MBI Capstone
2017

Can Natural Language Processing Improve Healthcare Quality by Early Identification of Coronary Artery Disease Risk Factors from Unstructured Clinical Data?

Raghav Mishra, M. D

Commonly used abbreviations

ACS- Acute Coronary Syndrome

ARRA- American Recovery and Reinvestment Act

BMI- Body Mass Index

CABG- Coronary Artery Bypass Grafting

CAD- Coronary Artery Disease

CPT- Current Procedural Terminology

CRF- Conditional Random Fields

DCT- Document Creation Time

DM- Diabetes Mellitus

EHR- Electronic Health Record

FN- False Negative

FP- False Positive

HITECH- Health Information Technology for Economic and Clinical Health

ICD- International Classification of Diseases

IOM- Institute of Medicine

LAD- Left Anterior Descending artery

ML- Machine Learning

NLP- Natural Language Processing

NPV- Negative Predictive Value

PDF- Portable Document Format

PPV- Positive Predictive Value

PRISMA- Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PTCA-Percutaneous transluminal coronary angioplasty

RF- Risk Factor

STEMI- ST Elevation Myocardial Infarction

SVM- Support Vector Machine

TG- Triglyceride

TN- True Negative

TP- True Positive

Background; the need for NLP to identify CAD risk factors in unstructured notes

The HITECH health information technology for economic and clinical health act was signed into law in 2009 as a part of the economic stimulus package formally known as American recovery and reinvestment (ARRA)(4). Drawing from the recommendations of IOM paper published in 1999 "To err is human", the primary focus of HIETCH act was to stimulate the adoption of Electronic Health Records (EHR) to promote patient safety(5). In addition to making clinical data available real time for a wide range of activities such as research, regulatory reporting and resource utilization, the transition of health records from paper to electronic medium also facilitated addition of key EHR functionalities such as clinical decision support for quality improvement.

Heart disease is the leading cause of deaths in the both men and women around the world(6). In the United States alone, 600,000 people die of heart disease every year(7). The annual cost of heart disease including medical costs and loss of productivity reportedly exceeds 109 billion dollars(8). Coronary Artery Disease (CAD) accounts for more than 60% of all incidents of adult heart disease(9). CAD can result in other chronic conditions such as heart failure, valve diseases, arrhythmia etc. (9). It has been shown that early detection and treatment of CAD risk factors such as smoking, hypertension, diabetes, dyslipidemia etc. can minimize the morbidity and mortality resulting from this condition (10).

By large, clinical information captured in the electronic health records fall under two broad categories; structured and unstructured. ICD codes, CPT codes, vital signs, laboratory results and electronic prescriptions are examples of structured data. This category of data has a high degree of organization and can be seamlessly integrated into various relational databases making them easily amenable for data querying, high scale computation and automation for other downstream clinical processes. However, structured data often lacks context and comprises only about 20% of the data in EHRs.

Conversely, unstructured health data is comprised of text-heavy, information rich clinical narratives such as patient history, their unique biopsychosocial details and provider rationale for clinical assessment/treatment are either typed or dictated by care providers in free text form(10). However, unstructured clinical data fields are not easily automatically processed. Thus, extracting accurate information from narrative notes, such as the risk factors for CAD as noted above, is a well-known challenge to clinicians and is typically obtained through laborious and time consuming manual review of the medical record.

Hence, there exists a tension between the way busy healthcare providers document clinical care and their need of wanting to re-use prior clinical data toward improving clinical decision-support for practice of evidence-based and personalized medicine.(9) While using structured data entry by clinicians would, in theory, make solving this problem easier, it is a valid argument that the clinical narrative best tells the patient's story and describes the provider's thought process. It is feared that critical information might be lost if providers switch from a narrative documentation format to documentation using simple drop-down options.

Natural language processing (NLP) has long been proposed as a solution to this problem. This paper explores the effectiveness of the domain of natural language processing tools for extracting information from unstructured clinical documents.

