Deep Learning For
Medical Knowledge Extraction From
Unstructured Biomedical Text*

Andrew Beam, PhD
Postdoctoral Fellow
Department of Biomedical Informatics
Harvard Medical School
05/10/2017

*work in progress
AI has the potential to fundamentally change healthcare and medicine…

… but how do we measure the progress of AI for general medical diagnosis*?

*outside of medical imaging
MDs often serve as **the** comparison for medical AI, but setting up a **fair** comparison is harder than it seems.

The Doctor Baseline

Doctors Don’t Predict

- Doctors *don’t*:
  - Predict appearance of diagnoses in the future
  - Provide calibrated probabilities
  - Optimize for AUC

- Doctors *do*:
  - Infer **current** disease state given symptoms
  - Triage patients given current estimate of disease state
Doctors Disagree

- Doctors often disagree about the correct diagnosis for a given patient
- Even the correct list of diagnoses to consider (e.g., the differential) is often not unanimous
- Thus, an objective “gold standard” dataset of labeled patients can be very hard to create in some instances.
**Healthcare Data is Messy**

- In most healthcare data (e.g. EHR/claims) you don’t observe the disease process directly, but instead the process of healthcare dynamics.
- Information leakage is inevitable.
- Doctor reasoning process is “baked in”, can’t take the doctor out of the data.
- How will an AI system trained on one EHR generalize to a new one?

Image credit: Griffin Weber, MD/PhD
Desirable Benchmark Properties

- Clarity: Unambiguous gold standard
- Portability: Easy to compare results across different healthcare environments and populations
- Comparability: Available metrics of human performance

Goal: Task that doctors actually do that also meets these criteria
USMLE STEP 1

United States Medical Licensing Examination

Exam administered in 3 “steps”
- Step 1 is taken after the 2nd year of medical school
- Requires several months of dedicated study
- Tests understanding of fundamentals of biology and clinical medicine
- Multiple-choice format
- Large influence on residency placement
- “SAT” for med students

Necessary (but not sufficient) condition for becoming a physician
STEP 1 AND AI

Step 1 is an attractive benchmark for medical AI
- Requires broad knowledge of medicine and biology
- Unambiguous right/wrong answers (clarity)
- Potentially free from healthcare data “messiness” (portability)
- 25,000 medical students take it each year -> good human performance numbers (comparability)
- It’s hard and will require methodological innovation
- **Con:** Unclear road to clinical tool
Can we train a deep learning system capable of passing step 1?

Step 1 Question
A full-term female newborn is examined shortly after birth … Which of the following mechanisms best explains this cytogenetic abnormality?

Answers
(A) Nondisjunction in mitosis
(B) Reciprocal translocation
(C) Robertsonian translocation
(D) Skewed X-inactivation
(E) Uniparental disomy
DATA RESOURCES

Biomedical Journal Articles
- PMC Open Access – 1.7M
- Elsevier – 2M
- Springer – 500K

Physician References
- Merck Manuals
- Mayo Clinic Disease Library
- MEDLINE
- DynaMed
- Emedicine/Medscape

Biomedical Knowledge Commons
- 4.3M articles
- 50,000 pages of reference material
- 15,000 flash cards
- Dozens of books
- 10,000 Step 1 style questions

All preprocessed and normalized against a common medical thesaurus

Test Preparation
- Flash cards
- High Yield Concept List
- Books

Step 1 Questions
- Open Osmosis
- Library Resources
- NBME
DATA PREPROCESSING

Raw Text → Normalization → MED2VEC
MED2VEC

What can we learn about medical concepts from 4.3 million journal articles?
MED2VEC

