

Research and Applications

Cohort selection for clinical trials: n2c2 2018 shared task track 1

Amber Stubbs,¹ Michele Filannino,^{2,3} Ergin Soysal,⁴ Samuel Henry ,² and Özlem Uzuner,^{2,3,5}

¹Department of Mathematics and Computer Science, Simmons University, Boston, Massachusetts, USA, ²Information Sciences and Technology, George Mason University, Fairfax, Virginia, USA, ³Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, ⁴School of Biomedical Informatics, University of Texas Health Science Center, Houston, Texas, USA, and ⁵Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

Corresponding Author: Amber Stubbs, PhD, Department of Mathematics and Computer Science, Simmons University, 300 The Fenway, Boston, MA 02115, USA; stubbs@simmons.edu

Received 18 July 2019; Revised 7 August 2019; Editorial Decision 11 August 2019; Accepted 18 September 2019

ABSTRACT

Objective: Track 1 of the 2018 National NLP Clinical Challenges shared tasks focused on identifying which patients in a corpus of longitudinal medical records meet and do not meet identified selection criteria.

Materials and Methods: To address this challenge, we annotated American English clinical narratives for 288 patients according to whether they met these criteria. We chose criteria from existing clinical trials that represented a variety of natural language processing tasks, including concept extraction, temporal reasoning, and inference.

Results: A total of 47 teams participated in this shared task, with 224 participants in total. The participants represented 18 countries, and the teams submitted 109 total system outputs. The best-performing system achieved a micro F1 score of 0.91 using a rule-based approach. The top 10 teams used rule-based and hybrid systems to approach the problems.

Discussion: Clinical narratives are open to interpretation, particularly in cases where the selection criterion may be underspecified. This leaves room for annotators to use domain knowledge and intuition in selecting patients, which may lead to error in system outputs. However, teams who consulted medical professionals while building their systems were more likely to have high recall for patients, which is preferable for patient selection systems.

Conclusions: There is not yet a 1-size-fits-all solution for natural language processing systems approaching this task. Future research in this area can look to examining criteria requiring even more complex inferences, temporal reasoning, and domain knowledge.

Key words: natural language processing, clinical narratives, machine learning, cohort selection, information extraction

INTRODUCTION

Track 1 of the 2018 National NLP Clinical Challenges (n2c2) aimed to answer the question “Can NLP systems use narrative medical records to identify which patients meet selection criteria for clinical

trials?” Finding eligible patients is a time consuming process, and often requires access to information that is not stored in structured data or cannot easily be turned into a database query. For example, questions about a patient’s living situation, ability to consent to

medical procedures, intentions for the future (eg, “patient is actively trying to become pregnant”), or underspecified requirements in the criteria (eg, using “severe disease” without specifically defining “severe”) all make cohort selection a difficult task; therefore, researchers must often examine clinical narratives by hand to identify potential research participants.

If a researcher does not have time to scour through clinical narratives, they may end up recruiting patients who seek out the clinical trials on their own, or who are encouraged to enroll in the study by their primary care physician. These practices can result in selection bias toward certain populations, such as those who are more likely to have a primary care physician, or people who have the knowledge and time to search for relevant clinical trials on their own. Having a biased population can in turn bias the results of the study.^{1,2} By developing natural language processing (NLP) to automatically assess whether a person is eligible to participate in a clinical trial, we can potentially reduce bias and the amount of time needed to find a suitable patient population.³

However, the complexity of the criteria and the clinical narratives themselves makes using NLP for these purposes a nontrivial task. For example, the criterion “Use of aspirin to prevent myocardial infarction” indicates that the patient must be not only taking a particular medication, but also taking it for a heart-related reason and not mild pain relief. To correctly select a patient that meets this criterion, an NLP system must be able to identify both the key medication and the reason for taking the medication—information that may be stated elsewhere in the narrative.

This shared task aimed to determine whether NLP systems could be trained to identify if patients met or did not meet a set of selection criteria taken from real clinical trials. The selected criteria required measurement detection (“Any HbA1c value between 6.5 and 9.5%”), inference (“Use of aspirin to prevent myocardial infarction”), temporal reasoning (“Diagnosis of ketoacidosis in the past year”), and expert judgment to assess (“Major diabetes-related complication”). For the corpus, we used the dataset of American English, longitudinal clinical narratives from the 2014 i2b2/UTHealth shared task.⁴

The use of computers to access information in clinical narratives has been going on for decades,⁵ and electronic screening has long been used to expedite the selection process for clinical trials, particularly by using computers to scan structured patient data. This can be done very effectively, often with little to no loss of sensitivity.^{6–8}

However, there is a limit to the amount of information that is stored in structured databases, and so more recent studies have focused on using computers to help screen clinical narrative text. Schmickl et al⁹ used Boolean text searches to identify patients eligible for a trial on chronic obstructive pulmonary disease. Their system prioritized recall, then medical workers reviewed the selected files. Their system greatly reduced the amount of time needed to screen patients for eligibility. Using a more NLP-focused approach, Ni et al¹⁰ used both a filter to remove patients based on structured data and a machine learning system to identify relevant information from clinical narratives.¹¹ Their system showed high recall for identifying eligible patients and reduced the time needed to find suitable recruits. Other researchers have also had success in using NLP for screening clinical records and reducing time spent recruiting patients.^{12–14}

