Published in final edited form as:
J Biomed Inform. 2015 December ; 58(Suppl): S67–S77. doi:10.1016/j.jbi.2015.07.001.

Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2

Amber Stubbs[1], **Christopher Kotfila**[2], **Hua Xu**[3], and **Ozlem Uzuner**[2]

[1]School of Library and Information Science, Simmons College, Boston, MA, USA

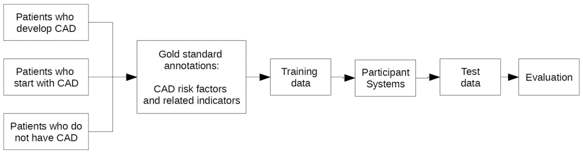
[2]Department of Information Studies, State University of New York at Albany, Albany, NY, USA

[3]Center for Computational Biomedicine, University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

The second track of the 2014 i2b2/UTHealth Natural Language Processing shared task focused on identifying medical risk factors related to Coronary Artery Disease (CAD) in the narratives of longitudinal medical records of diabetic patients. The risk factors included hypertension, hyperlipidemia, obesity, smoking status, and family history, as well as diabetes and CAD, and indicators that suggest the presence of those diseases. In addition to identifying the risk factors, this track of the 2014 i2b2/UTHealth shared task studied the presence and progression of the risk factors in longitudinal medical records. Twenty teams participated in this track, and submitted 49 system runs for evaluation. Six of the top 10 teams achieved F1 scores over 0.90, and all 10 scored over 0.87. The most successful system used a combination of additional annotations, external lexicons, hand-written rules and Support Vector Machines. The results of this track indicate that identification of risk factors and their progression over time is well within the reach of automated systems.

Graphical Abstract


Patients who develop CAD
Patients who start with CAD
Patients who do not have CAD
Gold standard annotations:
CAD risk factors and related indicators
Training data
Participant Systems
Test data
Evaluation

Corresponding author: Amber Stubbs, School of Library and Information Science, Simmons College, 300 The Fenway, Boston, MA 02115, USA, stubbs@simmons.edu, Phone: 617-521-2807.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest Statement
June 18, 2015
To whom it may concern,
This statement affirms that Amber Stubbs, Christopher Kotfila, Hua Xu, and Ozlem Uzuner, the authors of the paper “Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2” do not have any conflicts of interest relating to this paper or its results.

1. Introduction

In 2014, the Informatics for Integrating Biology and the Bedside (i2b2) project, in conjunction with University of Texas Health Science Center at Houston (UTHealth), sponsored a shared task in natural language processing (NLP) of narratives of longitudinal medical records. The second track of the i2b2/UTHealth shared task focused on identifying risk factors related to Coronary Artery Disease (CAD) in diabetic patients.

According to the World Health Organization, risk factors for a disease increase the chances that a person will develop that disease (WHO, 2014). Diabetes is a risk factor for cardiovascular diseases, including CAD (Dokken, 2008). Other risk factors include: hyperlipidemia/hypercholesterolemia, hypertension, obesity, smoking, and having a family history of CAD (NDIC, 2014). While the obvious way of detecting risk factors in a patient's medical record is to look for diagnoses of the aforementioned diseases, consultations with our medical advisors revealed that a more thorough analysis would go beyond diagnoses. It would consider *indicators of risk factors* which provide medical information that suggests the presence of risk factors. For example, a patient's medical record might not explicitly state that he is diabetic, but an entry of "insulin" in the patient's medication list would be a strong indication that the patient does, in fact, have diabetes. Additionally, indicators can provide evidence of the severity of risk factors. For example, a diagnosis of hypertension in conjunction with high blood pressure measurements and a prescription for blood thinning medication suggests that a patient is more at risk for CAD than a person who has hypertension but is managing it with only diet and exercise.

With these considerations in mind, we devised a shared task that invited participants to identify risk factors and their indicators in narratives of longitudinal medical records. In addition, participants were also asked to identify whether the risk factor or indicator was present before, during, or after the date on the record, giving the potential to create timelines of a patient's progress (or lack thereof) towards heart disease over the course of their longitudinal record.

This shared task differs from many others in the biomedical domain in two key areas: first, the records in the dataset are longitudinal, so they provide snapshots of the patients' progress over months and years. Second, the guiding concept when developing this task was to answer a clinical question about the patient, rather than focus on general syntactic or semantic categories. Specifically, we asked the question "How do diabetic patients progress towards heart disease, specifically coronary artery disease? And how do diabetic patients with coronary artery disease differ from other diabetic patients who do not develop coronary artery disease?" (Stubbs and Uzuner (a), this issue).

This paper provides an overview of the second track (also referred to as Track 2, or the "Risk Factor" or RF Track) of the i2b2/UTHealth 2014 NLP shared task. Section 2 discusses related work, Sections 3 and 4 provide brief descriptions of the data and the annotation process, Section 5 describes the metrics we used to evaluate the participants' systems, Section 6 provides an overview of the top-performing systems, and Sections 7 and 8 discuss the conclusions from the track.

2. Related work

Due to the difficulty of obtaining and sharing medical records (Chapman et al., 2011), few shared tasks have used medical narratives for training and testing. Recent shared tasks that have used medical narratives include the i2b2, the CLEF 2013[1] and 2014[2] shared tasks, and Task 7 from SemEval 2014 (Pradhan et al., 2014).

Previous i2b2 NLP shared tasks include identifying patient smoking status (Uzuner et al., 2007), identifying obesity and its co-morbidities (Uzuner, 2009), extracting information about medications and their dosages (Uzuner et al., 2010), extracting medical concepts, assertions, and their relations (Uzuner et al., 2011), coreference resolution in medical records (Uzuner et al., 2012), and temporal relations between events (Sun et al., 2013). Many of these shared tasks overlap with the RF track in the sense that many of the CAD-related risk factors identified by experts fall into categories examined by previous challenges. For example, diagnoses of "hypertension" were annotated in the medical concepts challenge (Uzuner et al., 2011), obesity and smoking status were the foci of two previous shared tasks (Uzuner et al., 2007; Uzuner, 2009), and medications were extracted in the medication challenge (Uzuner et al., 2010). The RF track builds on these shared tasks to the extent that they support a specific goal: identification of CAD risk factors and indicators.

3. Data

The corpus for this track consisted of 1,304 clinical narratives representing 296 patients (2–5 records per patient). These records were pulled from the Research Patient Data Repository of Partners Healthcare, and were scrubbed of any personal health information (Stubbs and Uzuner (a), this issue; Stubbs and Uzuner (b), forthcoming), which we replaced with realistic surrogates (Stubbs et al., forthcoming). Every patient in this corpus is diabetic, and each patient falls into one of three equally-represented groups: 1) patients who have been diagnosed with CAD starting with their first record in the corpus, 2) patients who develop CAD over the course of their records, and 3) patients who, up until the last record included in the corpus, do not have a diagnosis of CAD. These groupings make it possible for researchers to examine differences between the cohorts, and also help ensure that systems trained on the data are not biased towards one group or another. The training data consists of 60% of the total corpus (790 records), the testing data consists of the remaining 40% (514 records). All the records for a single patient are either in the training or the testing set, and each of the three cohorts are represented equally in the training and testing data. A full description of the data and the corpus selection process can be found in Kumar et al (this issue).

4. Annotation

Here we summarize the goal of the RF track and the annotations. Table 1 (a version of which also appears in Stubbs and Uzuner (a) (this issue)) provides a brief overview of the risk factors and their indicators.

[1] https://sites.google.com/site/shareclefehealth/data

[2] http://clefehealth2014.dcu.ie/task-1/2014-dataset

In Table 1, each risk factor (e.g., Hyperlipidemia) is followed by a list of indicators, i.e., other medical information that is indicative of the risk factor's severity, or that indirectly suggests that the risk factor may be present. For example, a total cholesterol measurement of over 240 indicates that the patient is, if not already hyperlipidemic, at great risk of becoming so.

RF track annotations were generated using a light annotation paradigm which optimizes the use of the annotator's time with the reliability of the annotations (Stubbs, 2013). More specifically, for each risk factor indicator, the annotators created document-level tags that show the presence of the risk factor and its indicator in the patient along with whether the indicator was present in the patient before, during, or after the date of the current record, a.k.a., the document creation time (DCT). For example, in the following text:

> "12/15/2014: 45yo diabetic male w/ history of hypertension admitted for confirmed STEMI. 11/15 A1c 5.5"

would produce the following document-level annotations:

- "diabetic": <DIABETES time="continuing" indicator="mention"/>
- "hypertension": <HYPERTENSION time="continuing" indicator="mention"/>
- "STEMI": <CAD time= "before DCT" indicator= "event"/>

Here, "continuing" is shorthand for the indicator being present before, during, and after the DCT. Also, the A1c level would not be annotated, as it is not over 6.5 (see Table 1).

Medications are treated as a separate category of risk factors, as there is some overlap between, for example, medications used to treat CAD and medications used to treat hypertension. We tracked medication categories such as insulins, beta blockers, and ACE inhibitors (Stubbs and Uzuner (a), this issue).

We hired seven annotators with medical training – one medical doctor, five registered nurses, and one medical assistant – to complete the annotations. Each file was annotated by three, and we used a majority rule to create the gold standard (Stubbs and Uzuner (a), this issue).

i2b2 released the training data to the shared task participants in two batches. The first batch was released in May 2014, and the remainder in June. In July, we released the test data. Participants were asked to stop system development upon accessing the test data and could submit up to three runs of their system for evaluation within three days of test data release.

