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

HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics

@AndrewLBeam

AI & MEDICINE

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

THE DOCTOR BASELINE

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

!=

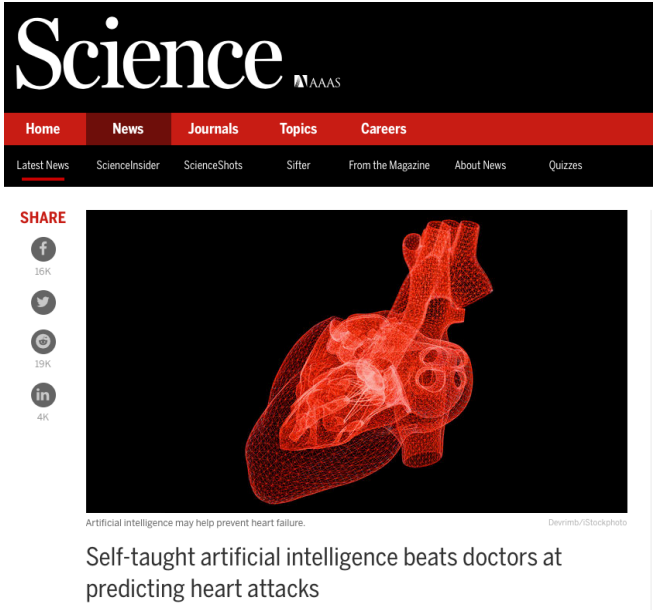
Image credit: http://www.bbc.com/news/magazine-28166019

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



THE DOCTOR BASELINE

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.

The Accuracy and Interobserver Agreement in Detecting the 'Gallop Sounds' by Cardiac Auscultation*

Charmaine E. Lok, MD; Christopher D. Morgan, MD; and Narasimhan Ranganathan, MD

Study objectives: To determine the observer accuracy and interobserver agreement in identifying S_4 and S_3 by cardiac auscultation and whether they improve with increasing observer experience.
Design: Prospective, blinded study.
Setting: Cardiology and general internal medicine wards in a university-affiliated teaching hospital.
Patients: Forty patients with a cardiac diagnosis and 6 patients without were studied.
Measurements and results: Two cardiologists, one general internist, three senior and two junior postgraduate internal medicine trainees, blinded to the patients' characteristics, examined the patients and documented their findings on a questionnaire. Computerized phonocardiogram was obtained in all patients as a gold standard and was interpreted by a blinded, independent cardiologist. The mean positive predictive values for S_4 and S_3 were 51% (range, 24 to 100%) and 71% (range, 50 to 88%), respectively. The mean negative predictive values for S_4 and S_3 were 82% (range, 67 to 94%) and 64% (range, 56 to 85%), respectively. The overall interobserver agreements for detecting S_4 was $K = 0.05$ (95% confidence interval [CI], 0.01 to 0.09) and S_3 was $K = 0.18$ (95% CI, 0.13 to 0.24). There was no apparent trend in the accuracy or interobserver agreement with regard to the level of observer experience.

Conclusion: The agreement between observers and the phonocardiographic gold standard in the correct identification of S_4 and S_3 was poor and the lack of agreement did not appear to be a function of the experience of the observers. The overall interobserver agreement for the detection of either S_4 or S_3 was little better than chance alone.

1998; 114:1283–1288)

phonocardiogram; PPV = positive predictive value; PCG = phonocardiogram; PPV = positive predictive value

THE DOCTOR BASELINE

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?