Most of the research conducted on the topic of using NLP to screen clinical narratives for eligibility criteria has been carried out internally in medical settings, using patient data that contains identifiable information. Because the Health Insurance Portability

Accountability Act restricts sharing identified patient records, other researchers are unable to replicate or improve these results. To open the problem of patient selection for clinical trials to the broader research community, we created Track 1 of the n2c2 2018 shared task.

To the best of our knowledge, the only other shared task to examine cohort selection for clinical trials was a medical track in the 2011 Text Retrieval Conference.¹⁵ There, participants evaluated records for relevance to 35 selected criteria on topics ranging from “Patients with hearing loss” to “Children admitted with cerebral palsy who received physical therapy.” This shared task made use of the entire 96 000 record Pittsburgh NLP Repository. Due to the size of the dataset, not all the records were judged for relevance for all criteria, so the judges used a bpref measure.¹⁶ The best system performance “was in the range 0.5-0.6, reasonably good performance for a general [information retrieval] task but not satisfactory for identifying patients in a set of clinical records.” Our shared task provides a specific gold standard for comparing system outputs per selection criteria, examines the progress in specific NLP tasks, and will add another data point for evaluating NLP systems on their ability to identify patient cohorts.

MATERIALS AND METHODS

The dataset for Track 1 of the 2018 n2c2 shared task used records of 288 patients from the corpus of the 2014 i2b2/UTHealth shared tasks,^{4,17} a collection of American English longitudinal records, with 2-5 records per patient. All the patients in this dataset have diabetes, and most are at risk for heart disease. The corpus contains 781 006 tokens, with an average of 2711 tokens per set of patient records. This provides a reasonably robust set of information about each patient.

This corpus was previously de-identified using a “risk averse” interpretation of the Health Insurance Portability Accountability Act guidelines.^{4,18} All information linked to a patient was removed and replaced with realistic surrogates, and dates were time-shifted a random amount for each patient.

Annotation guidelines

For each patient’s records, we annotated to indicate whether the patient meets or does not meet a set of selection criteria. We collected these criteria from real studies listed on ClinicalTrials.gov; however, we made slight modifications to make the criteria more relevant to patients represented in the corpus. All these patients are diabetic, and most exhibit risk factors for heart disease, so we selected criteria that were either related to these conditions, or that are common to most studies (eg, whether or not someone uses drugs, speaks English, or could make their own medical decisions). We selected 13 selection criteria:

- DRUG-ABUSE: Drug abuse, current or past
- ALCOHOL-ABUSE: Current alcohol use over weekly recommended limits
- ENGLISH: Patient must speak English
- MAKES-DECISIONS: Patient must make their own medical decisions
- ABDOMINAL: History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction
- MAJOR-DIABETES: Major diabetes-related complication. For the purposes of this annotation, we define “major complication” (as opposed to “minor complication”) as any of the following

that are a result of (or strongly correlated with) uncontrolled diabetes:

- Amputation
- Kidney damage
- Skin conditions
- Retinopathy
- nephropathy
- neuropathy
- ADVANCED-CAD: Advanced cardiovascular disease (CAD). For the purposes of this annotation, we define “advanced” as having 2 or more of the following:
 - Taking 2 or more medications to treat CAD
 - History of myocardial infarction (MI)
 - Currently experiencing angina
 - Ischemia, past or present
- MI-6MOS: MI in the past 6 months
- KETO-1YR: Diagnosis of ketoacidosis in the past year
- DIETSUPP-2MOS: Taken a dietary supplement (excluding vitamin D) in the past 2 months
- ASP-FOR-MI: Use of aspirin to prevent MI
- HBA1C: Any hemoglobin A1c (HbA1c) value between 6.5% and 9.5%
- CREATININE: Serum creatinine > upper limit of normal

Annotation procedure

Two annotators, both with medical expertise, independently annotated all the medical records. Each annotator examined every patient according to all thirteen criteria, and categorized each as met, not met, or possibly met. We asked that each annotator mark portions of the text that they used as evidence to determine the patient’s status, as a way to ensure consistency and close reading of the text. For criteria requiring temporal reasoning, we used the most recent record, regardless of date, as “now,” and used previous record dates to determine if the criterion was met.

A.S. adjudicated the disagreements in the annotations by examining the annotated evidence for each criterion, and consulting with an MD (E.S.) to create a gold standard. During the adjudication process, we discovered that all instances of “possibly met” could be resolved into “not met” or “met”; thus, the final gold standard contains only those labels.

