# Understanding Patients Better Using NLP Models

## But With No Text



Arpit Jain


Mathieu Sibué

**Faculty Advisor:** Dr. Robert Freund
**CVS Team:** Yannik Kumar, Sunanda Parthasarathy, Adam Saltz


**Project Timeline**

| Mar | Apr | May | Jun | Jul | Aug |
|---|---|---|---|---|---|
| Project Scoping, EDA, Doc2Vec analysis | | BERT pre-training | BERT & Word2Vec pre-training, validation on downstream tasks | | Handover |

CVS Health is the biggest retail pharmacy chain in the US with **10K** stores present in all the **50** states of the US

## Problem

**Prescription Personalization** (RxP) tasks are key to **increase prescription fills/pickups** and **cut return to stock costs**

These predictive tasks can benefit a lot from leveraging **patient longitudinal data** (transactions, outreach interactions)

However, **manually engineering** longitudinal data is **complex**, **time-consuming**, and might leave precious signal on the table

## Objective

**learn general-purpose patient features to improve performance and save dev. time on downstream tasks**

**Downstream tasks considered:**

- **FillNow** outreach success prediction
- **PickupNow** outreach success prediction

**Expected to impact** CVS Health Data Scientists, patients, and the broader CVS group itself

## Data

**Transactions** **Outreaches** **Patient Level**

### Analogy between patient sequences of events and text:

**Patient**

| Event type | Drug | Date |
|---|---|---|
| Fill | Albuterol | 10/21/20 |
| PickupNow outreach | Albuterol | 10/28/20 |
| Pickup | Albuterol | 10/28/20 |
| Fill | Amlodipine | 02/05/21 |

= {(Fill_Albuterol, 10/21/20), (PickupNow_Albuterol, 10/28/20), (Pickup_Albuterol, 10/28/20),…}

**Analogy**

- word **co-occurrences** convey meaning
- word **order** is important
- similar **long-tailed word distributions** observed

**Sentence**

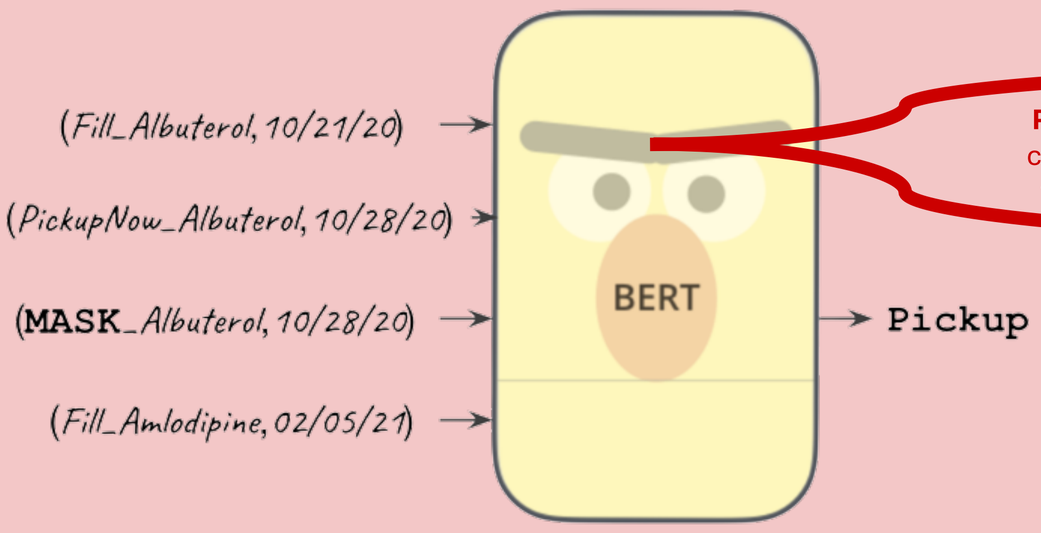
| Word | Position |
|---|---|
| the | 0 |
| cat | 1 |
| sat | 2 |
| on | 3 |

= {(the, 0), (cat, 1), (sat, 2),…}

## Methodology

### Step 1 - Unsupervised NLP-like pre-training

**Masked Language Modeling** over pharmaceutical events by **pre-training** a custom **BERT model**

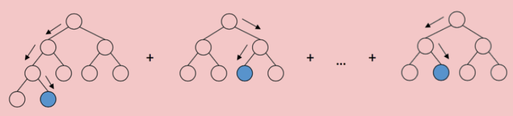



**Patient embeddings** created using our pre-trained BERT model

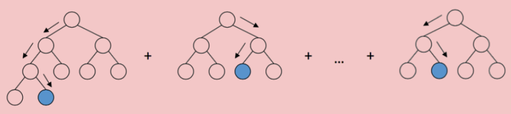
### Step 2 - Downstream tasks training

**Using the embeddings generated** by the pre-trained model (and other features) as inputs, we **train downstream task models**

**FillNow** success prediction



**PickupNow** success prediction



**Performance on Downstream Tasks:**

| Experiment | FillNow ROC AUC | PickupNow ROC AUC |
|---|---|---|
| Without Embeddings | 64.4 | 72 |
| With Embeddings | **65.7** | **83.9** |
| | **+2%** | **+17%** |

## Business Impact

After using patient embeddings on our downstream tasks, we estimate an impact of:

**10%**
**Avg. downstream performance increase**

**$1M/yr**
**Increase in Direct Revenue**

**40h/project**
**Time Saved for Data Scientists in RxP team**

## Next Steps

Exploring **pre-trained model explainability**

**Fine-tuning pre-trained BERT** model on downstream tasks

## Our Contributions

**Benchmarked several models** for embedding generation

Successfully **improved upon the baselines** to generate business impact

Created **custom Python Package** with our code