5. Evaluation

We evaluated systems on document-level annotations with information about risk factors, indicators, and times using micro-averaged precision, recall, and F1 measure on test data. We used F1 as our primary metric. The evaluation scripts are available on GitHub: https://github.com/kotfic/i2b2_evaluation_scripts/tree/v1.2.1

We used approximate randomization (Chinchor, 1992; Noreen, 1989) for significance testing. We tested for significance over micro-averaged P, R and F1, with N = 9,999 and alpha of 0.1.

6. Submissions

Overall, we received 49 submissions from 20 teams for the RF track. Table 6 in the Appendix contains an overview of the teams, the number of members, and their affiliations. Here we present overviews of the systems built by the top 10 teams (sorted alphabetically by team name). National Central University (ranked 8th) did not submit a paper to the workshop and are therefore not included in the overview.

The team from Harbin Institute of Technology Shenzhen Graduate School (Chen et al., this issue), ranked 2nd, divided the risk factors into three categories: phrase-based, logic-based, and discourse-based. Phrase-based risk factors are those that are identified simply by finding relevant phrases in the text, such as “hyperlipidemia” or the name of a particular medication. Logic-based risk factors are those that require a form of analysis after identifying the relevant phrase, such as finding a blood pressure measurement and comparing the numbers to see if they are high enough to count as a risk factor. Finally, discourse-based risk factors are ones that require parsing a sentence, such as identifying smoking status or family history. After pre-processing the texts with MedEx (Xu et al, 2010), the team developed an ensemble of Conditional Random Fields (CRF) and Structural Support Vector Machines (SSVMs) to identify phrase-based risk factors, they utilized rules and output from NegEx[3] for logic-based risk factors, and they studied Support Vector Machines (SVMs) to identify discourse-based ones. Finally, they used a multi-label classification approach to assign temporal attributes to risk factors.

The Kaiser Permanente team (Torii et al., this issue), ranked 3rd, treated the RF track as multiple text categorization tasks. In their words, “each combination of a tag and attribute-value pairs was regarded as an independent target category”. Therefore, the team generated feature sets for these pairs, centered around “hot-spot keywords” collected from the gold standard corpus, and fed them into Weka’s JRip classifier[4]. They built a second classifier for smoking status; this used an SVM whose output could be overruled by a set of regular expressions. Finally, they supplemented their classifier results with output from a third classifier based on Stanford’s Named Entity Recognition tool[5].

The Linguamatics and Northwestern University participants (Cormack et al., this issue), ranked 4th, used an existing text mining platform, I2E (Interactive Information Extraction) to create indexes that include syntactic information (part of speech, tokens, chunks, etc.), and to match terms to existing lexical resources such as SNOMED[6] and RxNorm[7]. They utilized the graphical user interface of I2E to construct queries related to the RF track, including contextual patterns to deal with negation. In order to create a list of appropriate synonyms

[3] https://code.google.com/p/negex/
[4] http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/JRip.html
[5] http://nlp.stanford.edu/software/CRF-NER.shtml
[6] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[7] http://www.nlm.nih.gov/research/umls/rxnorm/

and abbreviations, they used regular expressions and two existing tools based on distributional semantics. Finally, they used context and local dates to add temporal features to the extracted assertions. The candidates for annotation were passed through a series of post-processing steps, which utilized the RF track guidelines and the statistical properties of the training data to assign the final annotations with temporal attributes.

The team from the U.S. National Library of Medicine (NLM) (Roberts et al., this issue), ranked 1st, approached the RF track as a *mention-level* classification task. Using the spans highlighted by the i2b2/UTHealth annotators as a starting point for valid mentions, they re-annotated two-thirds of the training corpus, standardizing the mention spans and annotating both positive and negative mentions. Using that data for training, they pre-processed the documents to identify section headers, negation words, modality words, and output from ConText (Harkema et al., 2009). They used rules to locate trigger words stored in lexicons designed for each risk factor mention, medication, and measurement. They then examined the identified trigger words for each risk factor (with the exceptions of family history and smoking) and other contextual information in a series of SVM classifiers that identified the validity and polarity of each mention. The candidate medication and risk factor annotations were then run through three SVM classifiers that assigned temporal attributes. They identified smoking status by using a single 5-way classifier, and they also used a separate rule-based classifier for family history.

The team from The Ohio State University (Shivade et al., this issue), ranked 6th, identified concepts in the training data, belonging to a variety of openly available terminologies. They used these concepts to trigger rules consisting of regular expressions and UMLS concepts. They suppressed false positives by checking for negation, the experiencer of the event (the patient or someone else), and temporal markers. Further, they created terminology-restricted versions of their system by limiting rules to only those concepts that belonged to a specific terminology. Finally, they used the performance of these terminology-restricted systems as a measure to compare the utility of different terminologies for the RF track.

TMUNSW's team (Chang et al., this issue), ranked 7th, first identified section headers, and used those to classify the document as either a discharge summary or an email. For both types of records, they used both a dictionary-based and a CRF-based system to recognize mentions of the different risk factor concepts, and dictionary- and rule-based approaches to recognize medications and other risk factors such as measurements over a certain amount. The two document types had separate classifiers for identifying smoking status, family history, and time attributes, which used both rule-based and machine learning (Naïve Bayes) systems and made use of cTAKES output.

The participants from the University of Manchester (UNIMAN) (Karystianis et al., this issue), ranked 9th, implemented a rule-based approach, based on identifying semantic groups through the use of custom vocabularies designed for the RF track. The rules they designed aimed to be generic (e.g., for spotting risk factor mentions), while the vocabularies themselves were task-specific. The rules also assigned the temporal information about each identified relevant vocabulary item based on their specific entity class.

University of Nottingham's team (Yang and Garibaldi, (this issue)), ranked 5th, built a system that combines machine learning with dictionary-based and rule-based approaches. First, the team extracted several types of features (e.g., token, context, section, and task-specific features) from the text, which they then turned into features for a set of systems designed to identify disease risk factors: a CRF system that identified token-level entities, a set of three classifiers (CRF, Naïve Bayes, and Maximum Entropy) that identified sentence-level facts, and handwritten rules for identifying sentence-level measurements. For medication names, they used a CRF and dictionary lookup. Finally, a set of heuristic rules was applied to add temporal attributes to each tag.

The participants from the University of Utah (Khalifa and Meystre, this issue), ranked 10th, used combinations of existing tools and their own regular expressions, along with the UMLS Metathesaurus[8] to identify risk factors, implemented in the Apache UIMA[9] framework. First, they used cTAKES' built-in preprocessing tools, then ran the cTAKES smoking status identifier. They then used regular expressions to identify medications from a manually curated list and to identify applicable lab test results. They used the UMLS Metathesaurus module in Textractor (Meystre et al., 2010) to identify diseases and risk factors, then match them to the Concept Unique Identifiers (CUIs) that apply to the RF track. Finally, they performed a contextual analysis to remove risk factors that do not relate to the patient (i.e., negated), and identify family history of CAD using ConText. The time attributes were generated for each category of information based on most common time values found in the training data.

Overall, the systems built for the RF track vary widely, from exclusively rule-based systems to complex hybrids of rules and combinations of machine learning algorithms, and there was no consensus as to what features and algorithms would be best for this track: each team with a hybrid system used a different combination of features and algorithms.

However, there were some similarities between the top-performing approaches. For example, all the systems used pre-processing tools to gain syntactic information, all but one (Kaiser) added temporal attributes through a separate process at the end, and nearly all the systems used medical lexicons, either curated from the gold standard or from existing resources such as the UMLS, Drugs.com, Wikipedia, etc. Only Harbin Grad did not mention using a lexicon of medical terms.

Over half of the systems made use of section header information at some point during the process, often for the purpose of adding temporal information (i.e., NLM, Kaiser, Harbin Grad, Nottingham, Ohio, TMUNSW), and four of the systems assigned default temporal attributes based on the annotation categories (disease, medication, measurement) at least some of the time (i.e., Utah, UNIMAN, Linguamatics, NLM).

[8] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

[9] https://uima.apache.org/

7. Results

As discussed in Section 5, we used the micro-averaged F1 as our primary metric. Table 2 shows the precision, recall, and F1 at the micro level for the top 10 systems, along with a summary of the approach taken by that team. All of the systems achieved F1s over 0.87, with recall measurements being higher than precision for all systems. These results show that identifying risk factors, indicators, and their temporal labels is well within the reach of automated systems.

Table 3 shows the results of the significance testing between the top 10 systems. Note that we only show the lower half the table, as the upper diagonal would be symmetrically identical to the lower. Cells containing P, R, or F indicate that the two systems are *not* significantly different in precision, recall, or F1, respectively. Overall, we see a fair amount of similarity in the system outputs, especially among the top few systems.

Table 4 shows the micro-averaged F1s for each of the individual categories for the top run from each team. CAD and smoking status proved to be slightly more difficult than the other categories for most participants, with hypertension and family history having the best performance. However, part of the high performance on the family history is due to the fact that the majority of records either indicated no family history, or did not mention the family history at all (Stubbs and Uzuner (a), this issue).