Introduction

The development of Coronary Artery Disease (CAD) or “heart disease” for short is complex, and many factors are involved in determining whether a patient is at risk. The World Health Organization defines “risk factors” as “any attribute, characteristic or exposure of an individual that increases the likelihood of developing a disease or injury”(11). Smoking, Diabetes Mellitus, Hypertension, family history of heart disease, dyslipidemia and obesity are among the most common risk factors for CAD (9). While assessing a profile for risk of heart disease, providers also need to know if the patient had prior cardiac disease and what kind of intervention was performed (example; heart attack 10 years ago with coronary bypass surgery and subsequent graft stent placement 3 years ago). Yet another piece of information in this assessment is the patient’s current and past medication history. While some of this information (medications) may be present in the EHR in structured format, most of it remains buried in clinical notes in the form of narratives, requiring manual chart review as outlined in the background section.

Natural language processing (NLP) is a rapidly evolving interdisciplinary field combining computer science, artificial intelligence and linguistics. NLP is concerned with the interactions between computers and human (natural) languages. It is often synonymously used with natural language understanding that is, enabling computers to derive meaning from natural language input. Earliest models of NLP in healthcare used for information extraction were based on syntactic and semantic rules(10). These models were used dating back up to 30 years in life science research and health science databases to mine texts and minimize false positives search results(10). It wasn't until 1990s when statistical (also called machine learning or stochastic or probabilistic) language processing became the dominant methodology in the field due to the availability of fast computing hardware and plentiful text in an electronic format.

This paper will explore how natural language processing tools have been used to extract information from clinical notes, specifically in terms of identifying CAD risk factors documented in EHRs. We will also identify the unique challenges to using NLP in healthcare, compare numerical analyses for effectiveness of various NLP techniques and models to understand their current strengths and limitations. Finally, this paper will conclude with possible future directions for this area of work.

Materials and methods

Data sources and search

We conducted a systematic review of studies undertaken between 01/01/1996 and 06/01/2017 using MEDLINE (PubMed and OVID). The following keywords were used*: natural language processing, NLP, cardiovascular diseases, cardiovascular system, cardiovascular physiological* *phenomena, quality in healthcare, vital statistics and prognosis.* Both keywords and MESH (Medical Subject Headings) terms were used. In addition, a pearl string strategy was employed using frequently cited reviews of NLP being used in electronic health records for identifying heart disease. With the assistance of Oregon Health and Science Library, all qualifying references were downloaded in XML format and subsequently uploaded into Covidence, a web based software prior to initiating study selection. We also used this platform to store full articles in portable document format (PDF) and to track the search results at the title review, abstract review, article inclusion/exclusion levels. Of note, we involved subject matter experts from both NLP and Cardiology to oversee our search strategy and ultimately approve our study design.

Study selection

Inclusion and exclusion criteria were framed prior to the implementation of the search strategy. To evaluate NLP outcomes in identifying adult patients with heart disease risk factors documented in EHR, we included studies based on the following criteria:

1) A medical corpus with commonly identified heart disease risk factors was used to evaluate the NLP model between 01/01/1996 and 06/01/2017.

2) NLP techniques was described in sufficient detail.

3) Numerical analyses were reported to report effectiveness of the NLP techniques used.

Non-English publications, studies done on animals, non-CAD cardiovascular diseases (heart failure, pulmonary embolism, pulmonary hypertension etc.) and pediatric studies were excluded. Eligible titles and abstracts were reviewed by two reviewers for inclusion through voting independently on Covidence; disagreements were resolved via telephone discussion. Subsequently, an examination of the full-length articles was carried with the intent of eliminating duplicate studies or studies which did not meet the criteria outlined above.

Data extraction and outcome measures

Two reviewers independently identified data elements to be extracted from the selected full-length articles and resolved differences through consensus. The following data elements were finalized for extraction; Study year and country, Study name, Population studied; specifically, the type of CAD risk factors, NLP technique/s used and outcomes; effectiveness of NLP models in terms of numerical analyses (precision, recall, F1).