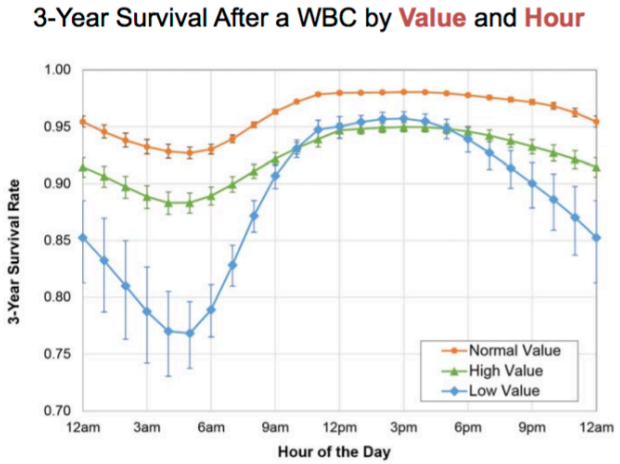


Image credit: Griffin Weber, MD/PhD

BENCHMARKING MEDICAL AI

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 **S**tates **M**edical **L**icensing **E**xamination

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

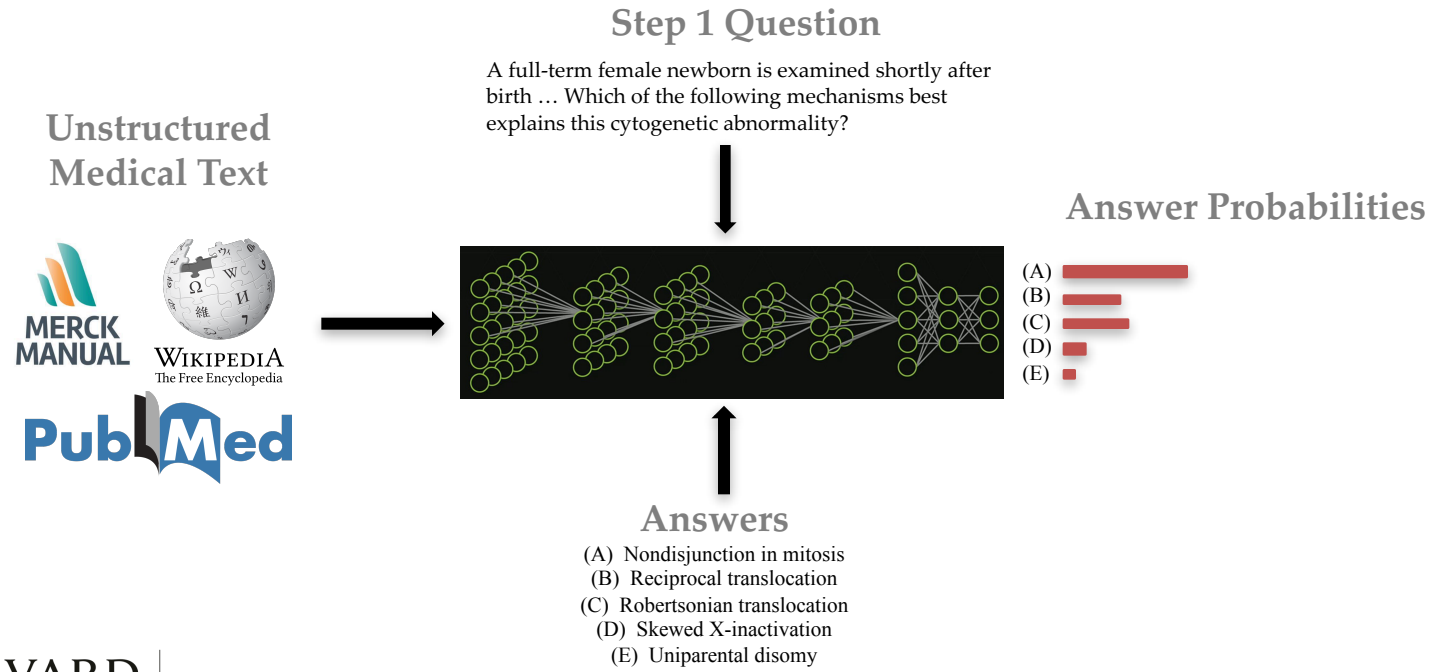
STEP 1 AND AI

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

OVERVIEW

Can we train a deep learning system capable of passing step 1?