Figure 1 in the Appendix shows the micro F1 for each risk factor for the top 10 submissions. Figures 2 through 7 in the Appendix show the breakdown of the different risk factor indicators by F1, along with the overall scores for each risk factor. We calculated these by evaluating each indicator individually, and then calculating the scores for all indicators for each risk factor combined. The figures show that, with the exception of CAD, most risk factors had significantly more mentions than any other type of indicator, meaning that the overall scores often shadow the mention scores. CAD indicators, on the other hand, were sufficiently varied and numerous to pull the overall score away from the scores for mentions. These trends indicate that in-text diagnoses are by far the most common risk factor indicators, and also that other indicators, such as blood test results, are much more difficult to identify accurately.

Figure 8 in the appendix shows the smoking categories, as well as the overall scores for that risk factor. Like the other risk factors, the overall scores closely mirror the scores for the best-represented category in the corpus, the “Unknown” category. The “Ever” category was the most difficult to identify, with many teams getting 0.

Due to the number of medication categories, we do not have a chart that shows a breakdown of every category, though Table 4 shows that overall, the F1 measures for medications range from 0.8585 to 0.9307.

Table 7 in the appendix shows the exact numbers for Figures 1–8, for more precise comparisons between teams and risk factors.

7.1 Comparison to other shared tasks

We noted in Section 3 the similarities between some of the previous i2b2 NLP shared tasks and the RF track. While most of the previous challenges are not identical to the RF track, we can still perform some basic comparisons.

7.1.1 Smoking Status—The risk factor annotations used the same smoking status categories as the 2006 i2b2 challenge (Uzuner et al., 2007): past smoker (quit over a year ago), current smoker (presently smokes), has ever smoked (smoked in the past, present status unknown, called "smoker" in the 2006 challenge), never smoked (called "non-smoker" in 2006) and unknown (smoking status not mentioned in the document).

The 2006 smoking shared task data set consisted of 502 medical records; the 2014 corpus had 1,304. The left side of Table 5 show the relative percentages of the different smoking categories in the two corpora.

The distributions of smoking categories in the two corpora differ substantially; this may in part be due to the 2014 data having been chosen for patients at risk for CAD, meaning that it is more likely that the doctor would make a note of whether or not the patient smokes, and possibly even more likely for the patients to have quit smoking out of concern for their health.

The right side of Table 5 shows the comparison of the micro-averaged F1 scores per smoking category of the top 2006 (Clark_3) and 2014 (Nottingham) systems. The Nottingham system performed significantly better in the "Past" category, and worse in the "Current" and "Smoker" categories. These changes are likely explained by the different distributions of the categories in the two corpora, as the 2014 corpus contained many more "Past" categories and fewer "Current" and "Smoker" categories. The only exception to this trend is the "Non-smoker" category, which was better represented in the 2014 corpus, but on which the Nottingham system performed slightly worse. Overall, the differences in these results suggest that the amount of training data for categories has the biggest impact on the success of the systems.

7.1.2 Obesity and comorbidities; concept extraction—The 2008 obesity shared task (Uzuner, 2009) presents a less straightforward comparison than the smoking status. The 2008 challenge focused not only on identifying obesity, but also its comorbidities, i.e., diseases that frequently occur in conjunction with obesity. This list of comorbid diseases included all of the diseases included in the RF track: CAD, hyperlipidemia, hypertension, and diabetes, along with other diseases such as asthma, GERD, and gout. Each document in the corpus was assigned a class for each disease: present, absent, questionable, and unmentioned. Annotation into these classes had two facets: a "textual" annotation that required a diagnosis of the disease in the text before the "present" label could be applied, and an "intuitive" annotation which allowed the annotators to take other information into account, similar to the way risk factor indicators are used in the 2014 annotation. For example, they could use a description of a person weighing 350lbs as the basis for an intuitive judgment that the patient was obese. For this reason, we compare the overall results from the 2014 shared task to the results of the intuitive annotation from the 2008 shared task.

In 2008, the highest F1 over the micro-averaged intuitive system result was 0.9654; in 2014 the best overall F1 was NLM's 0.9276 (Table 2). The discrepancy in scores is likely caused by the use of the "absent" annotations in the 2008 corpus. The 2008 scoring system counted correct "absent" annotations as true positives, thereby increasing the F1 metrics, and the number of "absent" annotations would have likely benefited systems which defaulted to that label. In contrast, the 2014 annotations only included diseases and risk factors that were present in the files, and the scoring system did not reward them for correctly leaving files unmarked. In addition, in the 2008 corpus the "absent" annotations greatly outnumbered the "present" annotations: the 2008 training corpus contained 3,267 "present" annotations and 7,362 "absent" annotations; the test corpus contained 2,285 "present" annotations and 5,100 "absent" annotations. The systems trained on the intuitive data scored best on the "absent" annotations, likely aided by the overabundance of those examples in the corpus.

The other i2b2 shared task that is similar to the "mentions" is the 2009 challenge on concepts, assertions, and relations (Uzuner et al., 2010). The concept extraction track asked participants to identify all "medical problems, treatments, and tests". The top-performing system on that track achieved an overall "inexact" F1 of 0.924, though given that the 2009 challenge encompassed a much broader range of entities, it is not surprising that the results from that challenge would be lower.

7.3 Medications

The 2008 i2b2 NLP shared task on medication extraction (Uzuner et al., 2009) asked participants to identify for all medications mentioned in a discharge summary "their names, dosages, modes (routes) of administration, frequencies, durations, and reasons for administration" at both the phrase and token levels. The 2014 RF track differs in that we asked participants to look only for particular classes of medications and their temporal markers, with annotations at the document level. Uzuner et al., (Uzuner et al., 2009) report that the best-performing system achieved an F1 of 0.884 at the phrase level (the entire, multi-word annotation) and 0.903 at the token level (looking at each individual token/word within the phrases) for identifying medication names in 2008. In comparison, Table 4 shows that the best-performing system from 2014 achieved an F1 of 0.9307. The document-level annotations of the 2014 challenge lend themselves to somewhat higher results, though the addition of temporal information does increase the complexity.

8. Discussion

In the Introduction to this paper, we discussed how this track and the accompanying annotations were designed with the following questions in mind: "How do diabetic patients progress towards heart disease, specifically CAD? And how do diabetic patients with CAD differ from other diabetic patients who do not develop CAD?" So we now ask: Based on the results of the RF track, can those questions be answered by automated systems?

Admittedly, this track did not link the individual records of a patient, so it does not directly address the question of progression. However, the longitudinal nature of the data provides snapshots of the patients throughout their treatments, and the annotations give information on what risk factors and indicators are present before, during, and after each document. So the tools for building timelines are present in the gold standard. But, can the systems recreate the gold standard?

Based on the overall results, it seems that they can: 6 of the top 10 achieved micro-averaged F1 measures of over 0.9, and all of the top 10 systems scored over 0.85 in precision, recall, and F1. While certain indicators were harder to correctly identify than others, overall the results from the RF track are positive and show that automated systems can, with appropriate training data, recognize patients who are at risk for CAD.

Some aspects of the RF track proved harder than others. Number-based indicators (i.e., A1c, glucose, cholesterol, LDL, blood pressure, and BMI measurements) all have significantly lower F1s than mentions. In part, this is likely due to simple sparsity of training data: compared to the mentions, the number-based indicators show up infrequently (Stubbs and Uzuner (a), this issue). However, a contributing factor is likely that many of these measurements appeared in tables of lab values, making it extremely difficult to construct feature sets or rules that could accurately determine which values were associated with which test and which date. This problem was compounded by the fact that often, structural integrity of the tables had been lost, making parsing even more difficult. For example, a file containing a table with tabs separating the columns may have had the tabs turned into s paces through a conversion error, making the table nearly unreadable.

The CAD indicators test, evaluation, and symptom, as well as files annotated as having a family history of CAD, also had comparatively low F1s in all the systems. While again, this could in part be due to sparsity of data, as there were relatively few examples of each of these indicators, it is likely that a contributing problem was the extreme variety of ways that these indicators were described in the text. While mentions of CAD, diabetes, and the other diseases can vary (for example, diabetes can be “diabetes mellitus”, “DM”, “DMII”, “DM2”, “t2dm”, and so on), these phrases alone, along with some basic polarity checking, is generally enough to identify positive diagnoses in the text. On the other hand, a CAD-related event could be as simple as “cardiomyopathy” or as complicated as “probable inferior and old anteroseptal myocardial infarction”, “s/p MI in 4/80”, “quadruple bypass 2096”, or “emergent catheterization”, all of which are examples from the gold standard annotations. Similar variations appear in the annotations for tests, evaluations, and family histories. The fact that these indicators were so much harder to correctly identify suggests the need for more accurate semantic matching of medical record text to resources such as the UMLS, so that the context of relevant text can be better evaluated.

9. Conclusion

This paper presents an overview of the 2014 i2b2/UTHealth NLP shared task track on identifying risk factors for CAD in longitudinal patient records. Evaluation consisted of precision, recall, and F1 at the micro level when comparing the system outputs to the document-level gold standard. 20 teams participated in this track, and the best-performing system achieved an F1 of 0.9276, and all of the top 10 systems achieving F1s over 0.87.

The high scores on this track suggest that it is feasible to train systems to identify diabetic patients who are at risk for CAD by identifying risk factors and indicators that are related to CAD, and that these systems can be trained with lightly annotated gold standards. However, the difficulty in processing complicated concepts and extracting certain types of numerical data suggests open questions that yet need to be addressed in future NLP research.

Overall, the results of the 2014 i2b2/UTHealth NLP shared task RF track are promising, and point the way towards using computers to help identify patients who are at risk for diseases.