Risk of Bias and Quality assessment criteria

We had originally intended to evaluate for quality using components of the RoBANS (Risk of Bias Assessment Tool for Nonrandomized Studies) scale. However, due to the inherent nature of informatics studies we found it difficult to apply this framework that is used to evaluate traditional clinical trials. We felt that almost all features of the RoBANS scale (the selection of participants, confounding variables, the measurement of exposure, the blinding of the outcome assessments, incomplete outcome data, and selective outcome reporting) did not apply to the included studies (where NLP models were deployed on a standard corpus provided by an external source) and hence this tool could not be appraised during our quality assessment.

However, two raters did independently determine the quality of the studies included through full text review and did not find any obvious methodological flaws.

Results

Based on initial search, 99 articles were obtained and reviewed independently by two reviewers. 81 articles were excluded based on title and abstract. A total of 18 potential studies were thus identified with our search strategy. 4 studies were further excluded, leaving 14 studies for final analysis (1-3, 9, 10, 12-20). The sequence describing the above process can be seen in Figure 1.

All except one study used NLP methods exclusively to extract CAD risk factor identifiers from unstructured notes. Liao et al (12), however, used both structured ICD codes and NLP techniques to identify CAD risk factor indicators; although in this study CAD risk factors were not as exhaustive as in other included studies.

As seen in Table 1, 10 out of 14 included studies used the micro-average method to report outcomes. Micro-averaged F1 was the primary metric. Micro-precision is ($TP_{\text{all sets}}$/ ($TP_{\text{all sets}}$ +$FP_{\text{all sets}}$)), Micro-recall is ($TP_{\text{all sets}}$ / ($TP_{\text{all sets}}$ +$FN_{\text{all sets}}$)) and Micro-F1 is simply the harmonic mean of Micro-precision and Micro-recall. However, 2 studies in this cohort did report their numerical outcomes and 1 study was a detailed description of the methodology of annotation process for machine learning with no outcomes to report. Liao et al, reported outcomes in terms of sensitivity, specificity, PPV and NPV.

Machine learning system by Roberts et al (3) using fine grained annotations and support vector machines was the top performer with a micro-precision of 89.51, micro-recall of 96.25 and micro-F1 of 92.76. It is worth noting that 8 out of 13 studies used either a machine learning NLP system or had a significant machine learning component (as a part of hybrid system). Further, 4 out of top 5 top performing systems were either machine learning systems or had a significant component of it. Similarly, Liao et al reported a high specificity and NPV and noted that including NLP into the CAD algorithm improved the sensitivity of the algorithm. (12)

Some aspects related to extracting information from clinical text proved harder than others. All number-based indicators (i.e., HbA1c, Glucose, Cholesterol, LDL, Blood pressure, and BMI measurements) have significantly lower F1s than "mentions". One 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. Yet another reason might be the sparsity of number related indicators were sparse in the training data (3, 20).

All "non-mention" CAD indicators (test, evaluation and symptom) also had comparatively low F1s in all the studies, including the top performing machine learning system listed in Table 1. Most likely this is due to the extreme variety of ways indicators specific to this risk factor were described in the corpus(3). For example, note the italicized terminologies below;

Event: "*s/p ant STEMI* + stent *LAD*", "*PTCA w/* Angioplasty to *LAD*"(3).

Test Result: “Stress (*3/88*): *rev*. *anterolateral* ischemia”, “normal ECG but a small *anteroseptal* zone of ischemia”(3).

Symptom: “occasional and very transient episodes of angina”, “*Since 11/19/2096* he has had complaints of increasing dyspnea on exertion and chest pain”(3).

While “mentions” of other risk factors such as Diabetes can also vary (for example, diabetes can be “diabetes mellitus”, “DM”, “DM1”, “IDDM”, “NIDDM”, “DMII”, “DM2”, “t2dm”, “MODY”, “Brittle DM” and so on), these phrases along with some basic polarity checking, is generally enough to identify positive diagnoses in the text. On the other hand, terminology for coronary artery related terminology could range from “ACS” to “probable inferior and old anteroseptal myocardial infarction”, “s/p MI in 4/80”, “quadruple bypass”, or “emergent cardiac catheterization” (20). Some argue that these terms can be identified with exhaustive tagging in rule/lexicon creation while other caution against such an approach for valid concerns of practicality. An exhaustive rule/lexicon can be difficult to maintain, hard to refine, has potential for duplicates and might overfit the data distribution.

