# Cluster analysis of low-dimensional medical concept representations from Electronic Health Records.

Fernando Jaume Santero
DS4DH group
Geneva, March 16th, 2022

UNIVERSITÉ DE GENÈVE
heg
Haute école de gestion Genève

# Summary

- **Introduction**
- **Medical concept representations**
- **Patient sequence embeddings**
- **Results**
- **Conclusions**

# Introduction

# Introduction

- Electronic Health Records (**EHR**) can be used to
  - **Monitor** and **diagnose** patients
  - Provide **personalized health** care
  - Explore **new treatments**

# Introduction

- Electronic Health Records (**EHR**) can be used to
  - **Monitor** and **diagnose** patients
  - Provide **personalized health** care
  - Explore **new treatments**

- However **EHR** data are very **heterogeneous** (categories, free-text, numerals, etc)

- There is a **scalability problem** when **experts** design new rule-based methodologies.

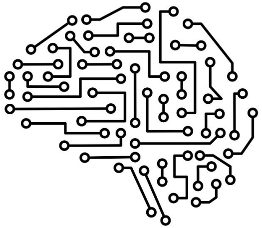
# Introduction

- Natural Language Processing (**NLP**) models designed to **extract information** from **documents**
- NLP can categorize and organize documents for **classification** and **translation** purposes
- Some NLP models **learn word associations** from text

# Introduction

- Natural Language Processing (**NLP**) models designed to **extract information** from **documents**
- NLP can categorize and organize documents for **classification** and **translation** purposes
- Some NLP models **learn word associations** from text

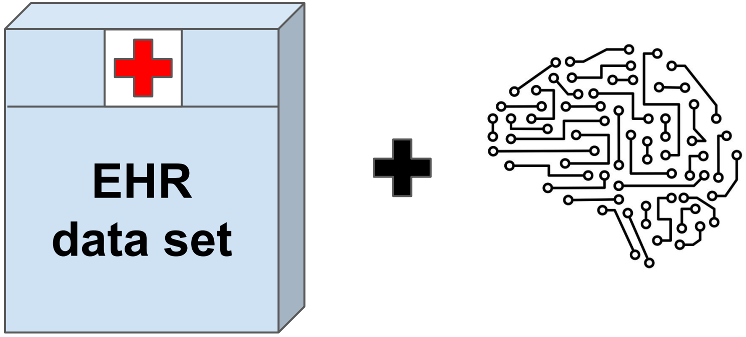
**Corpus**
(Wikipedia)



**Model**
(word2vec)

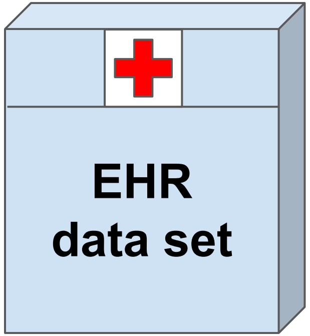
**King - Man + Woman = Queen**

# Introduction

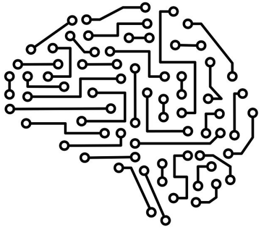



- **Heterogeneous data** set with health from hospitals
- **NLP models** that learn word associations

# Introduction




+



=

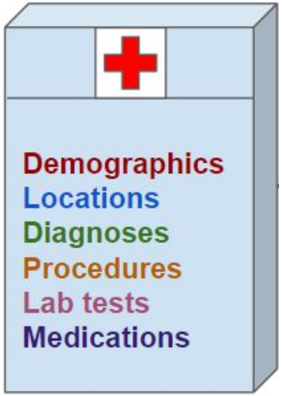
**Medical concept representations**
(in a low-dimensional vector space)

- **Heterogeneous data** set with health from hospitals
- **NLP models** that learn word associations
- **Main goal:**
    - **Study** of **relations** among **medical concepts** using NLP models.

# Medical concept representations

# Medical concept representations

EHR
(MIMIC-IV)