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

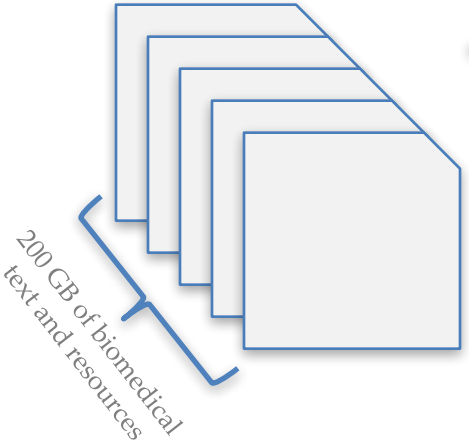
All preprocessed and normalized against a common medical thesaurus

Test Preparation
Flash cards
High Yield Concept List
Books

Step 1 Questions
Open Osmosis
Library Resources
NBME

DATA PREPROCESSING

Raw Text

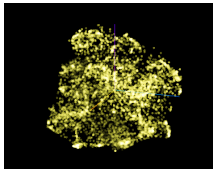


Normalization

Unified Medical Language System ®

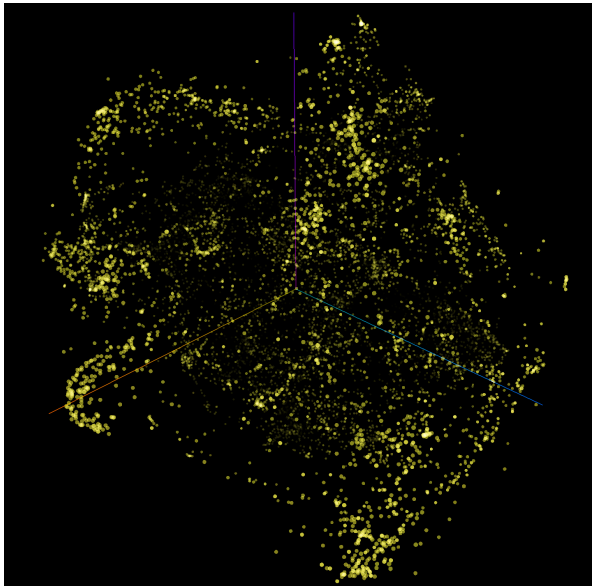
SNOMED CT
The global language of healthcare

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

CUI	SemanticType	String	X1	X2
C3872829	Biologically Active Substance	Adhesion protein	0.0056652557	-0.196249962
C3872700	Health Care Related Organization	Clinical pathology service	-0.4462876022	-0.013221873
C3872595	Nucleic Acid, Nucleoside, or Nucleotide	Human papillomavirus DNA	-0.0793692693	-0.198981807
C3872595	Biologically Active Substance	Human papillomavirus DNA	-0.0793692693	-0.198981807
C3872494	Manufactured Object	Device tip (physical object)	-0.1605362296	-0.325420946
C3872476	Medical Device	Body reference point marker	-0.0516590886	0.078756697
C3871203	Temporal Concept	At discharge	0.1280273497	0.100455143
C3864436	Medical Device	Anatomical structure separator	-0.0629289672	0.406454623
C3856907	Manufactured Object	Projector	0.1593338698	0.359946638

60,000 medical concepts

WHAT DRUGS ARE USED FOR BPD?

Query

bronchopulmonary dysplasia

Filter

Pharmacologic Substance

Rank

String	Similarity
Pulmonary Surfactants	0.3964883
palivizumab	0.3360302

HOW IS BPD MANAGED?

Query

bronchopulmonary dysplasia

→

Filter

Therapeutic or Preventive Procedure

Rank

String	Similarity
Oxygen Therapy Care	0.5373769
Mechanical ventilation	0.5018497
Intermittent Positive-Pressure Ventilation	0.4676242
High frequency oscillatory ventilation	0.4653350
Noninvasive Ventilation	0.4361196

DEEP LEARNING FOR QA

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

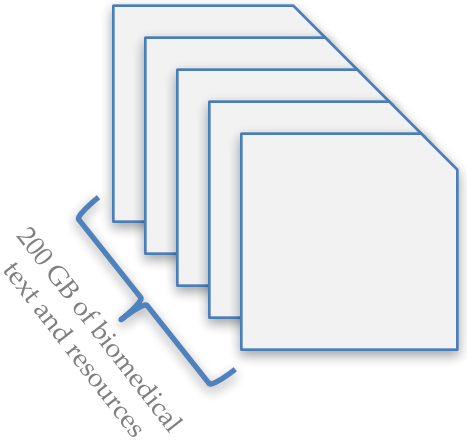
End-to-end deep learning QA systems need 100k – 1M QA pairs.

