# PRACTICE FUSION EXERCISE

Rose Glavin. Email rglavin2@usfca.edu

# DISCLAIMERS

- I am using this exercise to show skills and ability to think critically
- This is a proof of concept only.  I did not refine the model  or visualization
- I tried to show what tools I could leverage to gained insights about a particular condition given this dataset

# PRIOR TO MODEL CREATION

- Exploratory Data Analysis
  - Interested in chronic disease, access to biometrics and longitudinal data
  - Need to establish availability and quality/consistency
- Availability
  - Longitudinal : dataset covers a short window for chronic disease
  - Wide range of diagnoses, medications, labs
  - Wide range of providers
  - Good number of patients

# DATA QUALITY – HYPOTHYROIDISM

- Demographics and Diagnosis : Looks consistent.
  - Mainly women. Middle aged. Population in general : older
- Patients – 11% hypothyroid compared to 5% random national dataset
- Transcript – inconsistent height, BMI, BP, temp
- Medications – High use of antibiotics. Otherwise looks consistent
- Labs – missing data ?
  - > 1000 hyperthyroid patients . Labs :only 18 TSH , 62 T4. Ratio seems incorrect

# NLP MODEL FOR HYPOTHYROIDISM

- Build a Language Model specific to medical language
- Train a classifier for hypothyroidism on the LM
- Evidence of latent constellations present within diagnosis?
- This is self-referential but
  - It proves the concept of using diagnosis as a label and free text as the independent variable
  - Potential to use LM to find biometrics in the free text

# METHOD

- Collect all Diagnosis Description and Lab HL7Text from PF dataset (R,Tidyverse)
- Collect all hypothyroid patient by looking for "hypothyroid" in Diagnosis
- Sample roughly same number of non-hypothyroid patients
- Create a labelled dataset
  - Target: Hypothyroid status
  - Input : Diagnosis and Lab text ( 2 classifiers, 1 "hypothyroid" omitted in training)
  - Reserve 100 samples for testing
- Transfer learn  diagnosis text onto a wiki trained LM (fast.ai, colab)
- Build a classifier on top of this LM (fast.ai, colab)

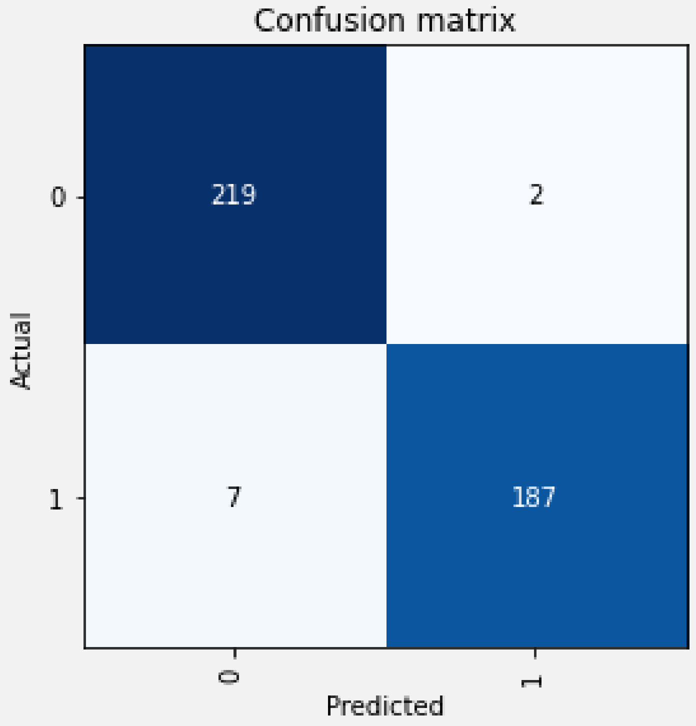
# RESULTS

Accuracy

- 57% when "hypothyroid" removed
- 98% when "hypothyroid" included

Interpretability

- Heatmaps show the words that are important



# DISCUSSION

- Clinical guideline labs missing for most hypothyroid patients . Why ? Is this dataset missing labs ?
- Correlation between diagnoses can drive a feedback loop
- Biometrics, and longitudinal data necessary for meaningful new insights
- The heatmap output provides interesting latent insight
- The LM learned words associated with a diagnosis

# CONCLUSION

- Either hypothyroidism is being poorly managed or
- Dataset is not suitable for studying hypothyroidism
- At a first glance – other chronic diseases likely to have similar issues
- Language Models can be a useful tool to see latent content
- Potential to extend the idea with more free text data
- Important because patients are getting labs directly
- Ability to enrich the dataset with patient-provided data