# Paper: Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data

Link: <https://psb.stanford.edu/psb-online/proceedings/psb20/Beam.pdf>

**Motivation**

What is the problem being solved?

* To construct a comprehensive set of embeddings for medical concepts
* By combining extremely large sources of multimodal healthcare data
* e.g. insurance claims, clinical notes and biomedical journals

Why is it important?

There is a lack of pre-trained resources for applications in healthcare and medicine relative to other areas of machine learning and NLP

What previous work exists?

* Word2vec
  + CBOW model: predicts the probability of the target word given its context defined within a window
  + Skip-gram model: predicts the surrounding context given the target word.

[Que: Why does skip-gram model consider conditional probability p(w|c) instead of p(c|w)?]

<https://github.com/rusheniii/LargeScaleClinicalEmbedding>

* GloVe: produces word embeddings by fitting a weighted log-linear model to co-occurrence statistics

Why is the previous work insufficient to solve the problem?

Healthcare data come in a variety of forms, but previous methods were originally developed only for text

**Approach (Cui2vec: 500-dimensional word2vec embedding on combined data)**

Mechanism

* Map all of concepts into a common concept unique identifier space (CUI)
* Construct CUI-CUI co-occurrence matrix
  + Non- clinical text data: preprocess -> chunk into fixed length windows of 10 words ->co-occurrence = the appearance of a CUI-CUI pair in the same window.
  + Claims data: ICD-9 codes are mapped to UMLS CUIs -> co-occurrence = the number of patients in which two CUIs appear in any 30-day period.
  + Clinical notes: co-occurrence = two CUIs appearing in the same 30-day `bin'
* Factor co-occurence matrix by GloVe / transformed into a SSPMI matrix and factor using SVD to create word2vec embeddings.
* Preprocessing: Word embeddings operate on tokens, and many medical concepts can span multiple tokens -> collapse multi-word concepts into a single token
* Benchmark: To leverage previous “known” relationships between medical concepts
  + Cosine similarity: may incorrectly penalized embeddings with higher ranking than query concept
  + Statistical power:
    - Compute the null distribution of scores by drawing 10,000 bootstrap samples (x\*, y\*) where x\* and y\* belong to the same category as x and y, respectively.
    - Compare the observed score and declare statistical significance if p< 0.05
    - Calculate the statistical power to reject H0: no relationship

Sources

* an un-identifiable claims database from a nationwide US health insurance plan with 60 million members over the period of 2008-2015
* a dataset of concept co-occurrences from 20 million notes at Stanford
* an open access collection of 1.7 million full text journal articles obtained from PubMed Central

**Results**

Cui2vec embedding results

* 500-dimensional embeddings for 108,477 unique concepts from combined data
* Most of the concepts appear in only one type of sources

Previous work comparison

* 300-dimensional embeddings for 15,905 concepts from claims
* 300-dimension embeddings for 28,394 concepts from clinical notes
* 200-dimensional embeddings for 59,266 concepts derived from PubMed abstracts

=> Cui2vec generally outperformed in terms of benchmarks

**Contributions**

|  | Current work | Previous work |
| --- | --- | --- |
| Data size | Larger | Small, Local |
| Embeddings to data source | map all concepts into a common co-occurrence space to produce a single set of embeddings that can be used on tasks with different kinds of clinical data | Different embeddings for each data source |
| Interpretability | Higher | Lower |
| Performance | Better | Worse |

Usefulness of Cui2vec

* Most healthcare data is unlabeled/weakly labeled -> ability to extract meaningful structure in an unsupervised manner is extremely important -> Cui2vec is successful in terms of benchmark measures [Que: I assume the so-called meaningful structure is measured by the benchmarks?]
* Most sources of healthcare data are not easily shareable -> limits some researchers to small sources of local data -> Cui2vec is created using large and national sources of healthcare data

**Other**

What aspects of the paper were unclear?

* The results section in the paper is a bit confusing, in particular 4.1 Benchmark results. To my understanding, they were trying to compare the performances of GloVe, word2vec, and PCA in terms of benchmarks. But Cui2vec is not included as one of the compared methods. It is unclear what’s the purpose for including this section.
* “Spearman correlation between human assessments of concept similarity and cosine similarity from the embeddings”: Did they bring in medical experts to conduct human assessments? I think the motivation to consider this correlation is to check to what extent the embedding approach can replicate human assessments?

What aspects of the paper did you enjoy and not enjoy?

* I enjoy reading how each embedding approach and power benchmarking method were constructed (the math)