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

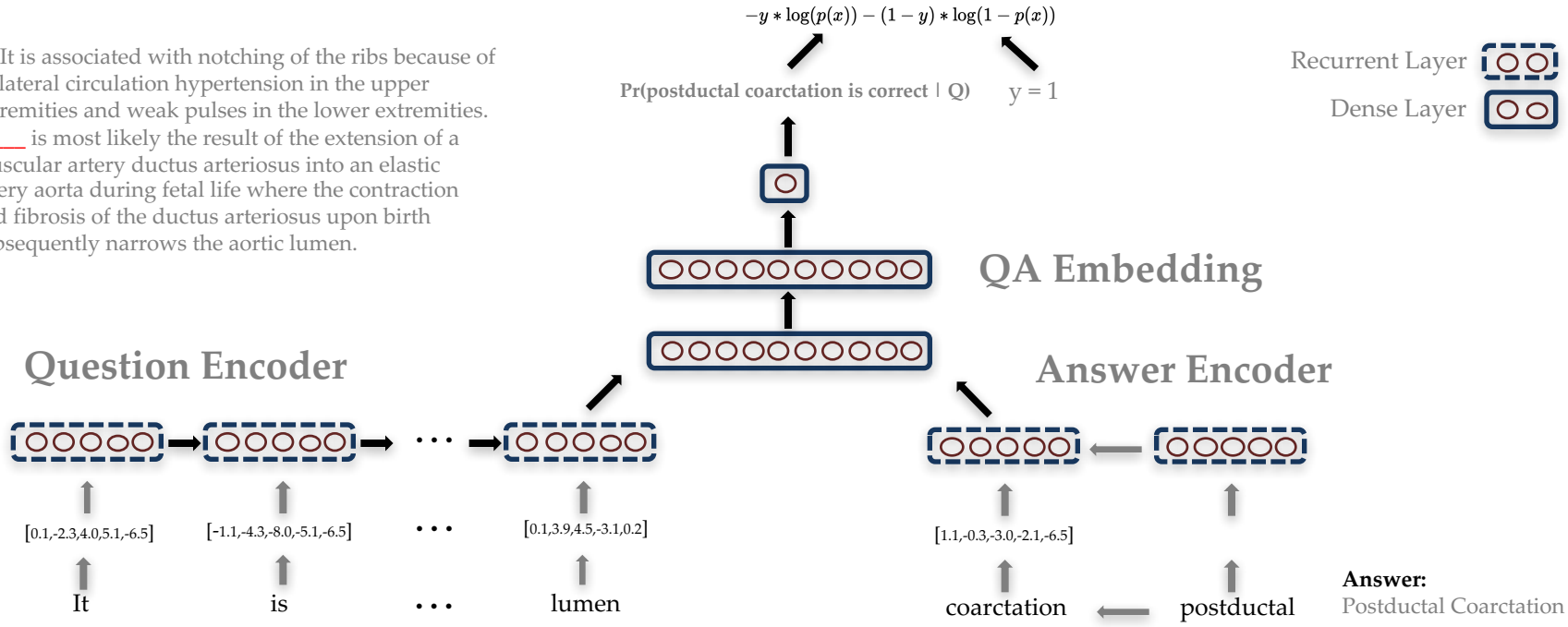
Compare semantic similarity of synthetic QA pairs against real ones.

Only keep high scoring synthetic QA pairs.

Results: 1 billion potential QA pairs

MODEL OVERVIEW

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

HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics

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

Osmosis
Open Osmosis

MAYO CLINIC

NBME
National Board of Medical Examiners

MERCK MANUAL