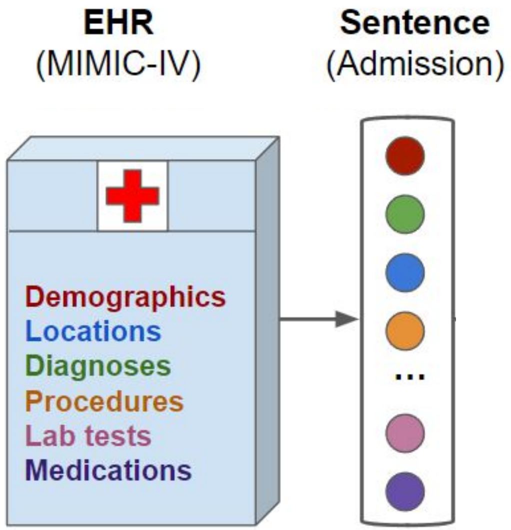



- **Extraction** of clinical information from MIMIC-IV

| Category | Labels | Description |
|---|---|---|
| Demographics | 14 | Gender, age, ethnicity, status after hospitalization |
| Locations | 36 | Locations within the hospital |
| Diagnoses | 19,735 | ICD-10 Clinical Modification |
| Procedures | 11,503 | ICD-10 Procedure Coding System |
| Lab tests | 929 | MIMIC-IV ItemID (OK: Normal, AB:Abnormal) |
| Medications | 4,770 | Generic Sequence Number |

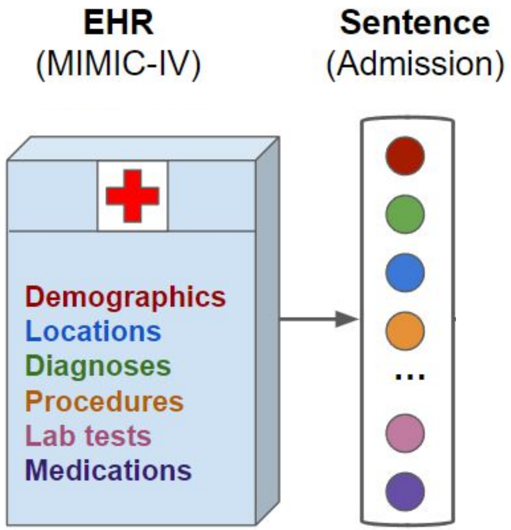
**Around 37,000 different medical concepts!**

# Medical concept representations




- **Extraction** of clinical information from MIMIC-IV
- **Sentence generation** from +500k hospital admissions

# Medical concept representations




- **Extraction** of clinical information from MIMIC-IV
- **Sentence generation** from +500k hospital admissions

