## Background

- Medical records are terse notes recorded by a doctor. Includes information such as:
  - Past medical history: Ex. "Patient is 50 yo w/ T2D. Prey received RRAA..."
  - Hospital course: Ex. "Presented hypertensive and UTI. He/she was treated with cipro, levofoxacin. Weeks 101, 106, 139/156.
  - Diagnoses, medications prescribed, discharge condition.
- Records are labeled with ICD codes, a comprehensive system which labels symptoms, diseases, medical procedures performed, etc:
  - ICD-10-AM
  - U10: Hypertensive heart and peripheral vascular disease.
  - U11: Hypertensive heart disease.
- We used the MIMIC-III dataset, which has emergency room records labeled with codes by doctors.

## Challenges

- Recent work has not produced usable results (~15-50 F1, depending on # of codes).
- Many features are hard-engineered rules for specific codes.
- Large # of labels (over 1,000 and 200 codes).
- Label distribution is very skewed, many codes have little training support.
  - Only 10% of labels in the top 500 codes come out of ~2000 in our dataset!
  - About 100 of those codes have >1 training example.
- Labels are extremely noisy.
  - One study showed that for Alzheimer’s disease, 65% of notes were not coded (false negative) and 5% coded without evidence in notes. Similar numbers for other diseases (Parkinson’s 75%, Breast 31% / 39%).
- Abbreviations, acronyms, and medical jargon are widespread.
  - Very domain-specific language, so we probably can’t use generic models trained on abundant sources of data such as Wikipedia.

## Baseline Model

- We trained a bag-of-words/phrases model with a multi-label linear classifier (logistic regression + L1 regularization).
  - Chosen for ease of explainability.
  - Short phrases were often more descriptive than individual words (ex. "renal tract infection")
- Results: 11 labels: 65% F1 score / 15 labels: 45% F1 score
- Learned features were generally quite relevant. Top features for "asthma":
  - Asthma related, singular, inhalers, fluticasone, mepolizumab, inhaler, inhalers.
- Revealed:
  - Sparsity: Relevant information is typically contained in 1-2 sentences in the entire document.
  - Multi-word understanding: Some labels are inspired by word orders, not just described.
  - We performed nearly worse on some codes with little training support.

## Full Model

- To tackle the first two problems, we have chosen to implement an simple attention-based model.
  - Sparsity
    - The document is segmented, and each segment is scored using either an RNN or a bag-of-words model. We use either the top scoring segment (non-differentiable, but more computationally efficient) or use the lowest as a weight for the classification model.
  - Multi-word understanding
    - The segment is encoded as word vectors, and we train either an RNN to classify the code from the segment.

## Conclusions/Future Work

- Problems left to tackle after this project:
  - Explaining hierarchy of codes
    - This has been explored in some recent work, and some ideas can be incorporated into our models.
  - Using word knowledge (ex. Wikipedia, code descriptions) to predict codes with little training support.
    - However, it is not clear whether there is significant practical benefit to predict rare codes accurately.
  - Noisy/Incomplete labels. We hope to explore bootstrapping methods and semi-supervised learning to deal with this.
    - This is a serious, and known problem.