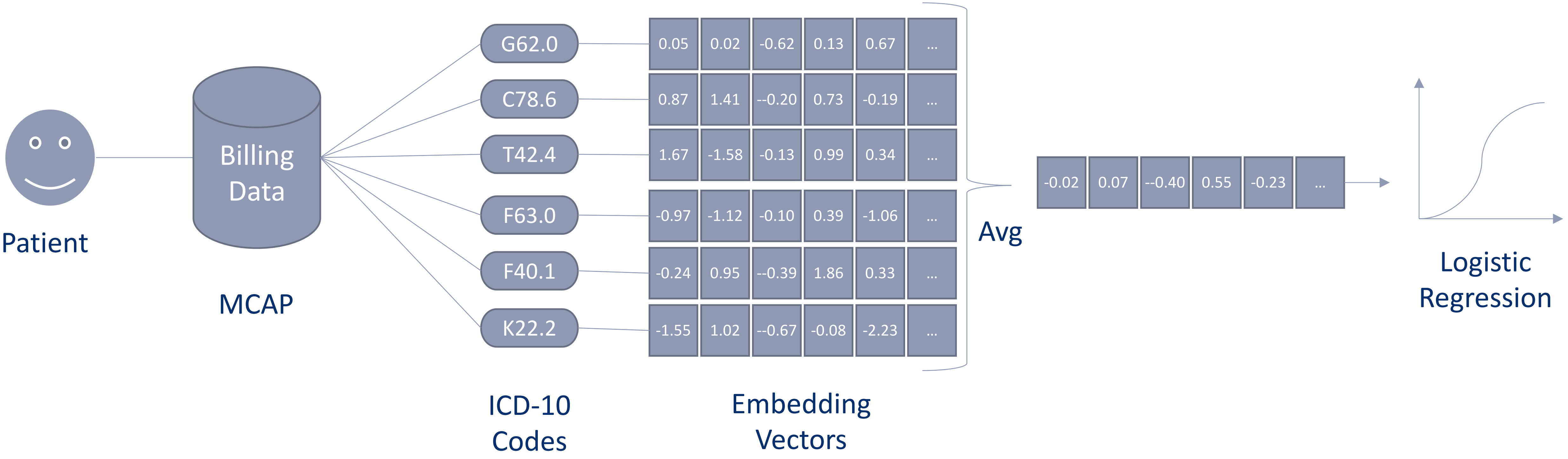
Of the remaining 1,573 patients, this method of calculating cachectic weight loss is 78% accurate.

	Has Cachexia Diagnosis	No Cachexia Diagnosis
Has Cachectic Weight Loss	913	156
Does Not Have Cachectic Weight Loss	185	319
Not Enough Weight Data	128	751

In MCAP since Oct 1, 2015, we found 61,899 patients with cachectic weight loss using our definition. Only 1,113 (1.8%) of these have a cachexia diagnosis in our billing data.

Using Co-Occurring and Prior Diagnoses

For a patient with six diagnoses in the billing data in the six months leading up to their index date, we have a matrix with six rows (one for each diagnosis) and 50 columns (one for each dimension of the embedding space). We simply took the average of each column so that each patient's diagnoses are represented by 50 numbers, regardless of how many diagnoses they have. This data was used to train a simple logistic regression classifier on the presence/absence of a cachexia diagnosis in the billing data (cachexia diagnoses were omitted from the training data).



0.05	0.02	-0.62	0.13	0.67	…
0.87	1.41	--0.20	0.73	-0.19	…
1.67	-1.58	-0.13	0.99	0.34	…
-0.97	-1.12	-0.10	0.39	-1.06	…
-0.24	0.95	--0.39	1.86	0.33	…
-1.55	1.02	--0.67	-0.08	-2.23	…

-0.02	0.07	--0.40	0.55	-0.23	…

You can find the full analysis, with code, here.

Using Co-Occurring and Prior Diagnoses

We trained a very simple logistic regression model on this data to classify each patient as cachectic or not.

The data set consists of 2,339 patients, 1,226 with a cachexia diagnosis in the billing data and 1,113 without.

We trained the model on 1,753 patients and held out 586 patients for validation.

On the validation set of 586 patients, the model is 88% accurate and has an ROC of 0.94.