**A sentence is generated for each admission**

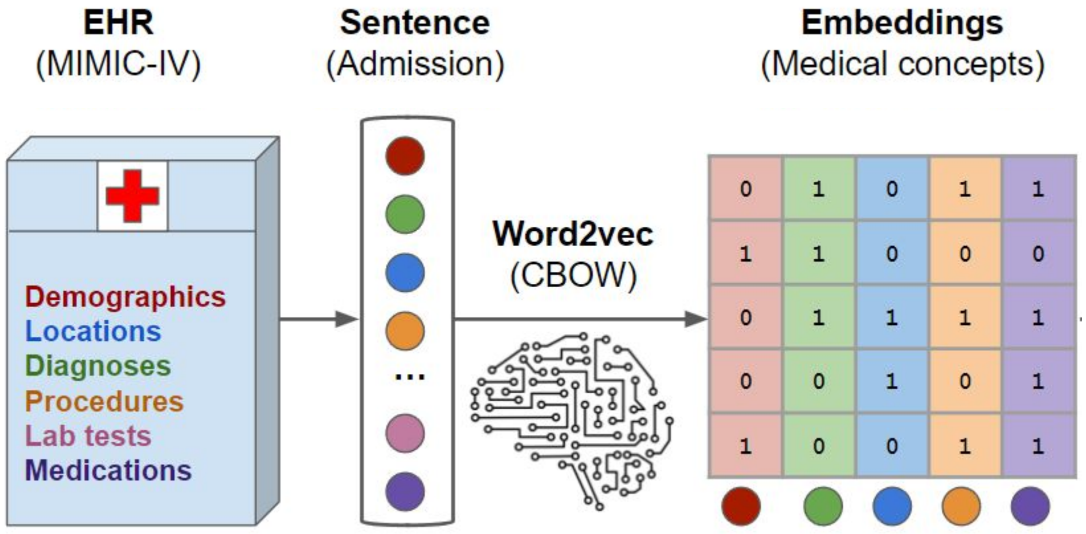
**Admission sentence example:**

**Female** patient **gave birth** with **epidural** in the **labor room**

Demographics ICD-10 GSN Location

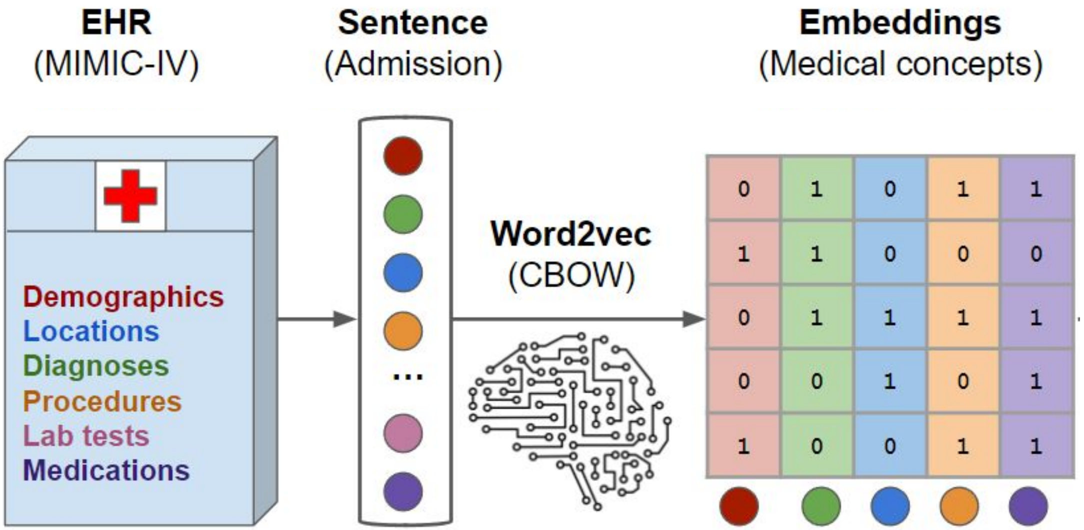
['F', 'Z3800', 'N01AH', 'LDR']

# Medical concept representations




- **Numerical** representations of medical concepts.
- Each concept has a numeric vector

# Medical concept representations




- **Numerical** representations of medical concepts.
- Each concept has a numeric vector

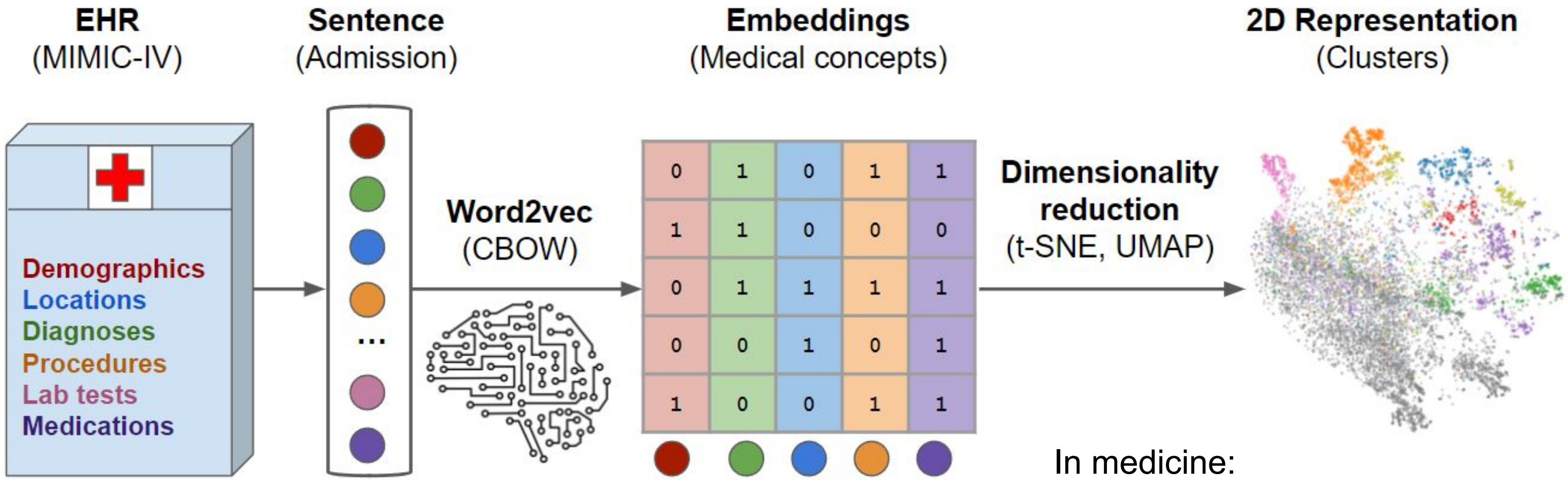
**Application example:**

In medicine:

**Patient sequence embeddings** →

- Diagnosis prediction
- Personal medicine

# Medical concept representations




In medicine:

**Patient sequence embeddings**

- Diagnosis prediction
- Personal medicine

# Patient sequence embeddings

# Patient sequence embeddings

- Patient sequence embeddings (**PSE**) generated by **aggregating** medical **concepts** vectors.
- PSE used to **predict** diagnosis, procedures and medications