Acknowledgments

We would like to thank the program committee for the 2014 i2b2/UTHealth NLP Shared Task, along with everyone who participated in the shared task and workshop.

Funding for this project was provided by:

- NIH NLM 2U54LM008748, PI: Isaac Kohane
- NIH NLM 5R13LM011411, PI: Ozlem Uzuner
- NIH NIGMS 5R01GM102282, PI: Hua Xu

Works Cited

Chang, Nai-Wen; Dai, Hong-Jie; Chen, Chih-Wei; Jonnagaddala, Jitendra; Chien, Chou-Yang; Kumar, Manish; Tsai, Richard Tzong-Han; Hsu, Wen-Lian. TMUNSW System for Risk Factor Recognition and Progression Tracking. Journal of Biomedical Informatics. this issue.

Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association. 2011; 18(5):540–543. [PubMed: 21846785]

Chen, Qingcai; Li, Haodi; Tang, Buzhou; Liu, Xin; Liu, Zengjian; Liu, Shu; Wang, Weida. risk factors for heart disease over time –HITSZ's system for track 2 of the 2014 i2b2 NLP challenge. Journal of Biomedical Informatics. this issue.

Chinchor, Nancy. The statistical significance of the MUC-4 results. Proceedings of the 4th conference on Message understanding. 1992:30–50.

Cormack, James; Nath, Chinmoy; Milward, David; Raja, Kalpana; Jonnalagadda, Siddhartha. Agile Text Mining for the i2b2 2014 Cardiac Risk Factors Challenge. Journal of Biomedical Informatics. this issue.

Dokken, Betsy B. The Pathophysiology of Cardiovascular Disease and Diabetes: Beyond Blood Pressure and Lipids. 2008 Jul; 21(3):160–165. http://spectrum.diabetesjournals.org/content/21/3/160.full.

Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics. 2009 Oct; 42(5):839–851. [PubMed: 19435614]

Karystianis, George; Dehghan, Azad; Kovačević, Aleksandar; Keane, John A.; Nenadic, Goran. Using Local Lexicalized Rules for Identification of Heart Disease Risk Factors in Free-text Clinical Notes. Journal of Biomedical Informatics. this issue.

Khalifa, Abdulrahman; Meystre, Stéphane M. Identification of Risk Factors for Heart Disease in Electronic Health Records of Diabetic Patients. Journal of Biomedical Informatics. this issue.

Kumar V, Stubbs A, Shaw S, Uzuner O. Creation of a new longitudinal corpus of clinical narratives. Journal of Biomedical Informatics. this issue.

Meystre SM, Thibault J, Shen S, Hurdle JF, South BR. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. Journal of the American Medical Informatics Association: JAMIA. 2010; 17(5):559–562. [PubMed: 20819864]

NDIC (National Diabetes Information Clearinghouse). [Last updated February 19, 2014] Diabetes, Heart Disease, and Stroke. http://diabetes.niddk.nih.gov/dm/pubs/stroke/index.aspx

Noreen, EW. Computer-intensive methods for testing hypotheses: an introduction. New York: Wiley; 1989.

Pradhan, S.; Elhadad, N.; Chapman, W.; Manandhar, S.; Savova, G. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014. Dublin, Ireland: Association for Computational Linguistics and Dublin City University; 2014 Aug. SemEval-2014 Task 7: Analysis of Clinical Text; p. 54-62.http://www.aclweb.org/anthology/S14-2007

Roberts, Kirk; Shooshan, Sonya E.; Rodriguez, Laritza; Abhyankar, Swapna; Kilicoglu, Halil; Demner-Fushman, Dina. NLM: Machine Learning Methods for Detecting Risk Factors for Heart Disease in EHRs. Journal of Biomedical Informatics. this issue.

Shivade, Chaitanya; Malewadkar, Pranav; Fosler-Lussier, Eric; Lai, Albert M. Comparison of UMLS Terminologies to Identify Risk of Heart Disease in Clinical Notes. Journal of Biomedical Informatics. this issue.

Stubbs, Amber. Doctoral Dissertation. Brandeis University; 2013 Feb. A Methodology for Using Professional Knowledge in Corpus Annotation.

Stubbs A, Uzuner Ö. (a). Annotating Risk Factors for Heart Disease in Clinical Narratives for Diabetic Patients. Journal of Biomedical Informatics. This issue.

Stubbs A, Uzuner Ö. (b). De-identifying longitudinal medical records. Journal of Biomedical Informatics. This issue.

Stubbs, A.; Uzuner, Ö. (c). De-identification of Medical Records Through Annotation. In: Ide, Nancy; Pustejovsky, James, editors. Chapter in Handbook of Linguistic Annotation. Springer. Anticipated publication; 2015.

Stubbs, A.; Uzuner, Ö.; Kotfila, C.; Goldstein, I.; Szolovitz, P. Challenges in Synthesizing Replacements for PHI in Narrative EMRs. In: Aris, Gkoulalas-Divanis; Loukides, Grigorios, editors. Chapter in Medical Data Privacy Handbook. Springer. Anticipated publication; 2015.

Sun W, Rumshisky A, Uzuner O. Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge Overview Jounal of the American Medical Association. 2013

Torii, Manabu; Fan, Jung-wei; Yang, Wei-li; Lee, Theodore; Wiley, Matthew T.; Zisook, Daniel; Huang, Yang. De-Identification and Risk Factor Detection in Medical Records. Journal of Biomedical Informatics. this issue.

Uzuner Ö. Recognizing Obesity and Comorbidities in Sparse Data Journal of the American Medical Informatics Association. 2009; 16:561–570. [PubMed: 19390096]

Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. Journal of the American Medical Informatics Association. 2012; 19(5):786–791. [PubMed: 22366294]

Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. Jounal of the American Medical Informatics Association. 2007; 15:14–24.

Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association. 2010; 17:514–518. [PubMed: 20819854]

Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text Journal of the American Medical Informatics Association. 2011; 18:552–556. [PubMed: 21685143]

WHO (World Health Organization). [Last updated 2014] Health Topics: Risk Factors. http://www.who.int/topics/risk_factors/en/

Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association? JAMIA. 2010; 17(1):19–24. [PubMed: 20064797]

Yang, Hui; Garibaldi, Jonathan. Automatic Extraction of Risk Factors for Heart Disease in Clinical Texts. Journal of Biomedical Informatics. this issue.

Appendix

Figure 1:

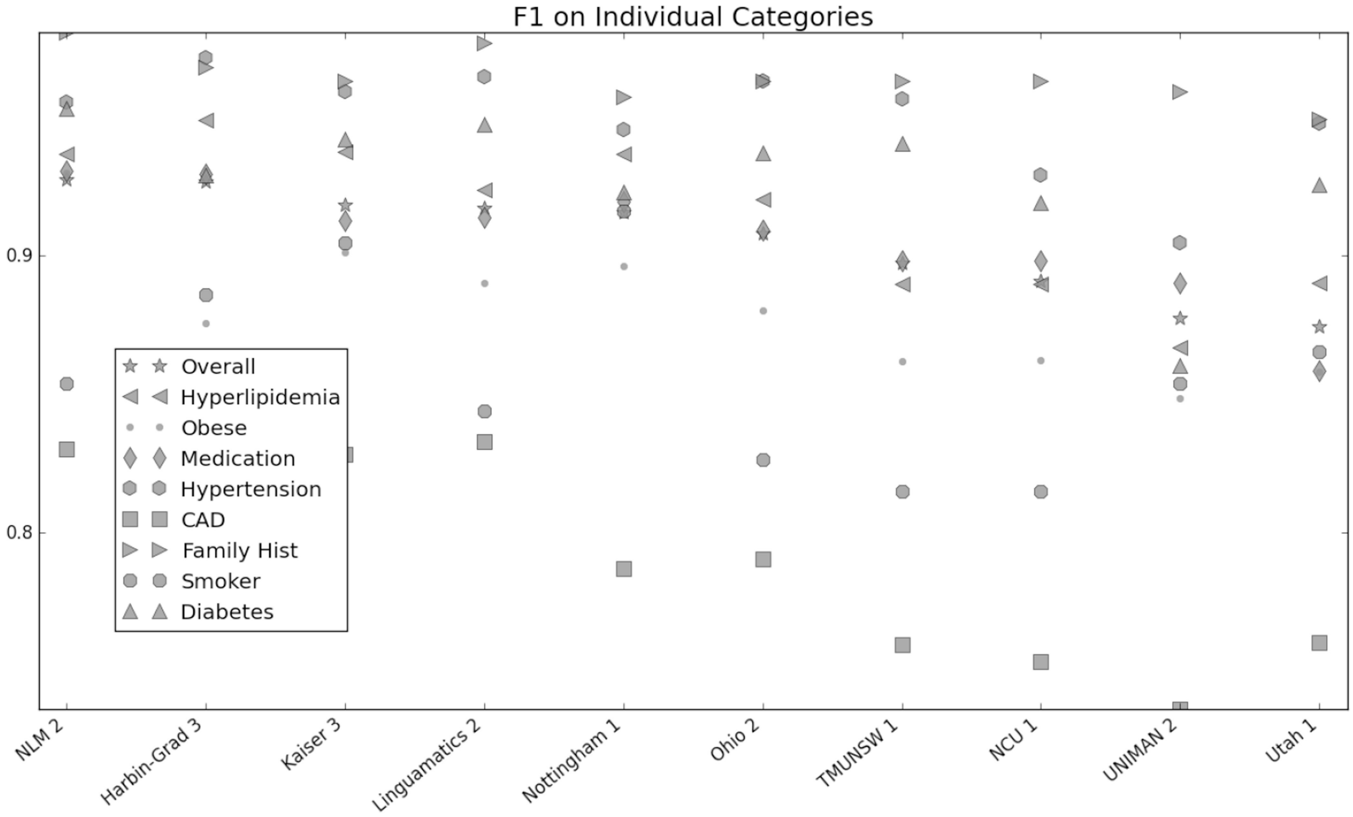


Figure 2:

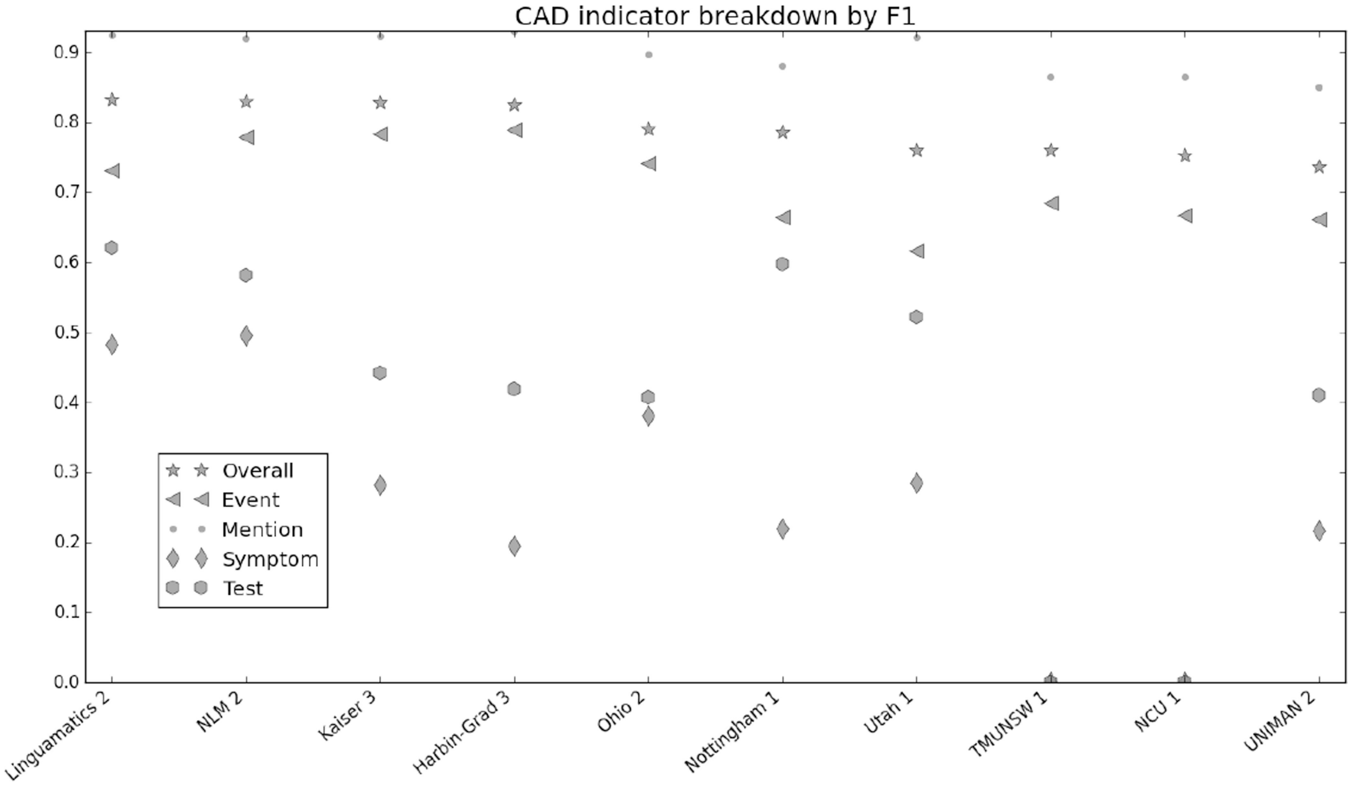

CAD indicator breakdown by F1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Overall
Event
Mention
Symptom
Test
Linguamatics 2
NLM 2
Kaiser 3
Harbin-Grad 3
Ohio 2
Nottingham 1
Utah 1
TMUNSW 1
NCU 1
UNIMAN 2

Figure 3:

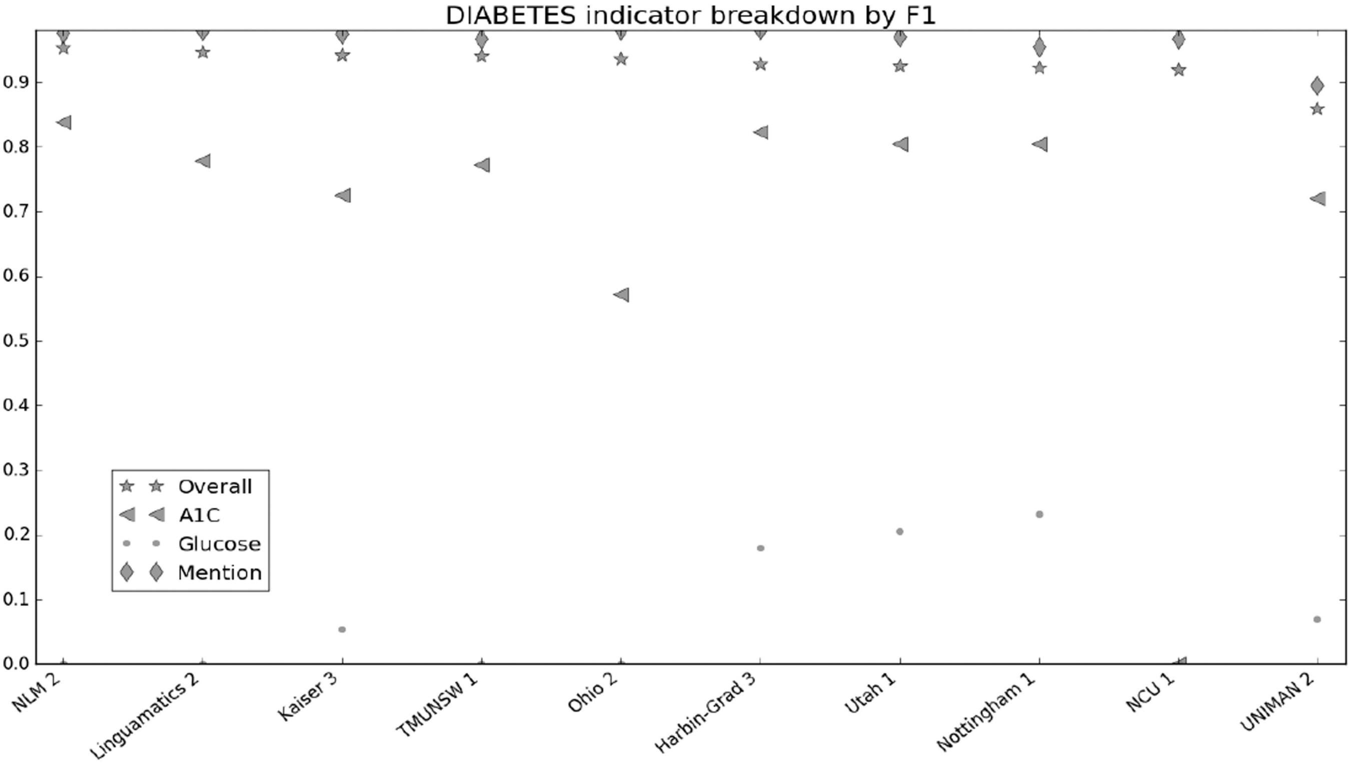

DIABETES indicator breakdown by F1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Overall
A1C
Glucose
Mention
NLM 2
Linguamatics 2
Kaiser 3
TMUNSW 1
Ohio 2
Harbin-Grad 3
Utah 1
Nottingham 1
NCU 1
UNIMAN 2

Figure 4:

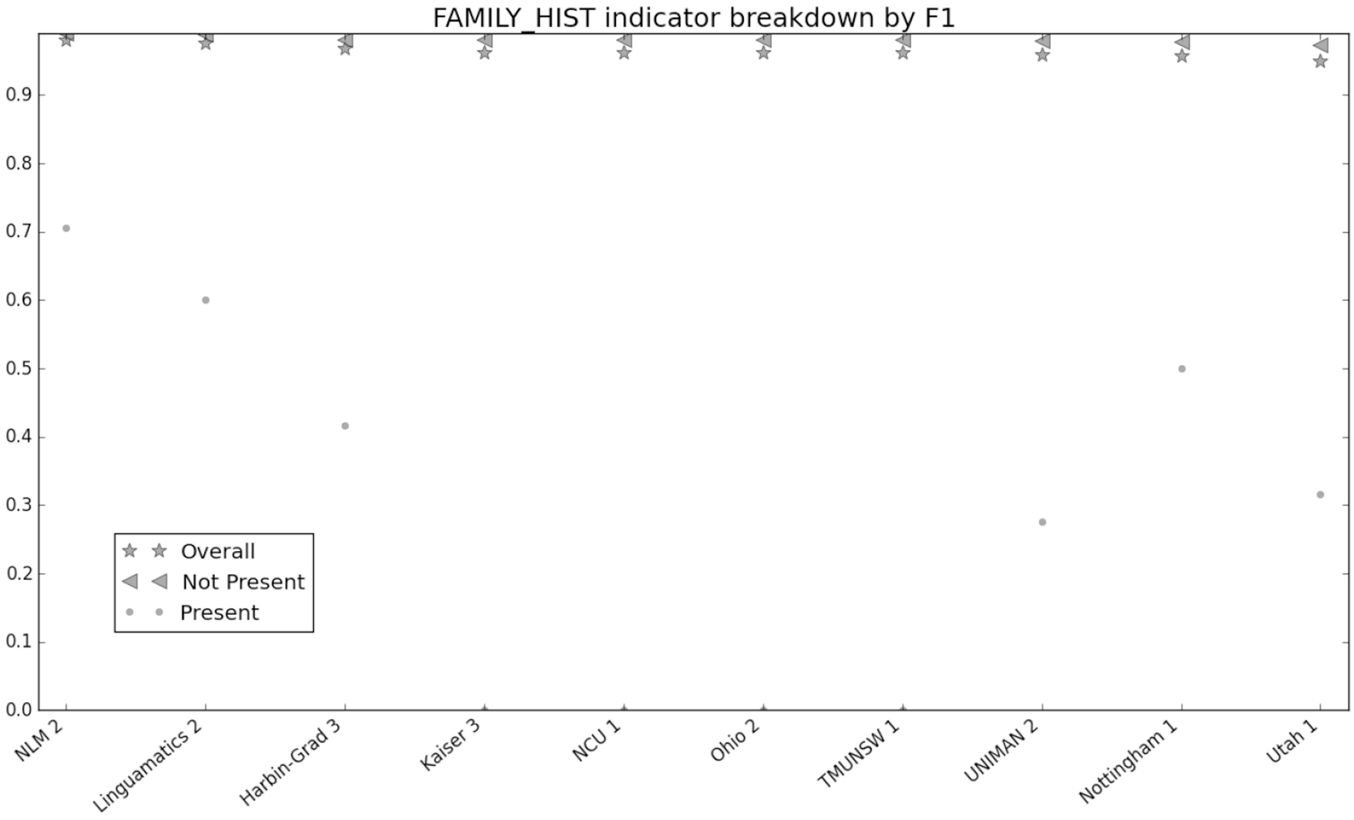


Figure 5:

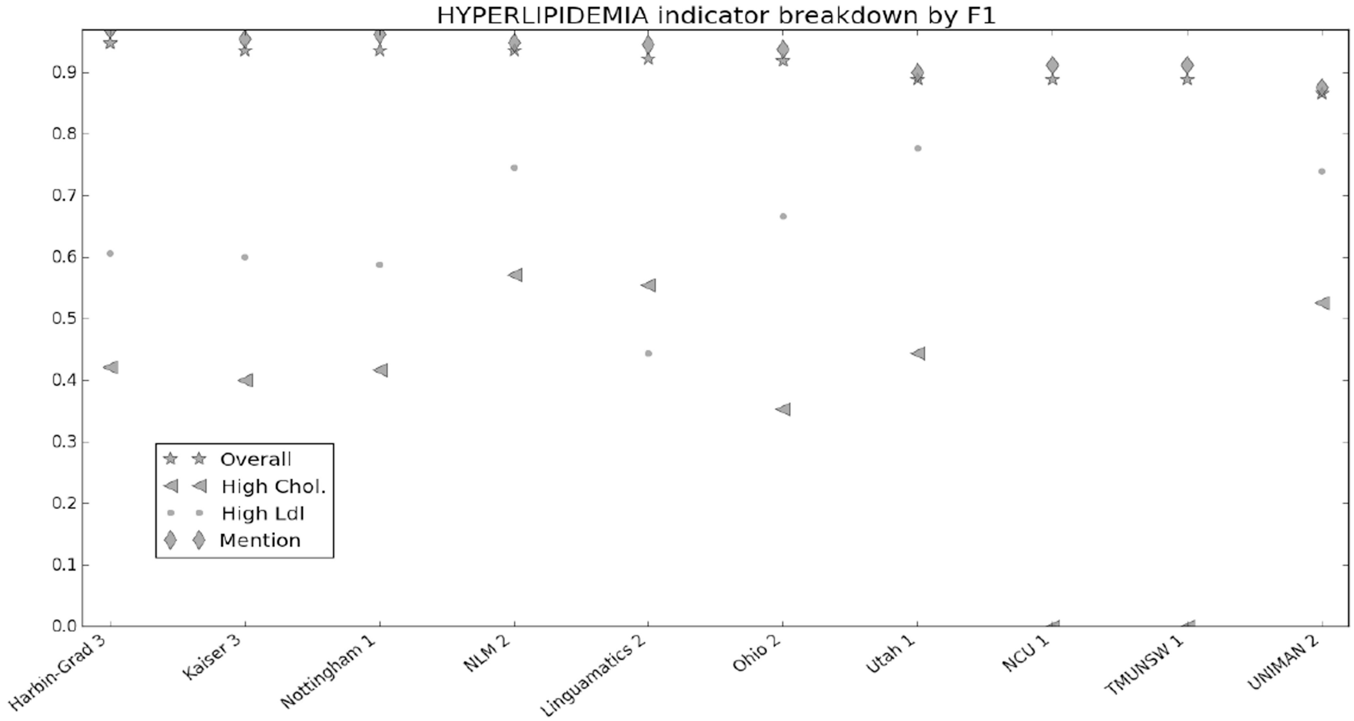


Figure 6:

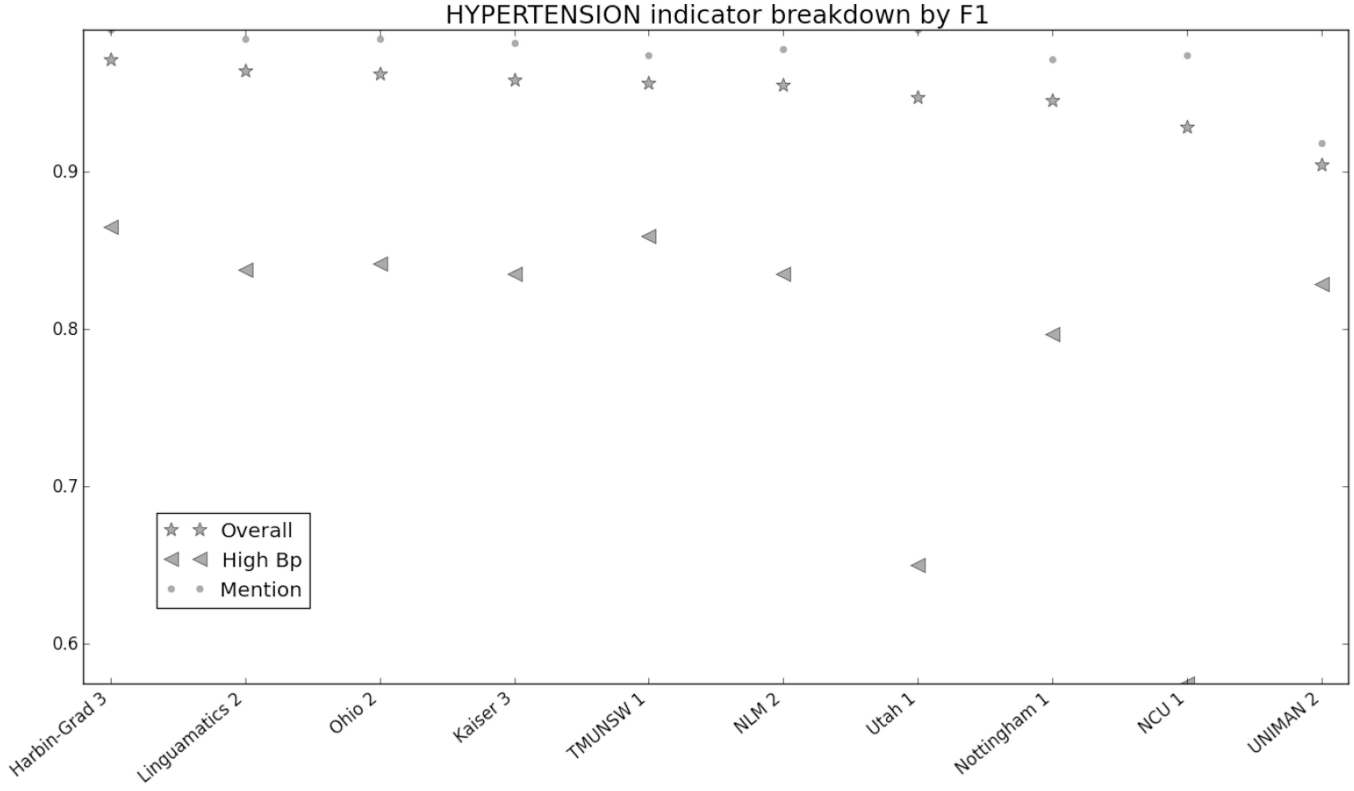


Figure 7:

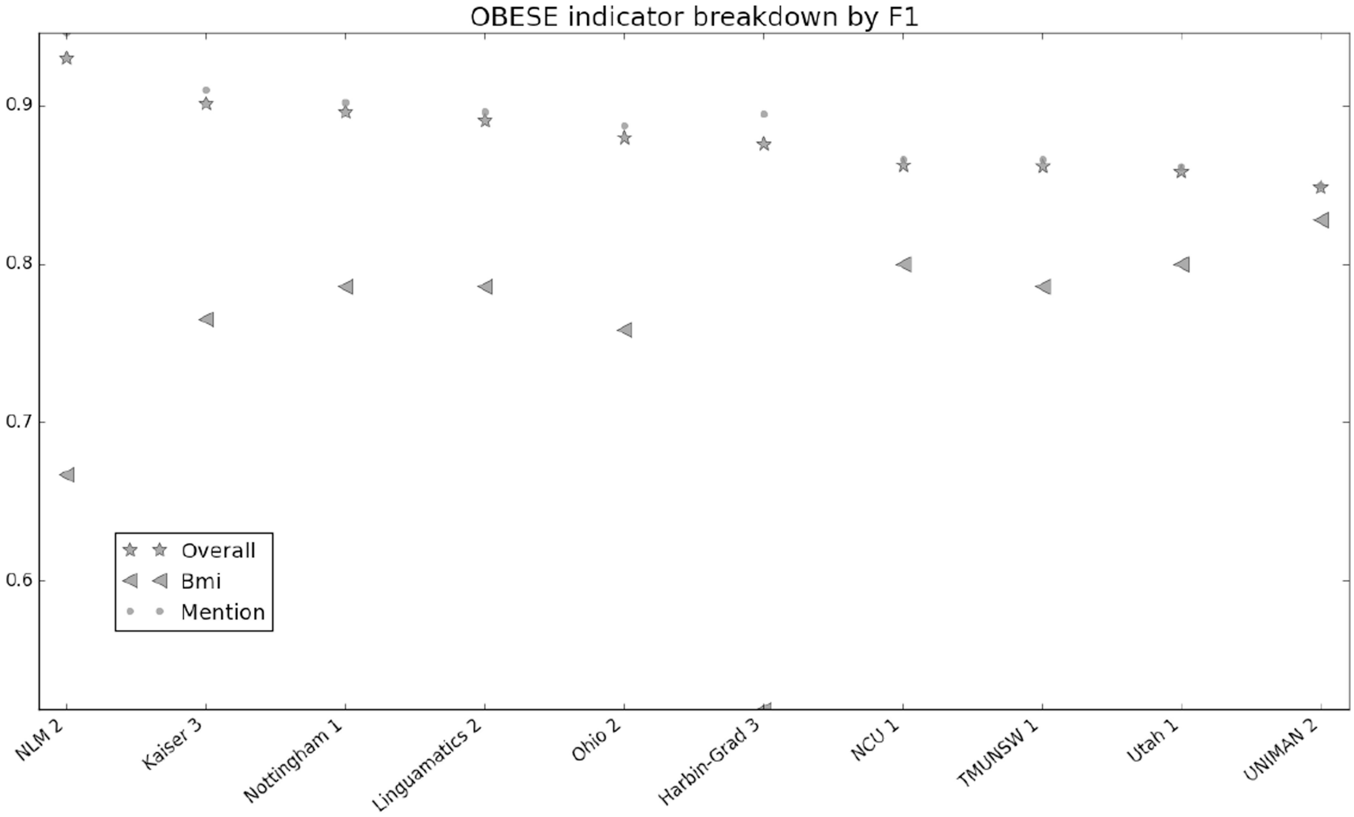

OBESE indicator breakdown by F1
0.9
0.8
0.7
0.6
Overall
Bmi
Mention
NLM 2
Kaiser 3
Nottingham 1
Linguamatics 2
Ohio 2
Harbin-Grad 3
NCU 1
TMUNSW 1
Utah 1
UNIMAN 2

Figure 8:

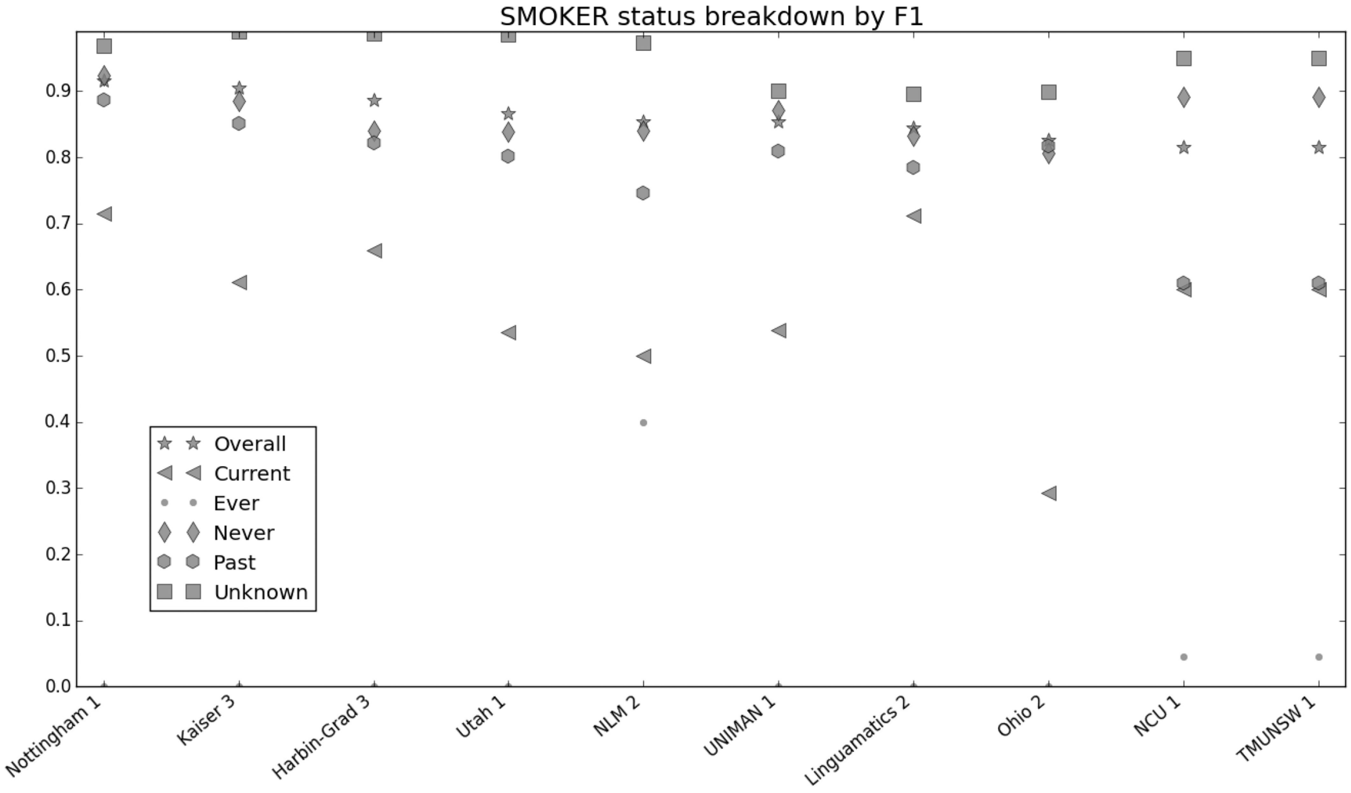


Table 6

Participants in Track 2 of the 2014 i2b2/UTHealth NLP shared task

Team name	affiliations	# of members	Countries
Harbin-Grad	Harbin Institute of Technology Shenzhen Graduate School	7	China
LIMSI-CNRS	Centre National de la Recherche Scientifique Universit e Paris-Sud	4	France
Linguamatics	Linguamatics Ltd. Northwestern University	5	UK USA
Kaiser	Kaiser Permanente Southern California	7	USA
NLM	U.S. National Library of Medicine	6	USA
Ohio	The Ohio State University	4	USA
Nottingham	University of Nottingham	2	UK
TMUNSW	Academia Sinica National Taiwan University Taipei Medical University University of New South Wales National Central University	8	Taiwan Australia
NCU	National Central University	2	Taiwan
UNIMAN	University of Manchester University of Novi Sad Health eResearch Centre	5	UK Serbia
Utah	University of Utah	2	USA
Mayo	Mayo Clinic	5	USA

Team name	**affiliations**	**# of members**	**Countries**
Zhejiang	Zhejiang University The Children's Hospital of Zhejiang University School of Medicine	4	China
UTHouston	University of Texas at Houston	1	USA
UTDallas	University of Texas at Dallas	2	USA
Milwaukee	Milwaukee School of Engineering Medical College of Wisconsin	1	USA
Seoul	Seoul National University	5	South Korea
Tongji	Tongji University Shibei District People's Hospital of Qingdao	4	China
Drexel	Drexel University	2	USA
Lira	Unknown	3	USA