Annotation quality

We calculated Cohen’s kappa¹⁹ over each criterion to determine interannotator agreement. Cohen’s Kappa measures agreement between annotators while taking chance agreement into consideration. The average kappa score across all criteria was 0.54. While this agreement metrics seems low, many of the criteria show distributions skewed toward either “met” or “not met,” which can affect these scores. Specifically, the scores for ALCOHOL-ABUSE (0.59), DRUG-ABUSE (0.63), ENGLISH (0.46), MAKES-DECISIONS (0.31), and KETO-1YR (-0.1) were greatly affected by the skewed distributions. In particular, for the KETO-1YR criterion, the annotators agreed that 275 of the 288 patients did not meet the criterion—such high agreement in a single category means that the Cohen’s kappa calculation assumes very high chance agreement, thereby causing a negative result.

The nonskewed criteria with the lowest agreement were ones that required multiple pieces of evidence or temporal reasoning to be considered “met,” specifically MAJOR-DIABETES (0.62 kappa), ADVANCED-CAD (0.37 kappa), MI-6MOS (0.63), and ASP-FOR-

Table 1. Distribution of “met” and “not met” labels for patients in the corpus, and their distribution over the training and testing data

Criterion	Met	Not met
ABDOMINAL	107 (77/30)	181 (125/56)
ADVANCED-CAD	170 (125/45)	118 (77/41)
ALCOHOL-ABUSE	10 (7/3)	278 (195/83)
ASP-FOR-MI	230 (162/68)	58 (40/18)
CREATININE	106 (82/24)	182 (120/62)
DIETSUPP-2MOS	149 (105/44)	139 (97/42)
DRUG-ABUSE	15 (12/3)	273 (190/83)
ENGLISH	265 (192/73)	23 (10/13)
HBA1C	102 (67/35)	186 (135/51)
KETO-1YR	1 (1/0) ^a	287 (201/86)
MAJOR-DIABETES	156 (113/43)	132 (89/43)
MAKES-DECISIONS	277 (194/83)	11 (8/3)
MI-6MOS	26 (18/8)	262 (184/78)

Values are total (train/test).

^aThe met group ketoacidosis was an annotation error. There were no instances of a patient meeting the ketoacidosis criterion in the corpus.

MI (0.62), as diagnosing these patients as “met” required closer reading and more inference. Measurements (CREATININE, 0.68; HBA1C, 0.72) were also somewhat difficult for the annotators to find, though these annotations were almost never disagreements, they were simply cases where one annotator found a qualifying measurement, and the other missed it. Finally, DIETSUPP-2MOS had a kappa score of 0.66, mostly due to some confusion over what constituted a supplement, which was cleared up over the course of the annotations. The agreement score for ABDOMINAL was 0.73.

Overall, the largest source of disparities in the annotations was that one annotator found evidence for a criterion that the other missed. This is likely due to the large amount of text for each patient, and how often relevant information was deep within paragraphs or spread across multiple records.

Table 1 shows the distribution of “met” and “not met” for each criterion in the data as it was distributed to the participants. The skewed distributions between “met” and “not met” are reflective of real-world patients—some conditions will be rarer in a population than others. For the training and test data split (70%/30%), we split the data to approximate the relative frequency of each criterion.

RESULTS

The training data for this shared task included 70% of the entire corpus—202 sets of patient records. These records all contained patient-level annotations of “met” or “not met” for each criterion. For 10 of the records, we shared with the community the evidence marked by the annotators.

Shared task participants had 2 months to build and test their systems. At that time, we released the remaining 86 unannotated patient records to the participants and allowed 3 days for systems to analyze the new data. Each team could submit up to 3 runs of results.

A total of 47 teams participated in this shared task, with 224 participants in total. The participants represented 18 countries, and the teams submitted 109 system outputs.

Evaluation metrics

We calculated the aggregate precision (P), recall (R), and F measure (F1) for all submissions. For each criterion, we calculated the correct

Table 2. Top 10 teams, best result from each team, ranked by micro F1 scores.

Rank	Team	Method and description	Medical professional	Micro F1
1	Medical University of Graz	Rule based; negation and family history detection; regular expressions ^{23,24}	Y	0.91
2	University of Michigan	Hybrid; used MetaMap, ²⁵ cTAKES, ²⁶ and RxNORM, ²⁷ separate processing for different types of criteria ^{28,29}	Y	0.9075
3	Sorbonne Université	Hybrid; external resources: Unified Medical Language System, ³⁰ cTAKES, ²⁶ Heideltime, ³¹ MIMIC II ³² dataset. used rule- and terminology-based approaches for some criteria, semi-supervised learning others ^{33,34}	Y	0.9069
4	Med Data Quest	Rule based; 3 parts: evidence extraction, assertion, logic ^{35,36}	N	0.9028
5	Cincinnati Children's Hospital Medical Center	Hybrid; used cTAKES, ²⁶ assertion detection, regular expressions for structured criteria and phrase-based detection for unstructured criteria ^{37,38}	N	0.9026
6	Arizona State University	Hybrid; rule-based systems for most criteria, semisupervised learning for criteria requiring more complicated rules. Used CLAMP, ³⁹ lists of medications and word2vec ⁴⁰⁻⁴²	N	0.9003
7	University of New South Wales/ National Cancer Institute	Rule based; 280 rules with 12 hand-crafted dictionaries ⁴³	N	0.8913
8	Harbin Institute of Technology	Hybrid; used rules, a CNN and CNN-highway-LSTM ^{44,45}	N	0.8855
9	University of Utah	Rule-based; Trie-based hash rule processor; components to identify sentence segments, context, temporality, etc. ^{46,47}	Y	0.8837
10	National Taitung, Taipei Medical, University of New South Wales	Hybrid; used rules and an SVM with a "multi-instance polynomial kernel" ⁴⁸	Y	0.8765