Other unique elements to extract were risk factors such as smoking and family history, which do not have attributes. Rather, family history is categorized as a binary; either as “present” or “not present”; however, this ultimately made it easier for family history to be extracted.

On the other hand, patients’ smoking history was much more complicated. Smoking history can be classified into five possible categories; never smoker, past smoker, current smoker, history of smoking but current status unclear and finally, unknown where the patient was ever smoker or not. Since this is a complicated extraction to make, most systems used a dedicated smoking classifier.

“Mentions “of obesity also proved to be challenging for all systems again due to the variations in how this risk factor and its indicators can be documented in clinical notes. Weight as a numerical value either in kilograms or pounds, BMI as a number without units, in plain text form e.g., “patient is overweight” or simply the word “obese” under clinical exam are possibilities. To the same effect, training set probably did not have all the variations of obesity available for tagging which made it a low scoring item for most systems.

Lastly, it is to be noted that although only 5 out of 13 systems were declared rule-based by their authors, rules were in varying proportion, an integral part of all hybrid systems and even the most dedicated machine learning systems.

Discussion

Broadly speaking, there were three different NLP approaches to extracting information related to CAD risk from clinical corpuses; machine learning systems, rule based systems and hybrid systems that combined the two in varying proportion (16, 20). Regardless of the approach, as a part of the system architecture most NLP clinical information extraction models used pre-processing tools to identify CAD risk factor concepts to be extracted using medical lexicons (either created from the gold standard or curated rules from trusted resources like UMLS), the output undergoes computation and finally the candidate annotations undergo post processing including temporal (time) attribution is performed (3, 20).

After NLP models extract information about CAD risk factors from a patient's longitudinal corpus, they are categorized under one of three groups; with CAD, without CAD but predisposed for future CAD due to presence of risk factors and lastly, without CAD or related risk factors. To be able to determine this, however, the NLP model must not only identify the CAD risk factor in a patient's note but also assess the number and severity of risk factors, the relationship between several notes that might belong to the same patient (longitudinal patient records often consists of clinic notes, hospital discharge summaries, letters of communication among providers etc.) and lastly, form time attributes for whether the risk factors were present before, during, and after the record's creation date, frequently abbreviated as DCT (Document Creation Time) (1, 2, 16). Assessing the time attribute can be particularly challenging in a clinical set up since records are written after the clinical appointment with the patient. For example; "patient's BP was 160/80" in the notes is likely a reading taken prior to the meeting, even though the tense seems to suggest the reading was taken in a comparatively distant past (16).

Rule based systems

The use of rule based NLP systems in life science research dates back up to 30 years(21, 22). As outlined in Figure 3 a rule based NLP system pipeline consists of the NLP rule engine, pre-and post-processing tools and a temporal analyzer(2). Creating a rule based system starts with identifying a set of lexical concepts related to the topic of interest, CAD risk factors in this case, on a metathesaurus like UMLS. While each risk factor (example, dyslipidemia) can be easily identified by its “mention”, there could be other indicators associated with it (Table 2). Here, “mention” is a statement identifying the risk factor (e.g., “patient has dyslipidemia”). On the other hand, “TG 1279 mg/dl” is a “test” indicator for dyslipidemia and “pancreatitis due to hypertriglyceridemia” is an “event” indicator for dyslipidemia (1).

In building these lexical concepts, a graph-search module is developed to traverse the UMLS graph from the identified concept to its children, along an IS-A relationship, until a leaf node was reached. Since IS-A relationships could also connect concepts which are not of the same semantic type, this can induce false positives (e.g., gestational diabetes is not a risk factor for CAD but is related to the diabetes concept). To eliminate such false positives, only concepts with the same semantic type as its parent were retained during a search. (2, 10, 16)

Subsequently, using pre-processing tools the note undergoes text segmentation. Various note sections like headers, medication lists, and tests are identified. The body of the note containing the narrative is broken into sentence chunks, tokens, and parts of speech tags. The pre-processed material is then fed into the NLP engine which houses the concept based rules. It is worth mentioning that while querying medical notes with concept identification is an integral functionality for rule based systems it is not an effective tool when it comes to detecting irregular clinical documentation patterns like abbreviations and numbers (e.g., A1C 8 gm%). To solve this problem, regular expressions are added to the NLP engine.