# Patient sequence embeddings

- Patient sequence embeddings (**PSE**) generated by **aggregating** medical **concepts** vectors.
- PSE used to **predict** diagnosis, procedures and medications

**Example:**

**Female** patient **gave birth** with **epidural** in the **labor room**

Demographics ICD-10 GSN Location

`[‘F’, ‘Z3800’, ‘N01AH’, ‘LDR’]`

# Patient sequence embeddings

- Patient sequence embeddings (**PSE**) generated by **aggregating** medical **concepts** vectors.
- PSE used to **predict** diagnosis, procedures and medications

**Example:**

**Female** patient **< … >** with **epidural** in the **labor room**

Demographics

GSN

Location

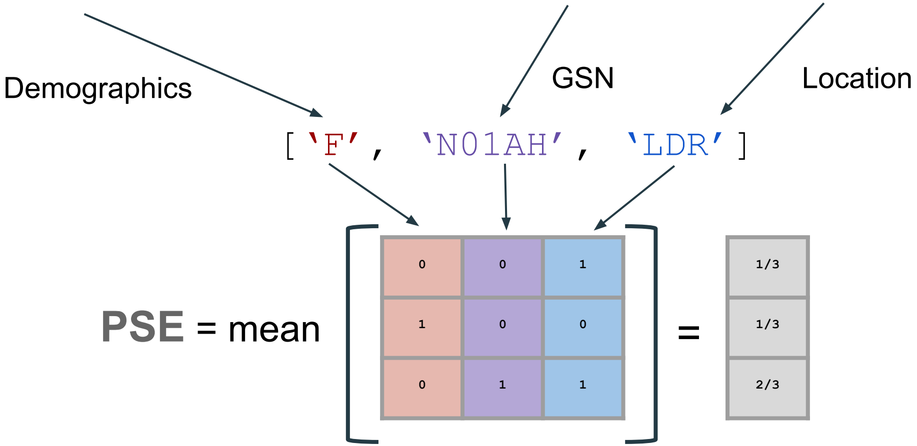
`['F', 'N01AH', 'LDR']`

# Patient sequence embeddings

- Patient sequence embeddings (**PSE**) generated by **aggregating** medical **concepts** vectors.
- PSE used to predict diagnosis, procedures and medications

**Example:**

**Female** patient **< … >** with **epidural** in the **labor room**



# Patient sequence embeddings

- Patient sequence embeddings (**PSE**) generated by **aggregating** medical **concepts** vectors.
- PSE used to **predict** diagnosis, procedures and medications

**Example:**

**Female** patient **< … >** with **epidural** in the **labor room**

**PSE**

**Labor & delivery**
**Diabetes**
**…**
**Broken limb**

**Score** PSE with all possible diagnoses using a given **metric** (e.g., cosine distance)

# Patient sequence embeddings

- Patient sequence embeddings (**PSE**) generated by **aggregating** medical **concepts** vectors.
- PSE used to **predict** diagnosis, procedures and medications

**Example:**

**Female** patient **< … >** with **epidural** in the **labor room**


Score PSE with all possible diagnoses using a given metric (e.g., cosine distance)

# Results

# Results

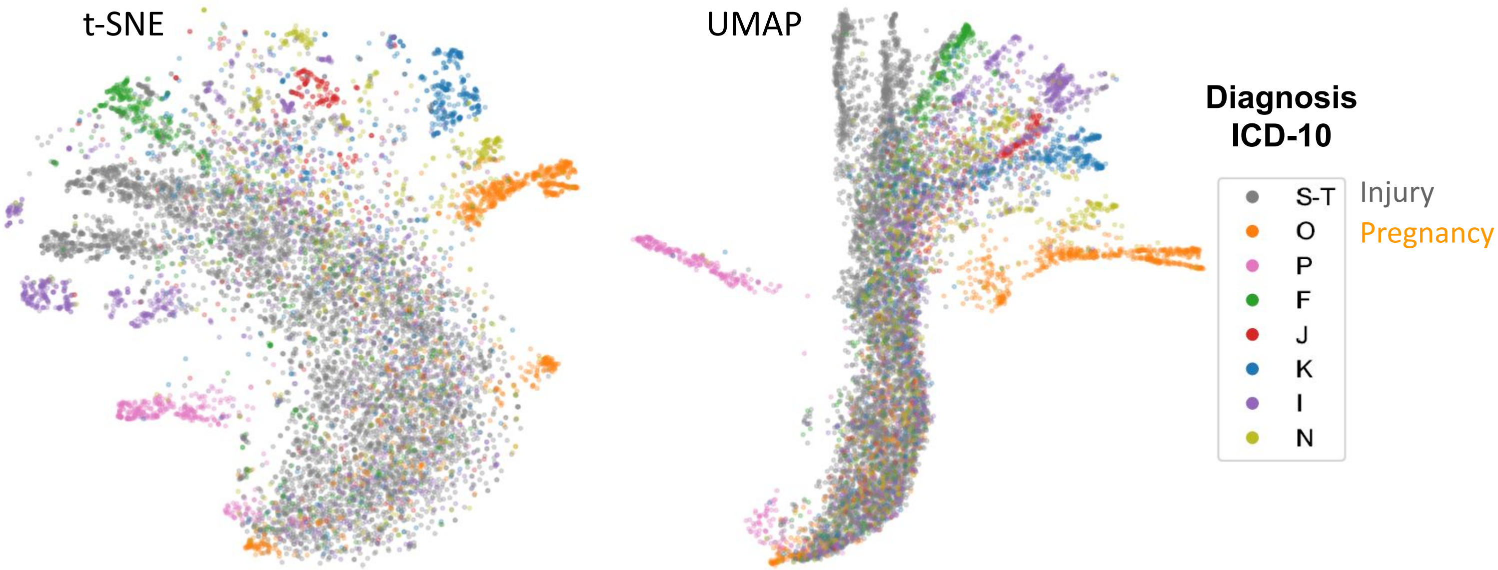
Do these models predict medical concepts correctly?