	Has Cachexia Diagnosis	No Cachexia Diagnosis
Predicted Cachexic	283	46
Predicted Not Cachexic	24	233

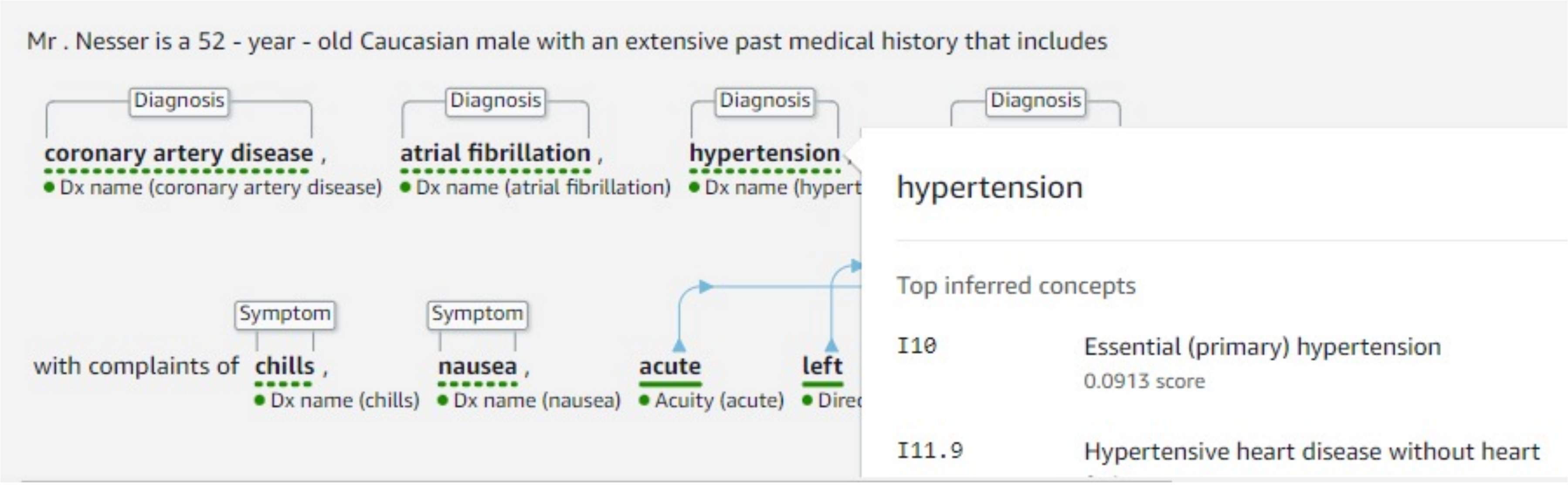
Using Patient Notes

We have hundreds of thousands of patient notes with an unknown amount of information buried in them. We tried using these to classify the patients in our sample as cachexic or not.

We fetched all Inpatient Progress Notes, History And Physical notes, and Ambulatory Care notes from *MCAP_PROD.STAGE.CERNER_CE_BLOB_EXTRACT* for each patient in the data set, from 50 days prior to their index date up to and including their index date.

This yields 7,960 notes and 1,692 of the 2,452 (69%) of the patients in our sample have at least one note.

We use AWS Comprehend Medical's [InferICD10CM API]to process each of these notes. This service identifies words/phrases that describe medical conditions and attempts to assign an ICD-10 code to each identified condition.



Mr . Nesser is a 52 - year - old Caucasian male with an extensive past medical history that includes