Regular expressions have been popularly described as a key word search on anabolic steroids. These search expressions can be words, numbers or part of a sentence and are referred to as 'strings'. The method has been described in detail elsewhere, but the principle behind RegEx is simple: RegEx provides a set of search rules that one can apply to their match. For example, search "s(ei)?z" will allow matching of words that follow the pattern "seiz", which can be followed by all types of endings, matching "seizure, seizing, seized, seizes" and so on. Additional features within RegEx include negative look behind, which can be used to filter out negation expressions (e.g. the string "no seizures") in this example. Many studies have successfully used RegEx for identifying irregular expressions of a medical identifiers. (21)

In addition to the concept based extraction and using regular expressions, additional rules can be written into the NLP engine to identify common phrases used in clinical notes in the form of semi-frozen lexical expressions, syntactic chunks and semantic placeholders (e.g., "patient was diagnosed with" or "patient underwent *x* procedure")(10).

After text mining is complete, the NLP engine generates candidate annotations which undergo post-processing through negation and context detection filters to remove false positives. For example, negation filter identifies and nullifies sentences like “patient is not obese”. Context detection filters help identify if the experiencer of the risk indicators is the patient (e.g., brother has heart disease would get filtered out as family history). All annotations past this stage are considered final and are then sent to the temporal marker which assigns a time (before, during or after DCT) to each annotation. Temporal markers generally use the structure of the document (section headers) to identify the determine the time. Regular expression can be used to identify language based time indicators (e.g., “chest pain today”) (2, 10). Less frequently, temporal attributes can be based on annotation categories (disease, medication, measurement).

Outcomes for this NLP approach can be studied using a training set using numerical analyses such as recall, precision and their mean to make improvements the system over time. The rule engine gets feedback on quality of rule changes by examining differences in results between higher recall (and lower precision) patterns with higher precision (and lower recall) patterns. Since it is common practice to start by designing for high recall, many authors have cautioned toward limiting the number of queries to a minimum as they can become increasingly difficult to maintain, refine and avoid duplicates (2, 16). Others have also correctly identified that an unintended adverse effect of such practice could be overfitting(16). Instead, a practical solution would be to use one query for all CAD risk identifiers by detecting commonalities between parameters, also known as parameterizing the rule. Ultimately, the model also needs to be refined for precision. (16)

Given their longstanding presence, minimal domain expertise requirements and no large-scale engineering needs, rule based NLP systems remain a viable option for extracting information from medical notes for clinical and epidemiological studies. (10)

Machine learning systems

Machine learning systems made an entry in the NLP scene around the late 1990s when statistical (also called stochastic or probabilistic) language processing became the dominant methodology in the field due to the availability of fast computing hardware and plentiful text in an electronic format. As such, machine learning found early adopters in applications related to internet search. (22)

Statistical models, a part of machine learning systems, make soft, probabilistic decisions based on attaching real-valued weights to the features derived from the input data. In contrast to rule based systems, using a probabilistic model, the user does not set up the rules. The process involves providing a training set of documents labeled as positive or negative into a tool, which uses algorithms that allow a statistical (or probabilistic)-based method to distinguish positive cases from the negative ones. The model generated can then be applied to a set of unclassified or unlabeled documents, and the probability of each document to be positive (a case) is reported. (22)

It must be noted that machine learning approaches in the context of extracting information from clinical notes often need a degree of rule-based modules to be effective. Figure 4 outlines the system architecture of a machine learning approach. Blue borders indicate rule-based modules while red borders (support vector machines are ML counterparts of NLP rule engine in rule based systems) indicate machine learning-based modules.