This table includes information about the types of systems each team built, and whether they consulted a medical professional in building their system. The systems are described more fully in [Supplementary Appendix A](#).

CNN: convolutional neural network; LSTM: long-short term memory; MIMIC: Multiparameter Intelligent Monitoring in Intensive Care; N: no; SVM: support vector machine; Y: yes.

P, R, and F1 for both "met" and "not met" answers, then averaged those to get the micro score for each criterion. Then we averaged all of those together to get the overall micro-averaged F1 score.

We used approximate randomization^{20,21} to test for statistical significance of P, R, and F1 between systems. We ran this test with 50 000 shuffles and the significance level $\alpha=0.1$, in keeping with results of previous shared tasks.²²

Top-performing systems

The mean micro F1 score for all submissions was 0.799. The maximum and minimum scores were 0.91 and 0.2117, respectively, with a median of 0.8227 and a standard deviation of 0.116.

[Table 2](#) shows the results of the best run from each of the top 10 teams, ranked by micro F1 score, as well as the type of system they built and involvement of medical professionals in system development. There were 4 rule-based systems and 6 hybrid rules or machine learning (ML) systems in the top 10. No top 10 system was purely ML-based. Five of the 10 teams consulted with medical professionals to build their systems.

Overall, many systems used external resources such as Unified Medical Language System,³⁰ cTAKES,²⁶ and word2vec⁴⁰ for processing, and many teams built custom dictionaries from a variety of sources to aid in concept extraction. The use of rules in all the top 10 systems suggests that for small datasets, and particularly for criteria such as lab results or simple presence/absence of a disease or medication, well-defined rules remain a more effective approach to NLP than supervised or semisupervised learning.

[Table 3](#) shows the results of significance testing between the top 10 systems. Because the upper diagonal would be symmetrically

identical to the lower, we show only the lower half of the table. Cells containing P, R, or F1 indicate that the 2 systems are significantly different in P, R, or F1, respectively. Overall, we see that the top 6 systems show no significant differences in output, and that lower-ranked systems are also similar to each other.

Error analysis

In the categories ABDOMINAL, CREATININE, DIETSUPP-2MOS, ENGLISH, HBA1C, and MAJOR-DIABETES, the majority of teams succeeded in scoring over 0.80 on the F1 measure, a reasonable level of performance for the tasks. The CREATININE and HBA1C criteria are lab measurements, so most teams used rules to make a determination about those criteria. Similarly, most teams addressed the DIETSUPP-2MOS criterion by checking medication lists in the narratives for diet supplements, then checking the dates on the records to see if it was listed in the timeframe. Since most medications and supplements are simply part of the list of medications, little context or negation assessment would need to be done to identify patients who "met" that criteria. In terms of information in easily-parsed lists, the ABDOMINAL and MAJOR-DIABETES criteria simply required that certain diseases or events have happened to the person at some point, and these would usually be found in the list of "major problems" (or similarly titled section header) found in many medical records. Finally, the ENGLISH criterion is one that, for most records, was presumed to be true unless a mention was made of the patient needing an interpreter or speaking another language, so most patients "met" that criterion.

[Table 4](#) shows the specific results for each criterion by team. For criteria in which most systems scored below 0.80, there are multiple

Table 3. Results of significance testing between system outputs.

	MedUniGraz	UMich	Sorbonne	MedDataQuest	CCHMC	ASU	UNSW/NCI	HIT	Utah	NTTMUNSW
MedUniGraz	–									
UMich		–								
Sorbonne			–							
MedDataQuest				–						
CCHMC					–					
ASU						–				
UNSW/NCI	P, R	P	P	P						
HIT	F, P, R	F, P, R	F, P, R	F, P, R	F, R					
Utah	F, P, R	F, P, R	F, P, R	F, P, R	F, R	F			–	
NTTMUNSW	F, P, R	F, P, R	F, P, R	F, P, R	F, P, R	F, P, R	F			–

Cells with P, R, or F indicate the differences for these systems are statistically significant for precision, recall, and F1, respectively.