Table 7

F1 scores for each team's best run, grouped by risk factors

	NLM	**Harbin-Grad**	**Kaiser**	**Lingua-matics**	**Nottingham**	**Ohio**	**TMUNSW**	**NCU**	**UNIMAN**	**Utah**
CAD Overall	0.8303	0.8253	0.8284	**0.8331**	0.7870	0.7904	0.7598	0.7536	0.7363	0.7603
CAD event	0.7795	**0.7900**	0.7832	0.7305	0.6638	0.7412	0.6852	0.6667	0.6610	0.6165
CAD mention	0.9205	**0.9299**	0.9227	0.9251	0.8813	0.8962	0.8653	0.8653	0.8501	0.9222
CAD symptom	**0.4957**	0.1957	0.2828	0.4821	0.2198	0.3817	0.0000	0.0000	0.2162	0.2844
CAD test	0.5814	0.4190	0.4421	**0.6207**	0.5972	0.4065	0.0000	0.0000	0.4098	0.5217
Diabetes Overall	**0.9533**	0.9291	0.9420	0.9473	0.9228	0.9369	0.9406	0.9191	0.8603	0.9256
Diabetes A1c	**0.8383**	0.8228	0.7251	0.7785	0.8047	0.5714	0.7730	0.0000	0.7205	0.8052
Diabetes glucose	0.0000	0.1803	0.0541	0.0000	**0.2319**	0.0000	0.0000	0.0000	0.0690	0.2056
Diabetes mention	0.9766	0.9802	0.9742	0.9790	0.9545	**0.9806**	0.9672	0.9672	0.8950	0.9696
Family hist FH) overall	**0.9805**	0.9681	0.9630	0.9767	0.9572	0.9630	0.9630	0.9630	0.9591	0.9494
FH not present	**0.9899**	0.9812	0.9812	0.9880	0.9776	0.9812	0.9812	0.9812	0.9790	0.9737
FH present	**0.7059**	0.4167	0.000	0.6000	0.5000	0.0000	0.0000	0.0000	0.2759	0.3158
Hyperlip. (HL) overall	0.9366	**0.9491**	0.9375	0.9236	0.9368	0.9204	0.8896	0.8896	0.8669	0.8903
HL high chol.	**0.5714**	0.4211	0.4000	0.5556	0.4167	0.3529	0.0000	0.0000	0.5263	0.4444
HL high LDL	0.7458	0.6071	0.6000	0.4444	0.5882	0.6667	0.0000	0.0000	0.7407	**0.7778**
HL mention	0.9498	**0.9700**	0.9552	0.9467	0.9622	0.9386	0.9130	0.9130	0.8767	0.9011
Hyperten (HT) overall	0.9555	**0.9715**	0.9591	0.9647	0.9455	0.9630	0.9567	0.9289	0.9047	0.9477
HT high BP	0.8353	**0.8651**	0.8350	0.8380	0.7970	0.8418	0.8592	0.5744	0.8289	0.6497
HT mention	0.9786	**0.9905**	0.9820	0.9851	0.9719	0.9846	0.9746	0.9746	0.9183	0.9904
Obesity overall	**0.9298**	0.8757	0.9011	0.8902	0.8961	0.8801	0.8620	0.8624	0.8486	0.8581

	NLM	**Harbin-Grad**	**Kaiser**	**Lingua-matics**	**Nottingham**	**Ohio**	**TMUNSW**	**NCU**	**UNIMAN**	**Utah**
Obesity BMI	0.6667	0.5185	0.7646	0.7857	0.7857	0.7586	0.7857	0.8000	**0.8276**	0.8000
Obesity mention	**0.9457**	0.8945	0.9098	0.8961	0.9019	0.8868	0.8657	0.8657	0.8498	0.8612
Smoker Overall	0.8538	0.8861	0.9045	0.8441	**0.9162**	0.8264	0.8148	0.8148	0.8538	0.8655
Smoker current	0.5000	0.6588	0.6111	0.7123	**0.7143**	0.2927	0.6000	0.6000	0.5385	0.5357
Smoker ever	**0.4000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0444	0.0444	0.0000	0.0000
Smoker never	0.8405	0.8397	0.8854	0.8318	**0.9237**	0.8061	0.8908	0.8908	0.8707	0.8384
Smoker past	0.7464	0.8219	0.8505	0.7839	**0.8869**	0.8173	0.6098	0.6098	0.8098	0.8020
Smoker unknown	0.9726	0.9876	**0.9896**	0.8955	0.9688	0.8993	0.9499	0.9499	0.9006	0.9856

Highlights

- First NLP shared task on identifying risk factors and related indicators in diabetic patients
- Twenty teams participated, submitted 49 system runs
- Six of the top 10 teams achieved F1 scores over 0.90; all 10 scored over 0.87

Table 1

List of risk factors and their indicators

Diabetes indicators: • **Mention**: A diagnosis of Type 1 or Type 2 diabetes • **Test**: An A1c test value of over 6.5 or 2 fasting blood glucose measurements of over 126 Hyperlipidemia/Hypercholesterolemia indicators: • **Mention**: A diagnosis of Hyperlipidemia or Hypercholesterolemia • **High cholesterol**: Total cholesterol of over 240 • **High LDL**: LDL measurement of over 100mg/dL Hypertension indicators: • **Mention**: A diagnosis of hypertension • **High blood pressure**: BP measurement of over 140/90 mm/hg Family History of premature CAD: • Patient has a first-degree relative (parents, siblings, or children) who was diagnosed prematurely (younger than 55 for male relatives, younger than 65 for female relatives) with CAD	CAD indicators: • **Mention**: A diagnosis of CAD • **Event**: An event indicative of CAD (MI, STEMI, NSTEMI, revascularization procedures, cardiac arrest, ischemic cardiomyopathy) • **Test**: Test results: exercise or pharmacologic stress test showing ischemia, or abnormal cardiac catheterization showing coronary stenoses • **Symptom**: Chest pain consistent with angina Obesity indicators: • **Mention**: A description of the patient as being obese • **High body mass index (BMI)**: BMI over 30 • **Large waist circumference**: Waist circumference measurement of: – men: 40 inches or more – women: 35 inches or more Smoker indicator: • Currently smoking or has smoked within the past year Medications: • Any medication used to treat the other risk factors or indicators

Table 2

Results for top 10 Risk Factor Track submissions, ranked by micro-averaged F1

Rank	Team name	Micro Precision	Micro Recall	Micro F1	System
1	NLM	0.8951	0.9625	0.9276	Additional annotation + lexicon and rules + SVMs
2	Harbin Institute of Technology Shenzhen Graduate School	0.9106	0.9436	0.9268	Ensemble systems + rules
3	Kaiser Permanente	0.8972	0.9409	0.9185	SVM+ensemble classifier+rules
4	Linguamatics and Northwestern University	0.8975	0.9375	0.9171	I2E interface + lexicon + existing tools + rules
5	University of Nottingham	0.8847	0.9488	0.9156	Lexicon + rules + ML (CRF, NB, ME)
6	The Ohio State University	0.8907	0.9261	0.9081	Lexicon + rules
7	TMUNSW	0.8594	0.9387	0.8973	Lexicon + rules + ML (CRF and NB)
8	National Central University	0.8586	0.9256	0.8909	unknown
9	UNIMAN	0.8557	0.9007	0.8776	Lexicon + rules
10	University of Utah	0.8552	0.8951	0.8747	Existing tools + regular expressions

Table 3

Significance tests for Risk Factor identification systems

	NLM	Harbin Grad	Kaiser	Linguamatics	Nottingham	Ohio	TMUNSW	NCU	UNIMAN	Utah
NLM										
Harbin Grad	F									
Kaiser	P	R								
Linguamatics	P	-	F,P,R							
Nottingham	-	R	F	F						
Ohio	P	-	P	-	P					
TMUNSW	-	R	R	R	-	-				
NCU	-	-	-	-	-	R	P			
UNIMAN	-	-	-	-	-	-	P	P		
Utah	-	-	-	-	-	-	P	P	F,P,R	

Table 4

Micro-averaged F1 for individual risk factor categories; highest scores for that category are bolded

	Overall	Hyperlip.	Obese	Med.	Hyperten.	CAD	Fam Hist.	Smoker	Diabetes
NLM	**0.9276**	0.9366	**0.9298**	**0.9307**	0.9555	0.8303	**0.9805**	0.8538	**0.9533**
Harbin-Grad	0.9268	**0.9491**	0.8757	0.9293	**0.9715**	0.8253	0.9681	0.8861	0.9291
Kaiser	0.9185	0.9375	0.9011	0.9126	0.9591	0.8284	0.9630	0.9045	0.9420
Linguamatics	0.9171	0.9236	0.8902	0.9138	0.9647	**0.8331**	0.9767	0.8441	0.9473
Nottingham	0.9156	0.9368	0.8961	0.9194	0.9455	0.7870	0.9572	**0.9162**	0.9228
Ohio	0.9081	0.9204	0.8801	0.9092	0.9630	0.7904	0.9630	0.8264	0.9369
TMUNSW	0.8973	0.8896	0.8620	0.8982	0.9567	0.7598	0.9630	0.8148	0.9406
NCU	0.8909	0.8896	0.8624	0.8982	0.9289	0.7536	0.9630	0.8148	0.9191
UNIMAN	0.8776	0.8669	0.8486	0.8900	0.9047	0.7363	0.9591	0.8538	0.8603
Utah	0.8747	0.8903	0.8581	0.8585	0.9477	0.7603	0.9494	0.8655	0.9256

Table 5

Smoking status in 2006 and 2014 corpora

Comparison of 2006 and 2014 for smoking status categories					
Distribution of smoking status categories			**Performance on smoking status categories**		
	% of 2006 corpus	**% of 2014 corpus**		**Top 2006 F1 scores**	**Top 2014 F1 scores**
Past	9.3	20.1	Past	0.67	.8869
Current	9.2	7.0	Current	0.82	.7143
Non-smoker (never smoked)	16.3	23.3	Non-smoker (never smoked)	0.94	.9237
Smoker (ever smoked)	2.4	.9	Smoker (ever smoked)	0.4	0
Unknown	62.7	47.1	Unknown	0.96	.9688
Unspecified (no consensus reached for gold standard)	-	1.6		-	-