Like rule based systems, the creation of machine learning system starts with identifying and creating appropriate lexical concepts for each CAD risk factor and its indicators. While this is a time-consuming exercise, machine learning systems using fine grained mention-level annotations have shown to outperform both rule based and hybrid NLP systems (3). To achieve this, unstructured notes are first processed with a collection of trigger lexicons targeting concepts related to CAD risk indicators, specifically the "mentions". Lexicons are built to identify both positive and negative annotations (and to further classify types of negative annotations detected). Efforts are also made to have consistent annotation boundaries and set an upper limit for annotation span as well as the maximum number annotations per document. Having more one than one annotator and an a priori inter-annotator conflict resolution agreement also helps avoid systemic errors. These processes significantly improve the performance of machine learning systems in comparison to coarse annotations(3). As mentioned earlier, rule-based modules are used in machine learning systems for extracting "non-mention" CAD risk indicators, e.g., family history, smoking history and test results.

After trigger extraction, candidate risk factors go through a series of support vector machine classifiers for detection of polarity (like negation detection) and validity (negative validity means a value detected in the annotated span with CAD risk indicator is not a valid association e.g., "A1C 19 days ago" ≠ A1C 19 gm%). In addition to above classifiers, support vector machines also include tools such as section classifiers (identify note type, lists within notes etc.), negation lexicons (e.g., "patient is *not* diabetic"), modality lexicons (e.g., "family history CAD *unknown,* was adopted") and context detection.

Finally, the time classifiers are applied with specified constraints and exceptions for assigning a time stamp for the information extracted from the document in relation to Document Creation Time (DCT).

A critical advantage of machine learning unlike rule based systems is it can afford to ignore the syntactic (parts of speech, dependencies etc.) as well semantic information (word senses, semantic roles and named entities) in the corpus and still achieve excellent recall and precision. (3). On multiple occasions it has been shown that the quality of the outcomes in machine learning systems depends on the quality of work put into annotating the training set.

Drawbacks of machine learning systems include the model becoming very sensitive to the distribution of data over time resulting in poor performance elsewhere. For example, ML model developed from institution A might not work as well for data from institution B. This is, in part, due to a lack of a corpus annotation standard across organizations. Preparing use cases of large annotated corpora conforming to a standard and limitation of data sharing in the domain can be very difficult if not unrealistic.

Compared to rule based systems, it has been observed that building machine learning systems requires additional expertise in computer sublanguages and statistics, skills which can be expensive to acquire.

Nevertheless, upon reviewing the included studies machine learning was noted to be very effective in its abilities to extract CAD risk factors from unstructured clinical notes.

Hybrid systems

Hybrid systems are simply a combination of rule based and machine learning systems in any variety of permutation and combinations. For instance, as noted earlier in the discussion, machine learning systems could themselves be classified under hybrid systems as they often tend to have components of rule based systems. Several popular open source NLP platforms like cTAKES (Clinical Text Analysis and Knowledge Extraction System), MedLEE (Medical Language Extraction and Encoding System) and caTIES (Cancer Tissue Information Extraction

System) are can also be classified under hybrid NLP systems as they use both machine learning and rule based components. In general, hybrid systems create rules by identifying and tagging risk factor indicators (using phrases, logic or discourse based tags), sometimes called "hot spotting"(13, 14, 17). These tags, either created manually or otherwise, are then used to train the system using machine learning engines such as vectors and conditional random field algorithms(13, 14, 17). The extracted annotations are subsequently fed into a variety of post processing tools including time attributers such as naïve Bayes for final annotations (9, 13). Unfortunately given the wide range of possibilities of how one can combine rule based and machine learning system components, an exhaustive review of hybrid systems is beyond the scope of this paper.

Conclusion

Our systematic review shows that natural language processing is very effective in extracting detailed information about risk factors for coronary disease from a variety of unstructured clinical notes, a majority constituent in a patient's longitudinal medical record. Rule based systems, machine learning systems as well as hybrids were all capable of performing above a micro-F1 score of 85.0 in the included studies. Trend suggests that machine learning systems are increasingly becoming popular in comparison with rule based systems. It is to be noted, however, that even the most advanced machine learning systems tend to benefit significantly from and hence often incorporate from components of rule based systems. Evidence suggests preparing high quality training data with lexicons, annotations and tags made the biggest difference in system performance. Difficulty in processing and extracting complicated medical concepts, irregular phrases, ad hoc abbreviations and certain types of numerical data suggests open questions that yet need to be addressed in future NLP research.