ASU: Arizona State University; CCHMC: Cincinnati Children's Hospital Medical Center; F: F1 score; HIT: Harbin Institute of Technology; MedUniGraz: Medical University of Graz; NTTMUNSW: National Taitung, Taipei Medical, University of New South Wales; P: precision; R: recall; Sorbonne: Sorbonne Université; UNSW/NCI: University of New South Wales/National Cancer Institute; UMich: University of Michigan; Utah: University of Utah

factors at play. First, these criteria required more complicated reasoning, as they included temporal modifiers (MI-6MOS, ADVANCED CAD, ALCOHOL-ABUSE) or inference (MAKES-DECISIONS, DRUG-ABUSE). Second, many of these categories had skewed distributions in the gold standard. Owing to the way we calculated micro F1 scores, low scores in these categories do not necessarily indicate a large number of incorrect answers on the criteria; rather, they may indicate wrong answers on a few key patients.

Table 5 shows the number of patients, that were incorrectly labeled by 5 or more systems in the top 10. Here we examine the commonalities between mislabeled patients, and identify the factors contributing to low scores. Patient screening systems are often calibrated to ensure maximum recall, and so errors of omission are more of a concern for cohort selection. Therefore we examine instances in which the gold standard is “met,” but the majority predicted label is “not met.” We group the criteria by the types of NLP they required: concept extraction, temporal reasoning, and inference.

Concept extraction

For the ABDOMINAL criterion, of the 8 records incorrectly labeled by 5 or more teams, 6 were incorrectly labeled as “not met.” The majority of systems incorrectly labeled patients with medical terms that were seen rarely, if at all, in the training data. For example, cesarean sections and bladder suspension surgeries are relatively rare in the dataset, and in the training data, cesarean sections were often present in patients who had other abdominal surgeries. The teams from the University of Utah and Sorbonne Université both labeled at least 3 of these 6 correctly.

Of the 9 patients mislabeled on the MAJOR DIABETES criterion, 6 should have been labeled “met.” In this case, the errors were sometimes caused by misspellings in the data (eg, “Fourier's gangrene” instead of “Fournier's gangrene”), but in many cases were due to annotators applying more of their own medical knowledge and inference to the data. For example, foot sores combined with loss of vision in a diabetic patient are likely indicators of severe diabetes.

Looking at the CREATININE and HBA1C criteria, we find that recall errors here (6 for CREATININE, 6 for HBA1C) were mostly caused by systems being unable to parse data phrased in an unusual manner, for example “Last HBGA1c 3/97-8.30” and “BUN and Cr are in the 25–40/1.3–1.7 range.” Systems that correctly labeled at least 3 of the HBA1C included National Taitung, Taipei Medical,

University of New South Wales (NTTMUNSW); Sorbonne Université; University of New South Wales/National Cancer Institute (UNSW/NCI); Arizona State University; Med Data Quest; and Harbin Institute of Technology. No team correctly labeled 3 or more of the 6 CREATININE errors.

Temporal reasoning

For ADVANCED-CAD, only 5 of the 18 incorrectly labeled patients were recall errors. The patient narratives that were incorrectly labeled by the systems were not cases of error in temporal reasoning. Instead, as with MAJOR-DIABETES, the annotators exercised a certain amount of professional knowledge. They therefore included patients with “apparently an MI” and “MIBI test showed very small amount of ischemia almost undetectable in amount”—phrases that may have been flagged as negative by a system checking for assertion or negation. University of Utah, University of Michigan, Sorbonne Université, NTTMUNSW, Med Data Quest, and Cincinnati Children's Hospital Medical Center each correctly labeled at least 3 of these 5.

For the criteria DIETSUPP-2MOS, the main sources of recall error (3 of the 8 errors) were supplements mentioned in a previous record, but not superseded in the most recent. In 1 case, the patient was taking the potassium supplement “Klor con,” which only 2 teams (UNSW/NCI, NTTMUNSW) correctly identified.

Of the 4 patients mislabeled on the MI-6MOS criterion, all 4 were recall errors. In 2 of these cases, the error does seem related to the need for temporal parsing and relations. For example, one record contains the description “Patient admitted to the hospital in middle of November after she presented with pulmonary edema. Ruled in for an MI [...]” Since the record this phrase comes from is the most recent one in the patient's file, and is dated for December, the patient does meet the criteria. Med Data Quest performed very well on this data, catching 3 of the 4 most frequently mislabeled patients.

Inference

For the criteria ASP-FOR-MI, ENGLISH, and MAKES-DECISIONS, all patients with errors made by 5 or more teams were cases where the gold standard labeled them “not met,” but the systems labeled them “met,” making these precision errors rather than recall. We surmise that this is due to the skewed distribution in the corpus of these criteria (ie, the vast majority of patients “met” all 3 criteria).

Table 4. Team F1 scores by criterion.

