UMassAmherst

UMASS MEDICAL SCHOOL

Collaborative Tagging of Phenotypic Data for Clinical and Translational Sciences

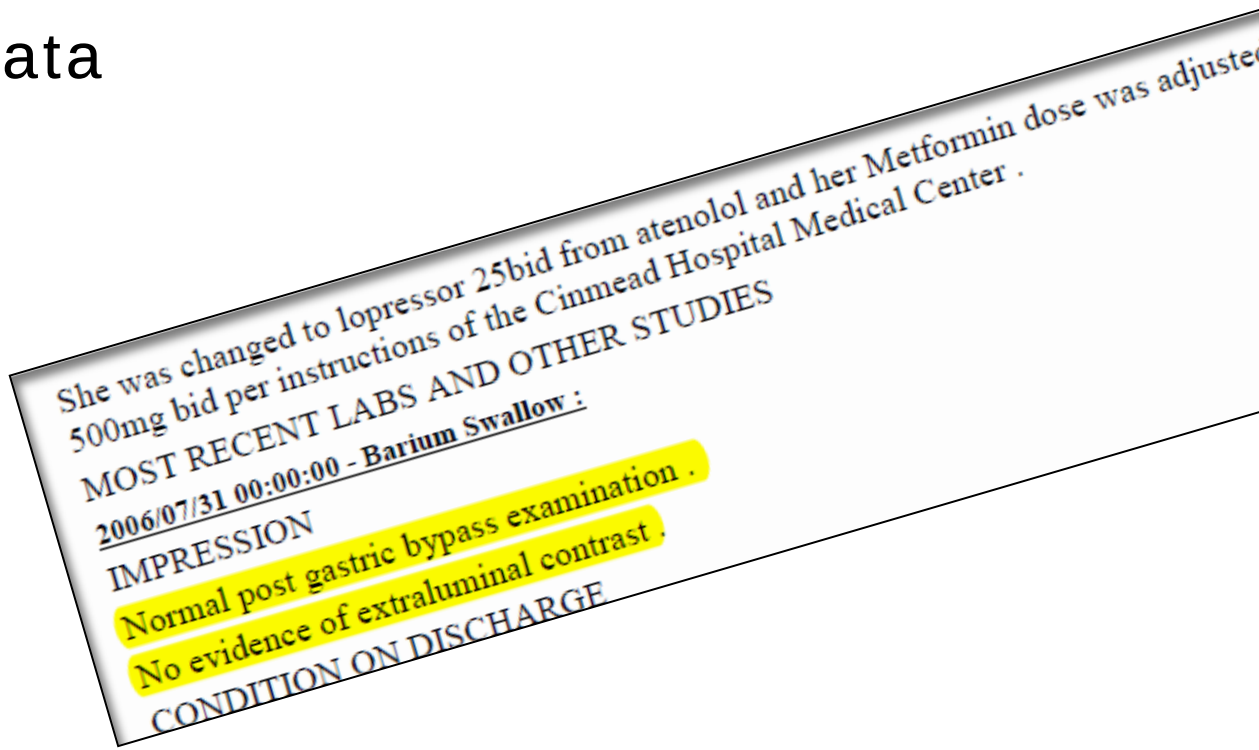
James Allan (presenting), W. Bruce Croft, Tom Houston, Rajani Sadasivam, Ariana Kamberi

Weize Kong, and Youngho Kim

May 8, 2013

Clinical Information

- Structured Data (labs, meds, ICD-9)
- Unstructured Data
  - Trapped
  - Not Easily
    - computer-interpretable
    - Organized
    - Retrievable


She was changed to lopressor 25bid from atenolol and her Metformin dose was adjuste
500mg bid per instructions of the Cinmead Hospital Medical Center .
MOST RECENT LABS AND OTHER STUDIES
2006/07/31 00:00:00 - Barium Swallow :
IMPRESSION
Normal post gastric bypass examination .
No evidence of extraluminal contrast .
CONDITION ON DISCHARGE

Outside Medicine

- Collaborative Tagging
  - Labels users create to represent topics in documents
  - Other users (and information retrieval systems) use these tags to explore information
  - Often unstructured, open-ended and interpretive

Wikipedia

Collaborative tagging of clinical notes

- Motivation
  - Structured clinical data using standard taxonomies are accurate but limited, relatively static, and represent a single view
  - Unstructured text is a rich source of information but NLP techniques are fragile, training and review is expensive
- Middle ground: tagging
  - Groups of individuals add or mark phrases (“tags”)
  - Resulting *folksonomy* may be simpler to use and can evolve quickly
- This work explores that middle ground