coronary artery disease , atrial fibrillation , hypertension ,

with complaints of chills , nausea , acute left

hypertension

Top inferred concepts

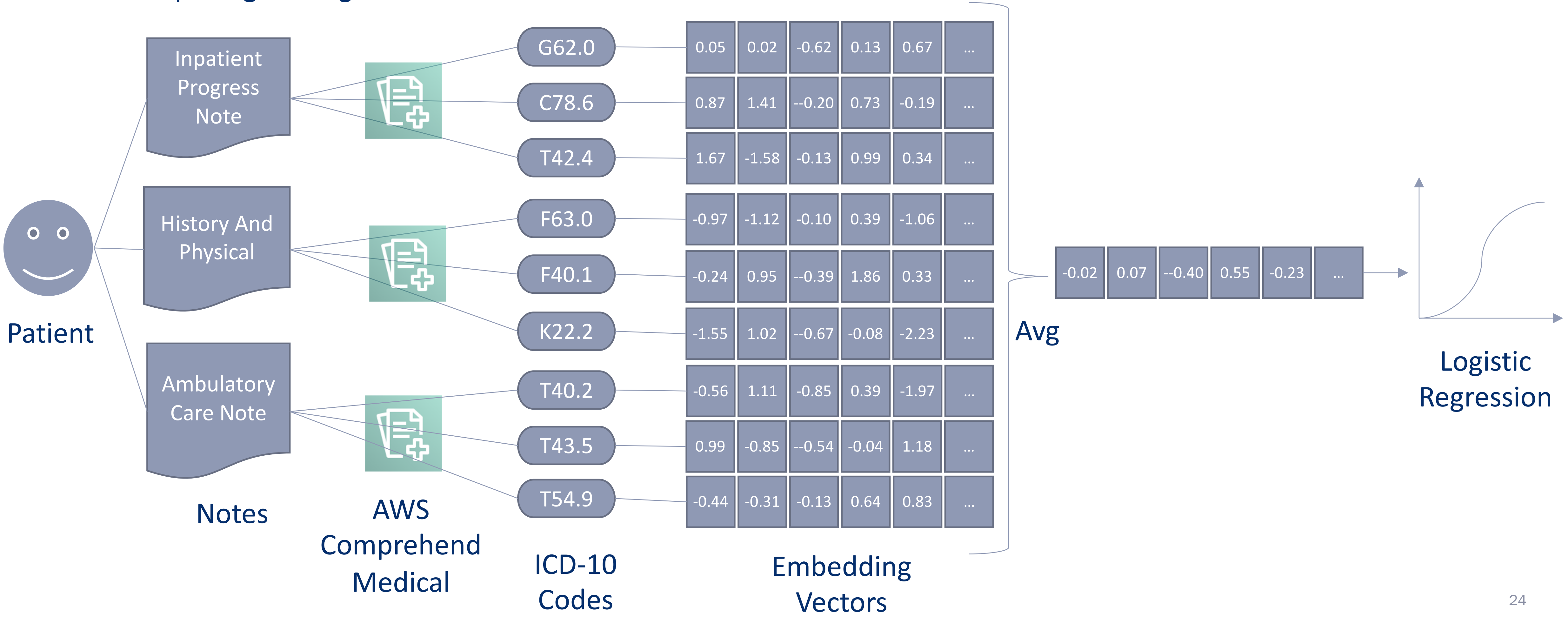
I10	Essential (primary) hypertension 0.0913 score
I11.9	Hypertensive heart disease without heart

You can find an exploratory analysis of AWS Comprehend Medical's capabilities, with code, [here].

Using Patient Notes

Extracting diagnoses from clinical notes yields many more diagnoses per patient than the billing data, though some of these are inaccurate, while the diagnoses in the billing data are close to 100% accurate.

We turned each patient’s note-derived diagnoses into embeddings the same way we did for billing diagnoses. Then we fit another simple logistic regression model on this data.



Patient

Notes: Inpatient Progress Note, History And Physical, Ambulatory Care Note

AWS Comprehend Medical

ICD-10 Codes

Embedding Vectors

G62.0	0.05	0.02	-0.62	0.13	0.67	…
C78.6	0.87	1.41	--0.20	0.73	-0.19	…
T42.4	1.67	-1.58	-0.13	0.99	0.34	…
F63.0	-0.97	-1.12	-0.10	0.39	-1.06	…
F40.1	-0.24	0.95	--0.39	1.86	0.33	…
K22.2	-1.55	1.02	--0.67	-0.08	-2.23	…
T40.2	-0.56	1.11	-0.85	0.39	-1.97	…
T43.5	0.99	-0.85	--0.54	-0.04	1.18	…
T54.9	-0.44	-0.31	-0.13	0.64	0.83	…