Team	Overall (micro)	ABDOMINAL	ADVANCED-CAD	ALCOHOL-ABUSE	ASP-FOR-MI	CREATININE	DIETSUPP-2MOS	DRUG-ABUSE	ENGLISH	HBA1C	KETO-1YR	MAJOR-DIABETES	MAKES-DECISIONS	MO-6MOS
MedUniGraz	0.91 ^a	0.87	0.790	0.4881	0.7095	0.8071	0.9185	0.691	0.8644	0.9382	0.5	0.8369	0.4911	0.8752
UMich	0.908	0.906 ^a	0.785	0.897 ^a	0.674	0.856	0.895	0.777	0.830	0.912	0.5	0.860	0.590	0.689
Sorbonne	0.907	0.912	0.763	0.744	0.606	0.898 ^a	0.895	0.827	0.844	0.899	0.5	0.884 ^a	0.485	0.767
MedDataQuest	0.903	0.906 ^a	0.708	0.488	0.710	0.898 ^a	0.860	0.660	0.977 ^a	0.938	0.5	0.825	0.476	0.876 ^a
GCHMC	0.903	0.847	0.742	0.827	0.616	0.868	0.919 ^a	0.691	0.830	0.925	0.491	0.883	0.897 ^a	0.792
ASU	0.900	0.894	0.802	0.488	0.648	0.847	0.872	0.827	0.925	0.913	0.5	0.835	0.827	0.710
UNSW/NCI	0.891	0.894	0.720	0.488	0.77 ^a	0.865	0.800	0.92 ^a	0.793	0.913	0.5	0.790	0.777	0.476
HIT	0.886	0.755	0.823	0.491	0.674	0.826	0.778	0.744	0.774	0.913	0.497	0.826	0.744	0.689
Utah	0.884	0.884	0.87 ^a	0.482	0.649	0.803	0.847	0.660	0.774	0.874	0.5	0.732	0.744	0.752
NTTMUNSW	0.877	0.780	0.775	0.655	0.743	0.726	0.785	0.744	0.752	0.95 ^a	0.494	0.767	0.655	0.767

^aHighest score for each category.

ASU: Arizona State University; CCHMC: Cincinnati Children's Hospital Medical Center; HIT: Harbin Institute of Technology; MedUniGraz: Medical University of Graz; NTTMUNSW: National Taitung, Taipei Medical, University of New South Wales; Sorbonne: Sorbonne Université; UNSW/NCI: University of New South Wales/National Cancer Institute; UMich: University of Michigan; Utah: University of Utah

Table 5. Error analysis for the top 10 teams over the gold standard test data, with the number of patients labeled incorrectly by 5 or more teams

Criteria	Incorrect by 5 or more teams (% of gold standard)
ABDOMINAL	8 (9.3)
ADVANCED-CAD	18 (20.93)
ALCOHOL-ABUSE	3 (3.49)
ASP-FOR-MI	12 (13.95)
CREATININE	10 (11.63)
DIETSUPP-2MOS	8 (9.3)
DRUG-ABUSE	3 (3.49)
ENGLISH	6 (6.98)
HBA1C	6 (6.98)
MAJOR-DIABETES	9 (10.47)
MAKES-DECISIONS	2 (2.33)
MI-6MOS	4 (4.65)

Finally, the ALCOHOL-ABUSE and DRUG-ABUSE criteria each only had 3 patients that 5 or more systems mis-labeled, and 5 of those incorrect labels were recall errors. The 2 DRUG-ABUSE recall errors were both cases where the patient admitted to past drug abuse—since the criterion stated the drug abuse could be past or present, these omissions on the part of the systems were likely because the information was negated, or not in a standard place in the document (eg, “H/O drug abuse” listed under “Problems” rather than “social history”). For ALCOHOL-ABUSE, the errors were much more subtle, and the texts were more open to interpretation. For example, “Drinks 1-2 black Russians per day, previously slightly more”—depending on which agency’s definition of “recommended limits” you use, this may or may not be alcohol abuse. UNSW/NCI did well on these patients.

DISCUSSION

By examining the information about the patients who should have been included in potential cohorts, but were excluded by most systems, we find some patterns. First, in many cases there is some room for interpretation of the data, and there is evidence that annotators used their own domain knowledge and intuition to label the patients. It is, of course, difficult to teach computers this type of reasoning, and for humans to always apply intuition consistently.

Second, looking at the teams who tended to correctly label patients that most other teams excluded, we see some names occur more frequently than others: University of Utah, Sorbonne Université, NTTMUNSW, and Med Data Quest. Each team used a unique combination of existing resources, rules, and (except for the University of Utah) machine learning. Three of the 4 teams all consulted medical experts as part of their system building, which may have contributed to their success.

Overall, the systems did very well on all the criteria, with no single criteria posing a problem for the majority of teams. We are excited to see that NLP can be applied so successfully to such a variety of NLP tasks. Looking toward the future of NLP, we suggest that future cohort selection tasks could incorporate more domain knowledge and even more complex temporal requirements.

CONCLUSION

The corpus for the 2018 n2c2 Track 1 provides a new dataset for researchers interested in cohort selection for clinical trials. The

corpus consists of longitudinal clinical narratives for 288 patients, with 2-5 records per patient. Each patient in the database has diabetes, and most are at risk for heart disease. As such, we selected 13 inclusion and exclusion criteria from existing clinical trials mostly related to diabetes and heart disease, and annotated the patient records to indicate whether each patient met or did not meet the criteria.