Clinician tagging of clinical notes

- Used existing, de-identified i2b2 collection

- Recruited clinicians to highlight and tag notes
- Approximate breakdown of resulting group
  - 50% family medicine doctor
  - 42% internal medicine doctor
  - 3% each nurse practitioner, physician assistant, senior resident

Their instructions

- Highlighting
  - Please use a highlighter to precisely select (highlight) tags, as many as you want, as few as you think you would need to best represent the **most important** aspects of the note you would like to share with others (could be two tags, could be ten tags, could be more).
- Tag generation
  - These tags are things you as a clinician might infer from the note, but are not explicitly stated (e.g., “missed diagnosis”, “good preventive care”, “depression prolonged hospital stay”).

Sample highlighted note

HISTORY OF PRESENT ILLNESS : Mr. Little is a 53 year old male who is under the care of Dr. Royendchaelmares , at Hend Geadcoastcar Hospital , with the diagnosis of coronary artery disease . He has a history of an old inferior myocardial infarction . He was well until three days prior to admission , when he developed an episode of shoulder and arm pain , with minimal exertion . The night prior to admission , he slept well , but the following day , he had a prolonged episode of chest pain .

He went to the Emergency Ward of Hend Geadcoastcar Hospital , where was found , on electrocardiogram , to have a right bundle branch block , and ST-segment elevations in the inferior and apical leads . He was treated with intravenous Streptokinase , intravenous heparin , intravenous nitroglycerin . He had a brief episode of bradycardia and hypotension , which responded to atropine and dopamine . He had some ventricular ectopy that responded to Xylocaine . He did well , without recurrent chest pain , congestive heart failure , or further arrhythmias . He ruled in for myocardial infarction , with a peak CPK of 660 units , 16% mB . An echocardiogram revealed an ejection fraction of 52% . He had cardiomegaly . He underwent an exercise tolerance test with Thallium , where he exercised for 2 minutes . The test was positive . Coronary angiography was performed on Sep 8 , which demonstrated a mean pulmonary capillary wedge pressure of 7 millimeters of mercury .

There was a 30% stenosis of the main left coronary artery . There was a 50% stenosis of the left anterior descending . The circumflex artery had a total occlusion . His right coronary artery had a severe 95% stenosis . The left ventricle has normal size , and an ejection fraction of 65% .

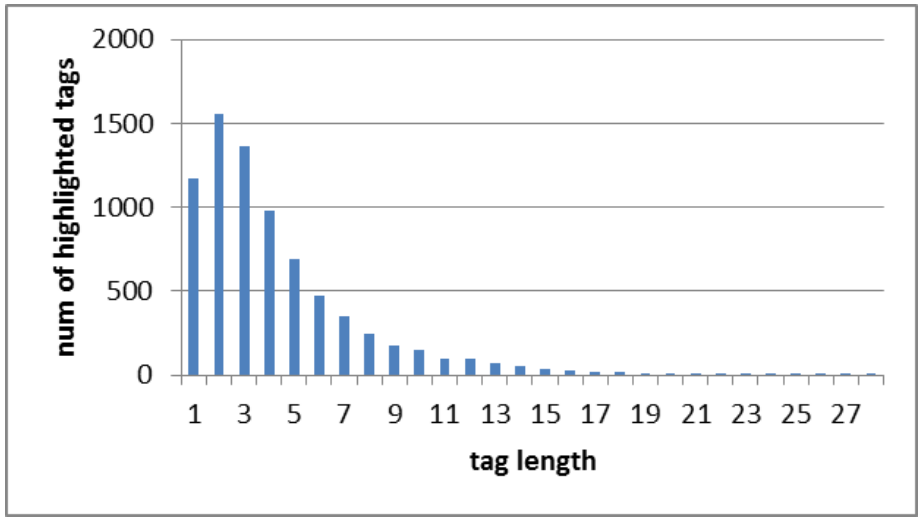
His PAST MEDICAL HISTORY is remarkable for an old Q-wave myocardial infarction .

He has a hiatus hernia .

He had prior surgery for hernia .