| Category | Concepts | Top 10 | Top 30 | Top 50 |
|---|---|---|---|---|
| Diagnoses | 19,735 | 47.07 % | 66.48 % | 72.74 % |
| Procedures | 11,503 | 58.46 % | 77.20 % | 83.82 % |
| Medications | 4,770 | 65.45 % | 80.45 % | 84.64 % |

**High prediction power** (accuracy) of **exact** medical **concepts**

# Results

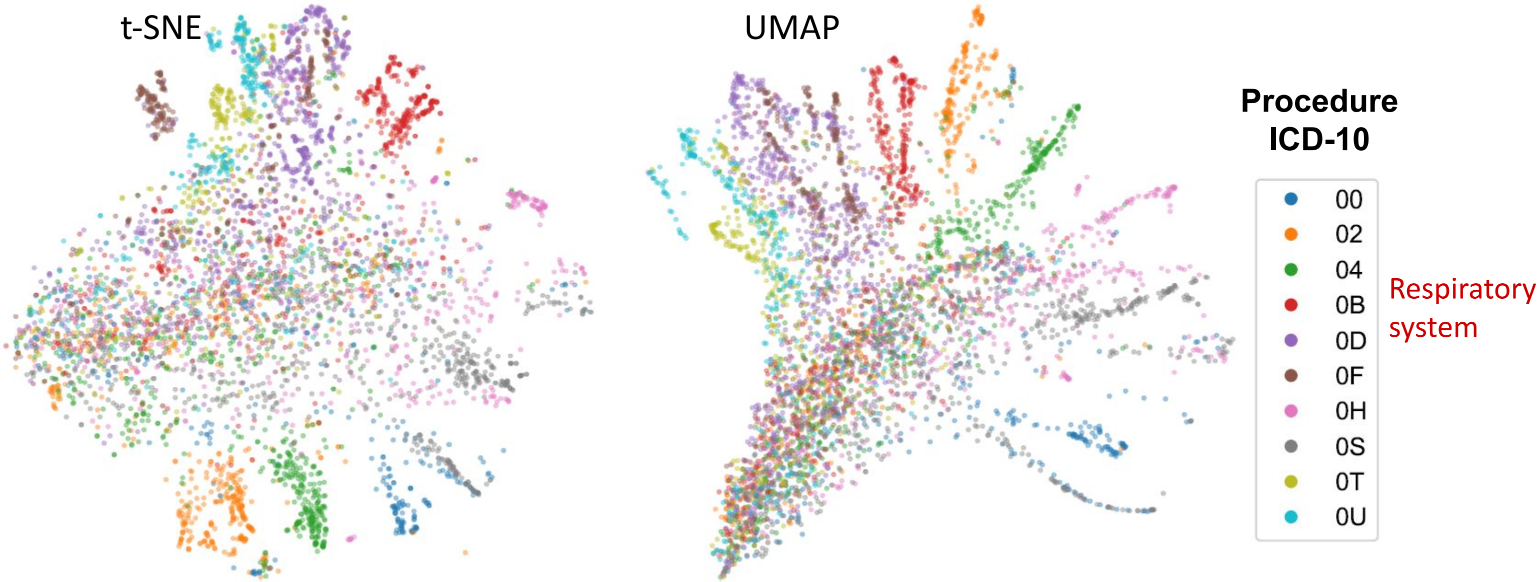
Are there relations between medical concept representations and their codes?