Track 1 of the n2c2 2018 shared task aimed to answer the question “Can NLP systems use narrative medical records to identify which patients meet selection criteria for clinical trials?” Based on the results of this shared task, we suggest that they can, though some criteria are easier to model than others.

A total of 47 teams participated in this shared task, with 224 participants in total. Of the top 10 team systems, all were able to achieve an overall micro F1 measure over 0.87, indicating high performance overall. While a rule-based system scored the highest overall (F1 score = 0.91), the scores of the top 6 teams were not significantly different. Most teams used combinations of rules and machine learning.

The error analysis shows us that, aside from the criterion requiring inference, there is no criterion that could not be assessed by at least 1 system with an F1 measure of 0.85 or higher. Systems that involved medical experts tended to have higher recall on patients that were difficult to label than systems that did not. Additionally, the error analysis shows that no one approach is inherently significantly better than others, but rather carefully building rules and crafting features for each criterion yielded the best successes.

Overall, the results of this shared task indicate that automated systems can be used to assist in cohort selection for clinical trials, but there is not yet a 1-size-fits-all solution for NLP systems approaching this task. Additionally, the error analysis shows that NLP systems are capable of tackling reasonably complex selection criteria, and future research in this area can look to examining criteria requiring even more complex inferences, temporal reasoning, and domain knowledge.

FUNDING

This work was supported by the National Library of Medicine of the National Institutes of Health under Award Numbers R13LM013127 (ÖU) and R13LM011411 (ÖU). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

AS and OU were the primary authors of the article, led the annotation effort, and coordinated the shared task. ES contributed to the annotation guidelines and adjudication process. MF and SH contributed to the data analysis.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank members of the organizing committee: Kevin Buchan Jr, Susanne Churchill, Isaac Kohane, and Hua Xu. We would also like to thank our annotators, Maren Muhlestein, Katie Grace Ruri, and Kathleen Ririe, and the

developers who built and maintained the web portal, Nathaniel Bessa, Rachel Eastwood, and Tommie Michael MacDuffie.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