Avg

-0.02	0.07	--0.40	0.55	-0.23	…

Logistic Regression

Using Patient Notes

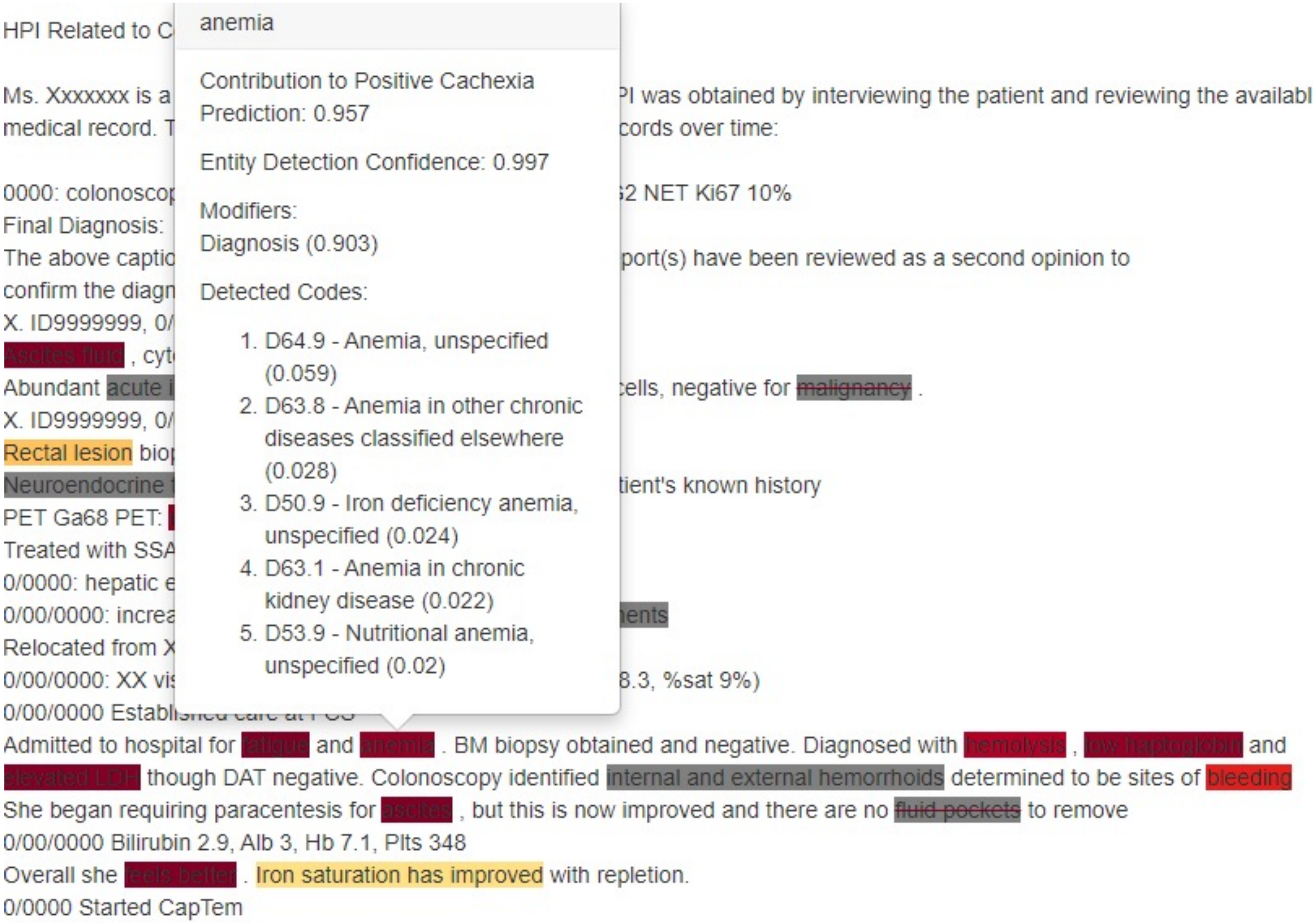
The model is 78% accurate on 424 patients held out during training:

	Has Cachexia Diagnosis	No Cachexia Diagnosis
Predicted Cachexic	246	73
Predicted Not Cachexic	22	83

Using Patient Notes

We calculated the contribution of each detected diagnosis to the model's predicted probability of cachexia and used that to highlight the note.

Darker red highlighting indicates a detected condition that increases the model's predicted probability of cachexia by a large amount. Lighter yellow highlighting indicates a detected condition that does not significantly increase the predicted probability of cachexia.
Grey highlighting indicates a detected diagnosis for which we do not have an embedding currently, and ~~strikethrough~~ indicates a condition that Comprehend Medical detected to be negated (e.g. "no fluid pockets"), hypothetical, or low confidence. These were not used by the model.