**Similar diagnoses** are **grouped together** matching the subcategories of ICD-10 codes

# Results

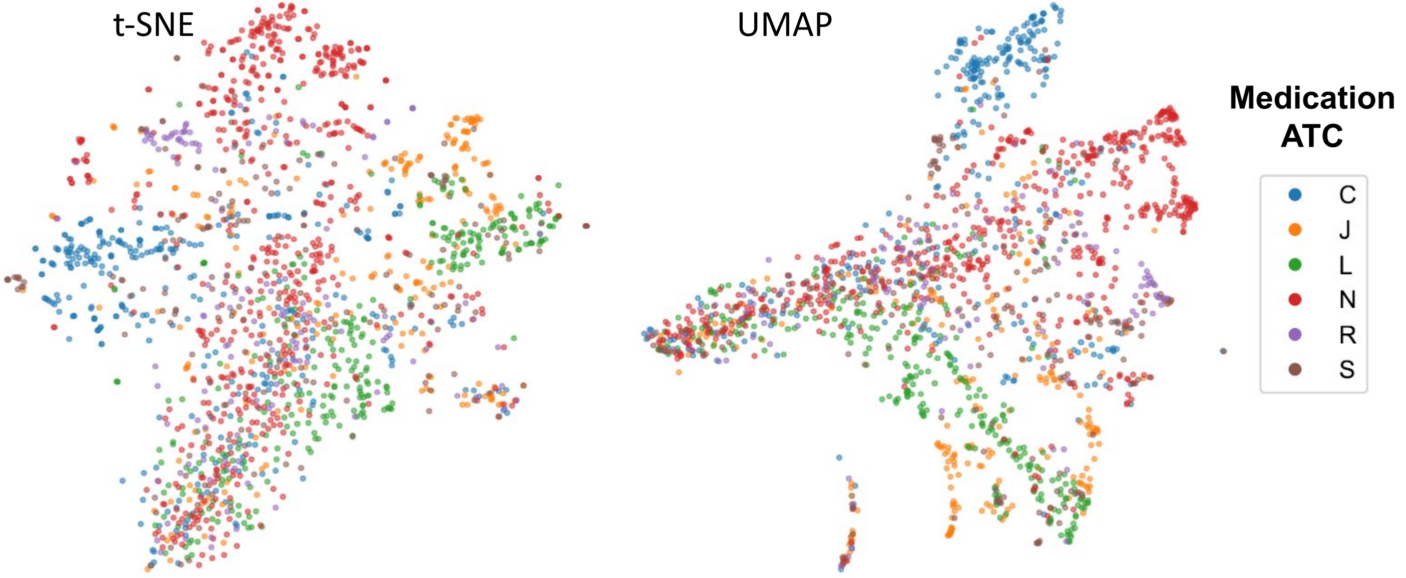
Are there relations between medical concept representations and their codes?




**Procedure representations** learn **body** parts where **surgical** operations ("0") take place.

# Results

Are there relations between medical concept representations and their codes?



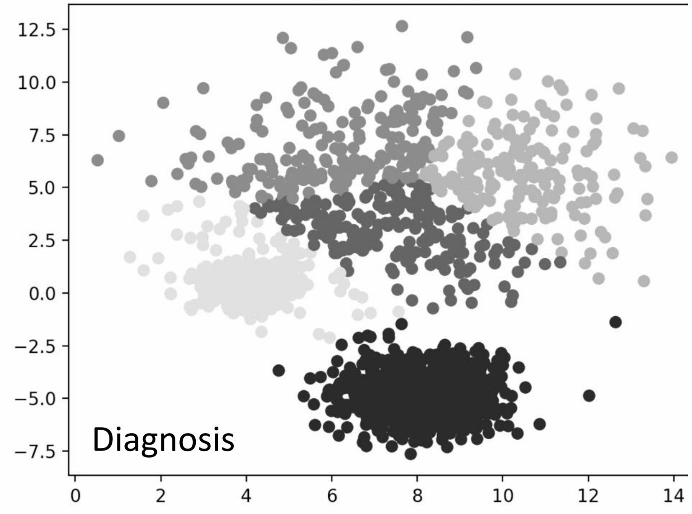

Consistent **match** between **medication** representations **and** their **anatomical** main **group**

# Results

- Are non-linear models such as word2vec necessary after all?
    - Study of relationships among different medical concepts using k-means

# Results

- Are non-linear models such as word2vec necessary after all?
  - Study of relationships among different medical concepts using k-means
  - **Example:** K-means clusters vs true label