- Mann CJ. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emerg Med J* 2003; 20 (1): 54–60.
- Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case-control studies. *Biostatistics* 2008; 10 (1): 17–31.
- Stubbs AC. *A Methodology for Using Professional Knowledge in Corpus Annotation* [PhD dissertation]. Waltham, MA, Brandeis University; 2013.
- Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. *J Biomed Inform* 2015; 58: S20–S29.
- Hripscak G, Friedman C, Anderson PO, et al. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995; 122 (9): 681.
- Thadani SR, Weng C, Bigger JT, et al. Electronic screening improves efficiency in clinical trial recruitment. *J Am Med Inform Assoc* 2009; 16 (6): 869–73.
- Embi PJ, Jain A, Clark J, et al. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. *AMIA Annu Symp Proc* 2005; 2005: 231–5.
- Grundmeier RW, Swietlik M, Bell LM. Research subject enrollment by primary care pediatricians using an electronic health record. *AMIA Annu Symp Proc* 2007; 2007: 289–93.
- Schmickl CN, Li M, Li G, et al. The accuracy and efficiency of electronic screening for recruitment into a clinical trial on COPD. *Respir Med* 2011; 105 (10): 1501–6.
- Ni Y, Wright J, Perentesis J, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. *BMC Med Inform Decis Mak* 2015; 15 (1): 28.
- Ni Y, Kennebeck S, Dexheimer JW, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. *J Am Med Inform Assoc* 2015; 22 (1): 166–78.
- Yuan C, Ryan PB, Ta C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc* 2019; 26 (4): 294–305.
- Koola JD, Davis SE, Al-Nimri O, et al. Development of an automated phenotyping algorithm for hepatorenal syndrome. *J Biomed Inform* 2018; 80: 87–95.
- Feller DJ, Zucker J, Yin MT, et al. Using clinical notes and natural language processing for automated HIV risk assessment. *J Acquir Immune Defic Syndr* 2018; 77 (2): 160–6.
- Edinger T, Cohen AM, Bedrick S, et al. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC medical records track. *AMIA Annu Symp Proc* 2012; 2012: 180–8.
- Buckley C, Voorhees EM. Retrieval evaluation with incomplete information. In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*; Sheffield, England: ACM Press; 2004: 25–32.
- Stubbs A, Uzuner Ö. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *J Biomed Inform* 2015; 58: S78–91.
- Stubbs A, Uzuner Ö, Kotfila C, et al. Challenges in synthesizing surrogate PHI in narrative EMRs. In: Gkoulalas-Divanis A, Loukides G, eds. *Medical Data Privacy Handbook*. Berlin: Springer; 2015: 717–35.
- Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20 (1): 37–46.
- Noreen EW. *Computer-Intensive Methods for Testing Hypotheses*. New York, NY: Wiley; 1989.
- Yeh A. More accurate tests for the statistical significance of result differences. In: *COLING '00: Proceedings of the 18th Conference on Computational Linguistics-Volume 2*. Stroudsburg, PA: Association for Computational Linguistics; 2000: 947–53.
- Nancy C. 1992. The statistical significance of the MUC-4 results. In: *Proceedings of the 4th Conference on Message understanding (MUC4 '92)*. Stroudsburg, PA: Association for Computational Linguistics; 1992: 30–50.
- Oleynik M, Kugic A, Kreuzthaler M, et al. Med Uni Graz at n2c2 Track 1. In: *proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks*; January 28, 2018; San Francisco, CA.
- Oleynik M, Kugic A, Kreuzthaler M, et al. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared-task on clinical text classification. *J Am Med Inform Assoc* 2019; doi: 10.1093/jamia/ocz149.
- Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010; 17 (3): 229–36.
- Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
- Nelson SJ, Zeng K, Kilbourne J, et al. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* 2011; 18 (4): 441–8.
- Vydiswaran VGV, Agarwal M, Bagazinski E, et al. Hybrid bag of approaches to characterize selection criteria for cohort identification. In: *proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks*; January 28, 2018; San Francisco, CA.
- Vydiswaran VGV, Strayhorn A, Zhao X, et al. Hybrid bag of approaches to characterize selection criteria for cohort identification. *J Am Med Inform Assoc* 2019; doi: 10.1093/jamia/ocz079.
- Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–70.
- Strötgen J, Gertz M. Multilingual and cross-domain temporal tagging. *Lang Resour Eval* 2013; 47 (2): 269–98.
- Saeed M, Villarreal M, Reisner AT, et al. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Crit Care Med* 2011; 39 (5): 952–60.
- Bréant S, Cisneros H, Daniel C, et al. Participation to n2c2 challenge: a variety of approaches for cohort selection for clinical trials. In: *proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks*; January 28, 2018; San Francisco, CA.
- Tannier S, Paris N, Cisneros H, et al. Hybrid approaches for our participation to the n2c2 challenge on cohort selection for clinical trials. *arXiv* 2019 Mar 19.
- Chen L, Gu Y, Ji X, et al. Clinical trial cohort selection based on rule-based and machine learning hybrid NLP Systems. In: *proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks*; January 28, 2018; San Francisco, CA.
- Chen L, Gu Y, Ji X, et al. Clinical trial cohort selection based on multi-level rule-based natural language processing system. *J Am Med Inform Assoc* 2019; doi: 10.1093/jamia/ocz109.
- Yizhao N. Automated clinical trial eligibility screener. In: *Proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks*; January 28, 2018; San Francisco, CA.
- Ni Y, Bermudez M, Kennebeck S, Liddy-Hicks S, Dexheimer J. A Real-Time Automated Patient Screening System for Clinical Trials Eligibility in an Emergency Department: Design and Evaluation. *JMIR Med Inform* 2019; 7 (3): e14185.
- Soysal E, Wang J, Jiang M, et al. CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018; 25 (3): 331–6.
- Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, et al, eds. *Proceedings of the 26th International Conference on Neural Information Processing Systems, Vol. 2. (NIPS'13)*. Red Hook, NY: Curran Associates Inc; 2013: 3111–9.

41. Adhya S, Kulkarni S, Prakash A, *et al.* A hybrid approach to cohort selection for clinical trial: ASU at 2018 n2c2 challenge track 1. In: proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks; January 28, 2018; San Francisco, CA.
42. Rawal S, Prakash A, Adhya S, *et al.* Automated clinical lexicon induction and its use in cohort selection from clinical notes. *arXiv*: 1902.09674; 2019.
43. Karystianis G, Florez-Vargas O. Application of a rule-based approach to identify patient eligibility for clinical trials. In: proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks; January 28, 2018; San Francisco, CA.
44. Shi X, Xiong Y, Jiang D, *et al.* Cohort selection for clinical trials using CNN and CNN-highway-LSTM. In: proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks; January 28, 2018; San Francisco, CA.
45. Xiong Y, Shi X, Chen S, *et al.* Cohort selection for clinical trials using hierarchical neural network. *J Am Med Inform Assoc* 2019; doi: 10.1093/jamia/ocz099.
46. Shi J, Shao J, Graves K, *et al.* A generic rule-based pipeline for patient cohort identification. In: proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks; January 28, 2018; San Francisco, CA.
47. Shi J, Graves K, Hurdle JF. A generic rule-based system for clinical trial patient selection. *arXiv* 2019 Jul 16 [E-pub ahead of print].
48. Wang F-D, Chen C-W, Dai H-J, *et al.* NTMUNSW system for n2c2 track1: cohort selection for clinical trials. In: proceedings of the 2018 National NLP Clinical Challenges (n2c2) Workshop Shared Tasks; January 28, 2018; San Francisco, CA.