HPI Related to C

Ms. Xxxxxxx is a PI was obtained by interviewing the patient and reviewing the availabl
medical record. T cords over time:

0000: colonoscop ;2 NET Ki67 10%
Final Diagnosis:
The above captio port(s) have been reviewed as a second opinion to
confirm the diagn
X. ID9999999, 0/
Ascites fluid , cyt
Abundant acute i cells, negative for ~~malignancy~~ .
X. ID9999999, 0/
Rectal lesion bio
Neuroendocrine tient's known history
PET Ga68 PET:
Treated with SSA
0/0000: hepatic e
0/00/0000: increa ents
Relocated from X
0/00/0000: XX vis 8.3, %sat 9%)
0/00/0000 Establ
Admitted to hospital for fatigue and anemia . BM biopsy obtained and negative. Diagnosed with hemolysis , low haptoglobin and elevated LDH though DAT negative. Colonoscopy identified internal and external hemorrhoids determined to be sites of bleeding
She began requiring paracentesis for ascites , but this is now improved and there are no ~~fluid pockets~~ to remove
0/00/0000 Bilirubin 2.9, Alb 3, Hb 7.1, Plts 348
Overall she feels better . Iron saturation has improved with repletion.
0/0000 Started CapTem

anemia

Contribution to Positive Cachexia Prediction: 0.957

Entity Detection Confidence: 0.997

Modifiers:
Diagnosis (0.903)

Detected Codes:

1. D64.9 - Anemia, unspecified (0.059)
2. D63.8 - Anemia in other chronic diseases classified elsewhere (0.028)
3. D50.9 - Iron deficiency anemia, unspecified (0.024)
4. D63.1 - Anemia in chronic kidney disease (0.022)
5. D53.9 - Nutritional anemia, unspecified (0.02)

You can find an interactive version of this illustration here.

Conclusion

TL;DR

- Ontologies such as ICD-10 standardize medical concepts in a way that leverages domain expertise and facilitates the unification of disparate data sources.
- It makes a lot of sense to represent ICD-10 codes, and other medical ontologies, using embeddings. Embeddings capture semantic relationships and contextual meanings, and have much greater information density than one-hot-encoding.
- We identified claims where cachexia might have been left out using embedding representation of billed ICD-10 codes. This worked better than a rule-based approach of identifying cachexia using patient weight.
- We used AWS Comprehend Medical to associate ICD-10 codes to patient notes, encoded those using embeddings, then trained a cachexia classifier. This allows us to highlighting specific words and phrases in patient notes that may indicate cachexia, even if the word “cachexia” is not present.
- This is just one use case for ICD-10-CM embeddings, there are many others.

Future Work

- We are working on training ICD-10 code embeddings using Moffitt data.
    - We are seeking claims data sets to pre-train these embeddings. If anyone is aware of such a data set, we would love to hear about it.
    - A simple model implemented with keras produced good preliminary results, but the equivalent model in PyTorch is producing what looks like total nonsense. If anyone has experience porting keras models to PyTorch, please save me from a diagnosis of Z56.6!
- We would like to embed other ontologies (e.g. CPT procedure codes and NDC drug codes) into a single embedding space with ICD-10 codes. It seems reasonable that the diagnosis Z56.6 (mental strain related to work) and the procedure 90837 (60 minute individual psychotherapy session performed by a licensed mental health provider) contain similar information about a patient.
- It is intuitive to think about the embedding space such that the origin (0, 0, 0, ..., 0) represents perfect health and farther away from the origin corresponds to deteriorating health. E.g. the vectors for chronic kidney disease stage one and stage two both point away from the origin in the same direction, but stage two is further from the origin than stage one. But the embeddings do not work this way currently. We would like to experiment with imposing this constraint on the model while training embeddings.
- Generating accurate English descriptions of the learned embedding dimensions would further improve interpretability in downstream use cases of medical ontology embeddings, and there is some interesting recent work in this space.

Calls to Action

Please reach out if you are interested in

- Collaborating with our team on training ICD-10/CPT/NDC code embeddings
- Using ICD-10/CPT/NDC code embeddings trained on Moffitt data for your own downstream use cases
- Calling the AWS Comprehend Medical API from R
- Using the text-highlighting R package shown earlier in this presentation

Questions?