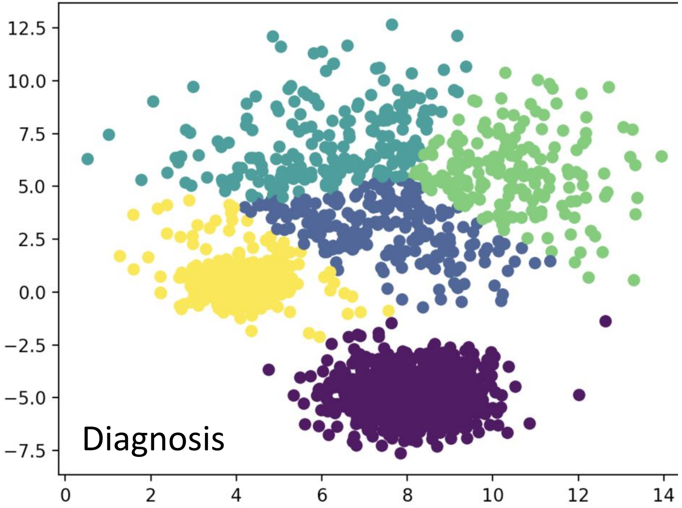



www.ml-science.com/k-means-clustering

1. Generate vector space

# Results

- Are non-linear models such as word2vec necessary after all?
    - Study of relationships among different medical concepts using k-means
    - **Example:** K-means clusters vs true label

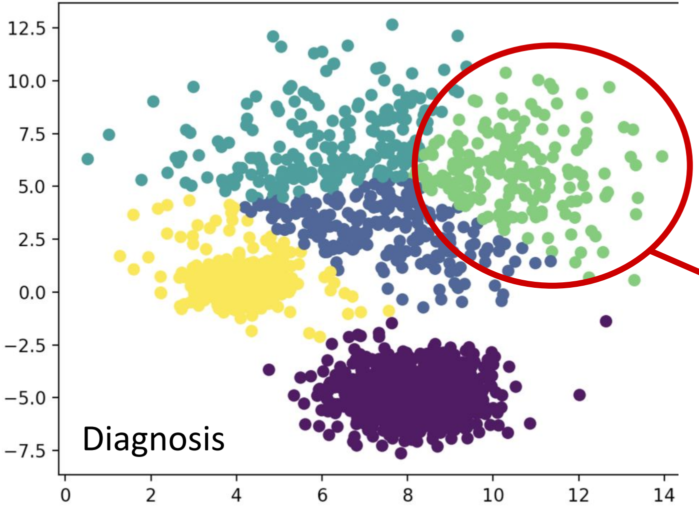



www.ml-science.com/k-means-clustering

1. Generate vector space
2. K-means clustering

# Results

- Are non-linear models such as word2vec necessary after all?
  - Study of relationships among different medical concepts using k-means
  - **Example:** K-means clusters vs true label




www.ml-science.com/k-means-clustering

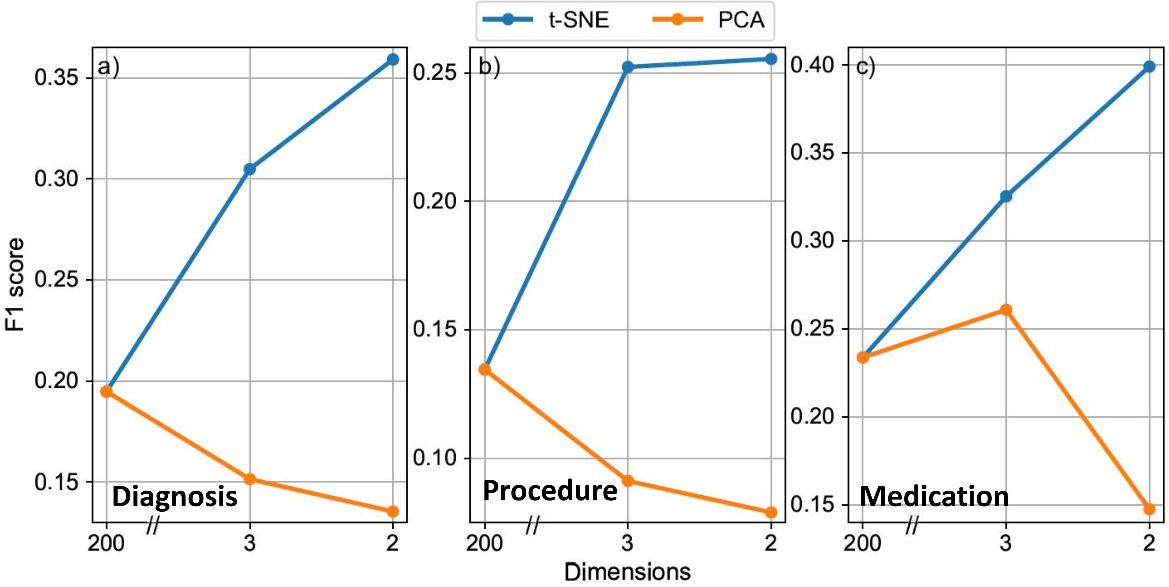
1. Generate vector space
2. K-means clustering
3. True label comparison