Query bronchopulmonary dysplasia

Compute Similarity
Medical Concept Vector Database

<table>
<thead>
<tr>
<th>CUI</th>
<th>SemanticType</th>
<th>String</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3872829</td>
<td>Biologically Active Substance</td>
<td>Adhesion protein</td>
<td>0.0056652557</td>
<td>-0.196249962</td>
</tr>
<tr>
<td>C3872700</td>
<td>Health Care Related Organization</td>
<td>Clinical pathology service</td>
<td>-0.4462876022</td>
<td>-0.013221873</td>
</tr>
<tr>
<td>C3872595</td>
<td>Nucleic Acid, Nucleoside, or Nucleotide</td>
<td>Human papillomavirus DNA</td>
<td>-0.0793692693</td>
<td>-0.198981807</td>
</tr>
<tr>
<td>C3872595</td>
<td>Biologically Active Substance</td>
<td>Human papillomavirus DNA</td>
<td>-0.0793692693</td>
<td>-0.198981807</td>
</tr>
<tr>
<td>C3872494</td>
<td>Manufactured Object</td>
<td>Device tip (physical object)</td>
<td>-0.1605362286</td>
<td>-0.325420946</td>
</tr>
<tr>
<td>C3872476</td>
<td>Medical Device</td>
<td>Body reference point marker</td>
<td>-0.0516590886</td>
<td>0.078756697</td>
</tr>
<tr>
<td>C3871203</td>
<td>Temporal Concept</td>
<td>At discharge</td>
<td>0.1280273497</td>
<td>0.100451143</td>
</tr>
<tr>
<td>C3864436</td>
<td>Medical Device</td>
<td>Anatomical structure separator</td>
<td>-0.0629289672</td>
<td>0.406454623</td>
</tr>
<tr>
<td>C3856907</td>
<td>Manufactured Object</td>
<td>Projector</td>
<td>0.1593338698</td>
<td>0.359946638</td>
</tr>
</tbody>
</table>

60,000 medical concepts
WHAT DRUGS ARE USED FOR BPD?

**Query**
bronchopulmonary dysplasia

**Filter**
Pharmacologic Substance

**Rank**

<table>
<thead>
<tr>
<th>String</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulmonary Surfactants</td>
<td>0.3964883</td>
</tr>
<tr>
<td>palivizumab</td>
<td>0.3360302</td>
</tr>
</tbody>
</table>
HOW IS BPD MANAGED?

Query
bronchopulmonary dysplasia

Filter
Therapeutic or Preventive Procedure

Rank

<table>
<thead>
<tr>
<th>String</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygen Therapy Care</td>
<td>0.5373769</td>
</tr>
<tr>
<td>Mechanical ventilation</td>
<td>0.5018497</td>
</tr>
<tr>
<td>Intermittent Positive-Pressure Ventilation</td>
<td>0.4676242</td>
</tr>
<tr>
<td>High frequency oscillatory ventilation</td>
<td>0.4653350</td>
</tr>
<tr>
<td>Noninvasive Ventilation</td>
<td>0.4361196</td>
</tr>
</tbody>
</table>
DEEP LEARNING FOR QA

**Approach:** Deep neural network that maps word vectors in question -> correct answer

End-to-end deep learning QA systems need 100k – 1M QA pairs.

Existing SOTA operate in an “easier” domain (e.g. Who is Obama’s wife?)

10,000 questions are not enough. We need a way to generate more questions.
SYNTHETIC QUESTIONS

Scan through entire corpus

Extract Potential QA pair

Using UMLS NLP/POS tagger:
- Tag noun-phrases that mention medical concepts as potential answers
- Surrounding sentences as potential question
- Each QA pair becomes a potential fill in the blank question.

Score Synthetic QA Pairs

Compare semantic similarity of synthetic QA pairs against real ones.
Only keep high scoring synthetic QA pairs.

Results: 1 billion potential QA pairs
Q: It is associated with notching of the ribs because of collateral circulation hypertension in the upper extremities and weak pulses in the lower extremities. ____ is most likely the result of the extension of a muscular artery ductus arteriosus into an elastic artery aorta during fetal life where the contraction and fibrosis of the ductus arteriosus upon birth subsequently narrows the aortic lumen.

Answer: Postductal Coarctation

Work is on going!
CONCLUSIONS

- Thoughtful metrics of progress for medical AI are vitally important
- Head to head comparisons with doctors can be tricky
- Step 1 may be a good benchmark for medical AI
- Unsupervised learning on large sources of biomedical text can automatically extract relationships between medical concepts
- Deep learning has promise for answering step 1 questions
ACKNOWLEDGEMENTS

**Harvard Medical School**
- Inbar Fried
- Sam Finlayson
- Nathan Palmer
- Isaac Kohane

**Google Brain**
- Jasper Snoek
- Alex Wiltschko

**Funding**
- Robert Wood Johnson Foundation

**Hardware**
- NVIDIA®

**Data**
- Open Osmosis
- MAYO CLINIC
- NBME
- MERCK MANUAL

@AndrewLBeam