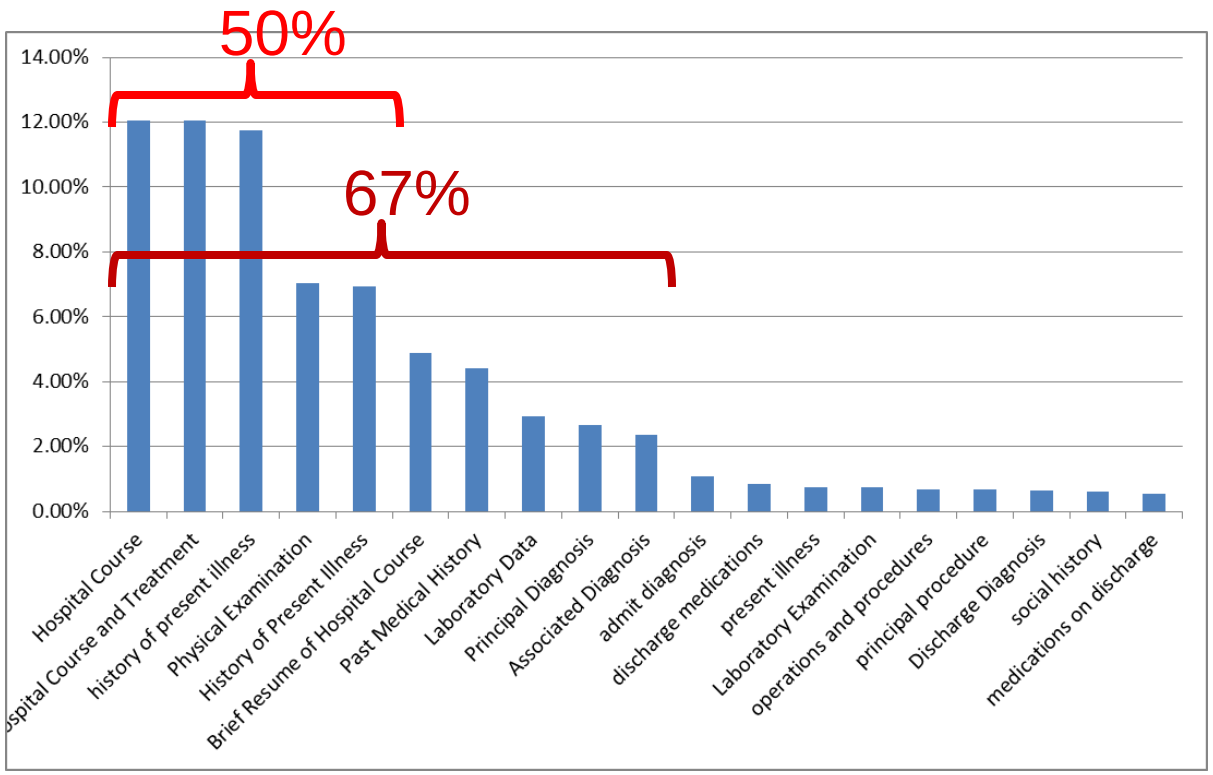
Summary of collected data

- 366 notes highlighted
  - Average of 2.1 annotators per note (766 notes)
  - Average of 5.8 notes per annotator (132 annotators)
  - From 16 to 496,506 words long
- 7,642 highlighted tags
  - Average of 20.9 highlighted tags per note
  - Average length is 4.36 words



Where do highlighted tags occur?

- Heavily skewed to a few sections of note



25% of tags in tail that is not shown

Use highlights to train *automatic* highlighting?

- cTakes, named entity tagger (small sample)
  - Identifies average of 232 tags
  - Here, 24 tags per notes, 17 of which overlapped
  - 70% recall, 7% precision

- SVM classifier
  - 21 features: length of tag, frequency, frequency in medical domain, which section, where in section, etc.
  - Pilot evaluation (27 train, 26 test)
    - 37% of top 5 words match
    - 16% of top 50 words match

Retrieving related medical records

- Query: medical record
- Goal: past medical records including related diseases, conditions, or treatments/interventions

- Pilot evaluation
  - 9 medical record “queries”
  - Average of 5.7 related past medical records

	Prec @ 20	Recall@20	MAP
Original	12%	43%	32%
Tags	13%	47%	37%
Expansion	17%	61%	44%

Example Applications

- Similarity
  - Identify a record with a medical error THEN
    - Find OTHER records with similar errors
  - Identify a specific hospital course THEN
    - Find OTHER records with similar

- Prioritize
  - Information within a note
  - OR within a patient record