**Are all concepts from the same subcategory?**

**F1 score**
(the higher the better)

# Results

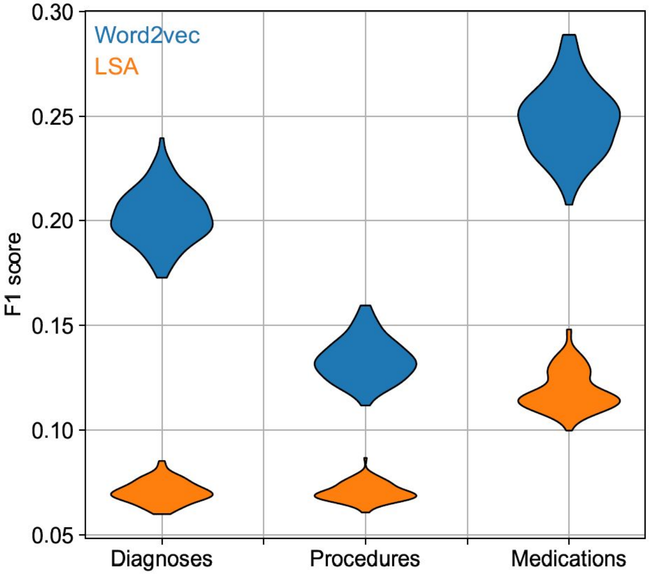
- Are non-linear models such as word2vec necessary after all?
  - Study of relationships among different medical concepts using k-means




  - **Non-linear** (t-SNE) **> Linear** (PCA) dimensionality-reduction methods

# Results

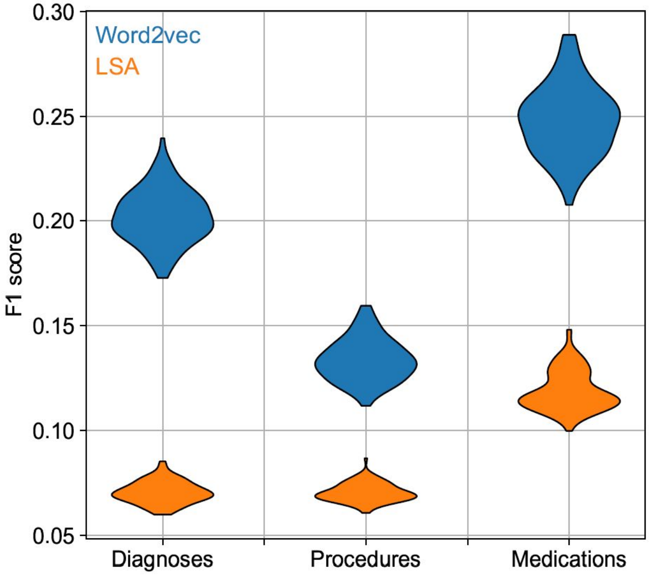
- Are non-linear models such as word2vec necessary after all?
  - Study of medical concept relationships: Linear (LSA) vs non-linear (word2vec)




- LSA stands for Latent Semantic Analysis
- **Linear representation** of medical concepts
- Co-occurrence matrix + Singular Value Decomposition

# Results

- Are non-linear models such as word2vec necessary after all?
    - Study of medical concept relationships: Linear (LSA) vs non-linear (word2vec)




- LSA stands for Latent Semantic Analysis
- **Linear representation** of medical concepts
- Co-occurrence matrix + Singular Value Decomposition
- **Word2vec** has **higher F1** scores than LSA
- **Relationships** among medical concepts are **non-linear**

# Conclusions

# Conclusions

- **Robust** numeric **representations** of medical concepts extracted from **electronic records**
- Representations exhibited **high predictive power**
- **Similar concepts** are located **nearby** within the vector space

# Conclusions

- **Robust** numeric **representations** of medical concepts extracted from **electronic records**
- Representations exhibited **high predictive power**
- **Similar concepts** are located **nearby** within the vector space
- **Complex relationships** among medical concepts
- **Importance** of using **non-linear models** such as word2vec