2018 n2c2 shared task

Track 1: Cohort selection for clinical trials: overview

Amber Stubbs, Michele Filannino, Ozlem Uzuner

Assistant Professor, Simmons University

November 2, 2018

Corpus

- Records for 288 patients
  - 2-5 records per patient
  - Based on the corpus from the 2014 i2b2 shared task
    - Patients have diabetes, most are at risk for heart disease
- Annotated to indicate whether each patient meets or does not meet a set of selection criteria

Selection criteria

- Drug abuse, current or past
- Current alcohol use over weekly recommended limits
- Patient must speak English
- Patient must make their own medical decisions
- History of intra abdominal surgery, small or large intestine resection or small bowel obstruction
- Major diabetes-related complication.
- Advanced cardiovascular disease
- Myocardial infarction in the past 6 months
- Diagnosis of ketoacidosis in the past year
- Taken a dietary supplement (excluding Vitamin D) in the past 2 months
- Use of aspirin to prevent myocardial infarction
- Any HbA1c value between 6.5 and 9.5%
- Serum creatinine > upper limit of normal

Gold Standard creation

- Annotation
  - 2 annotators (with medical expertise) per record
  - Every patient assigned “meets”, “does not meet” or “possibly meets” status for each criteria, with accompanying text highlighted
- Adjudication
  - Resolve “possible” into “meets” or “not meets”
  - Resolve conflicts between annotators, in consultation with MD (Thank you, Ergin!)

Annotator agreement metrics

- Created redux files of annotator’s work
  - just tag + their annotation
- Simple % agreement calculation
  - Used one annotator as the “gold standard”
  - Check if other annotation matched or not

Inter-annotator agreement

Overall agreement:

84.9%

Criterion	% agree
Abdominal	85.8
Advanced CAD	60.1
Alcohol abuse	95.1
Aspirin for MI	84.4
Creatinine	82.6
Diet supplement w/in 2 months	81.6
Drug abuse	95.1
Speaks English	87.8
HBA1C	83.7
Ketoacidosis within 1 year	93.4
Major diabetes	77.7
Makes own decisions	86.1
MI in 6 months	89.6

Annotator disagreements

- Criteria that require multiple pieces of evidence had lower agreement
  - Major diabetes (77.7%)
  - Advanced CAD (60.1%)
- Diet supplement w/in 2 months (81.6%)
  - Some confusion over what constituted a supplement
- Measurements
  - Creatinine - 82.6%
  - HBA1C - 83.7%

Gold standard statistics

These are the original numbers, not the numbers from the corrections found during the shared task

Criterion	Met	Not met
Abdominal	107	181
Advanced CAD	170	118
Alcohol abuse	10	278
Aspirin for MI	230	58
Creatinine	106	182
Diet supplement w/in 2 months	149	139
Drug abuse	15	273
Speaks English	265	23
HBA1C	102	186
Ketoacidosis within 1 year	1	287
Major diabetes	156	132
Makes own decisions	277	11
MI in 6 months	26	262

Track 1

Goals:

- Answer the question “Can NLP systems use narrative medical records to identify which patients meet selection criteria for clinical trials?”

Setup:

- Training: 202 patient records with document-level annotations, 10 records with textual spans indicating annotator’s evidence for their annotations
  - 2 months to build/test systems
- Testing: 86 patient records
  - 3 days to run tests
  - Submit up to 3 runs

Participants

- 45 teams participated
  - Over 200 participants
  - 18 countries represented
- 109 system outputs submitted

Evaluation metrics

- Evaluation using standard micro-averaged precision, recall, and F1
- Micro F1 is the primary evaluation metric

Aggregate metrics - micro f1, all runs

	Track 1
Minimum	0.2117
Maximum	0.91
Average	0.799
Standard deviation	0.116
Median	0.8227

The slide you’ve been waiting for...

::drumroll::

Top 10 teams (best run only)

Rank	Team	Micro F1
1	MedUniGraz	0.91
2	University of Michigan	0.9075
3	Sorbonne Université	0.9069
4	Med Data Quest	0.9028
5	Cincinnati Children’s Hospital Medical Center	0.9026
6	Arizona State University	0.9003
7	University of New South Wales / National Cancer Institute	0.8913
8	Harbin Institute of Technology	0.8855
9	University of Utah	0.8837
10	NTTMUNSW	0.8765

Top 10 teams (best run only)

* = Presentation # = Poster

Rank	Team	Micro F1
1	* MedUniGraz	0.91
2	* University of Michigan	0.9075
3	* Sorbonne Université	0.9069
4	Med Data Quest	0.9028
5	Cincinnati Children’s Hospital Medical Center	0.9026
6	Arizona State University	0.9003
7	University of New South Wales / National Cancer Institute	0.8913
8	* Harbin Institute of Technology	0.8855
9	* University of Utah	0.8837
10	* NTTMUNSW	0.8765

Thank you!

Organizing committee:

- Ozlem Uzuner, co-chair, George Mason University
- Amber Stubbs, co-chair, Simmons University
- Michele Filannino, co-chair, MIT
- Kevin Buchan, SUNY at Albany
- Susanne Churchill, Harvard Medical School
- Isaac Kohane, Harvard Medical School
- Hua Xu, UTHealth
- Ergin Soysal, UTHealth

National Library of Medicine, National Institutes of Health, NIH NLM 5R13LM011411, PI: Ozlem Uzuner

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

GEORGE MASON UNIVERSITY | Volgenau School of Engineering