References

1. Stubbs A, Uzuner O. Annotating risk factors for heart disease in clinical narratives for diabetic patients. J Biomed Inform. 2015;58 Suppl:S78-91.
2. Shivade C, Malewadkar P, Fosler-Lussier E, Lai AM. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. J Biomed Inform. 2015;58 Suppl:S103-10.
3. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform. 2015;58 Suppl:S111-9.
4. (OCR) OfCR. HITECH Act Enforcement 2010 [Available from: https://www.hhs.gov/hipaa/for-professionals/special-topics/HITECH-act-enforcement-interim-final-rule/index.html.
5. Medicine Io. To Err is Human 1999 [Available from: http://www.nationalacademies.org/hmd/Reports/1999/To-Err-is-Human-Building-A-Safer-Health-System.aspx.
6. Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Blaha MJ, et al. Heart disease and stroke statistics--2014 update: a report from the American Heart Association. Circulation. 2014;129(3):e28-e292.
7. Sherry L. Murphy BSJX, M.D.; and Kenneth D. Kochanek, M.A. National Vital Statistics Reports US DEPARTMENT OF HEALTH AND HUMAN SERVICES, Centers for Disease Control and Prevention
8. Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD, et al. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation. 2011;123(8):933-44.
9. Chang NW, Dai HJ, Jonnagaddala J, Chen CW, Tsai RT, Hsu WL. A context-aware approach for progression tracking of medical concepts in electronic medical records. J Biomed Inform. 2015;58 Suppl:S150-7.
10. Karystianis G, Dehghan A, Kovacevic A, Keane JA, Nenadic G. Using local lexicalized rules to identify heart disease risk factors in clinical notes. J Biomed Inform. 2015;58 Suppl:S183-8.
11. Organization WH. WHO; risk factors. 2017.
12. Liao KP, Ananthakrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS, et al. Methods to Develop an Electronic Medical Record Phenotype Algorithm to Compare the Risk of Coronary Artery Disease across 3 Chronic Disease Cohorts. PLoS ONE. 2015;10(8):e0136651.
13. Chen Q, Li H, Tang B, Wang X, Liu X, Liu Z, et al. An automatic system to identify heart disease risk factors in clinical texts over time. J Biomed Inform. 2015;58 Suppl:S158-63.
14. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. J Biomed Inform. 2015;58 Suppl:S128-32.
15. Jonnagaddala J, Liaw ST, Ray P, Kumar M, Chang NW, Dai HJ. Coronary artery disease risk assessment from unstructured electronic health records using text mining. J Biomed Inform. 2015;58 Suppl:S203-10.

16. Cormack J, Nath C, Milward D, Raja K, Jonnalagadda SR. Agile text mining for the 2014 i2b2/UTHealth Cardiac risk factors challenge. J Biomed Inform. 2015;58 Suppl:S120-7.
17. Torii M, Fan JW, Yang WL, Lee T, Wiley MT, Zisook DS, et al. Risk factor detection for heart disease by applying text analytics in electronic medical records. J Biomed Inform. 2015;58 Suppl:S164-70.
18. Yang H, Garibaldi JM. A hybrid model for automatic identification of risk factors for heart disease. J Biomed Inform. 2015;58 Suppl:S171-82.
19. Urbain J. Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models. J Biomed Inform. 2015;58 Suppl:S143-9.
20. Stubbs A, Kotfila C, Xu H, Uzuner O. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. J Biomed Inform. 2015;58 Suppl:S67-77.
21. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011;18(2):181-6.
22. Kimia AA SG, Landschaft A, Harper MB. An Introduction to Natural Language Processing. How You Can Get More From Those Electronic Notes You Are Generating. Pediatric Emergency Care. 2015;Volume 31(Number 7):6.

Supplementary Materials

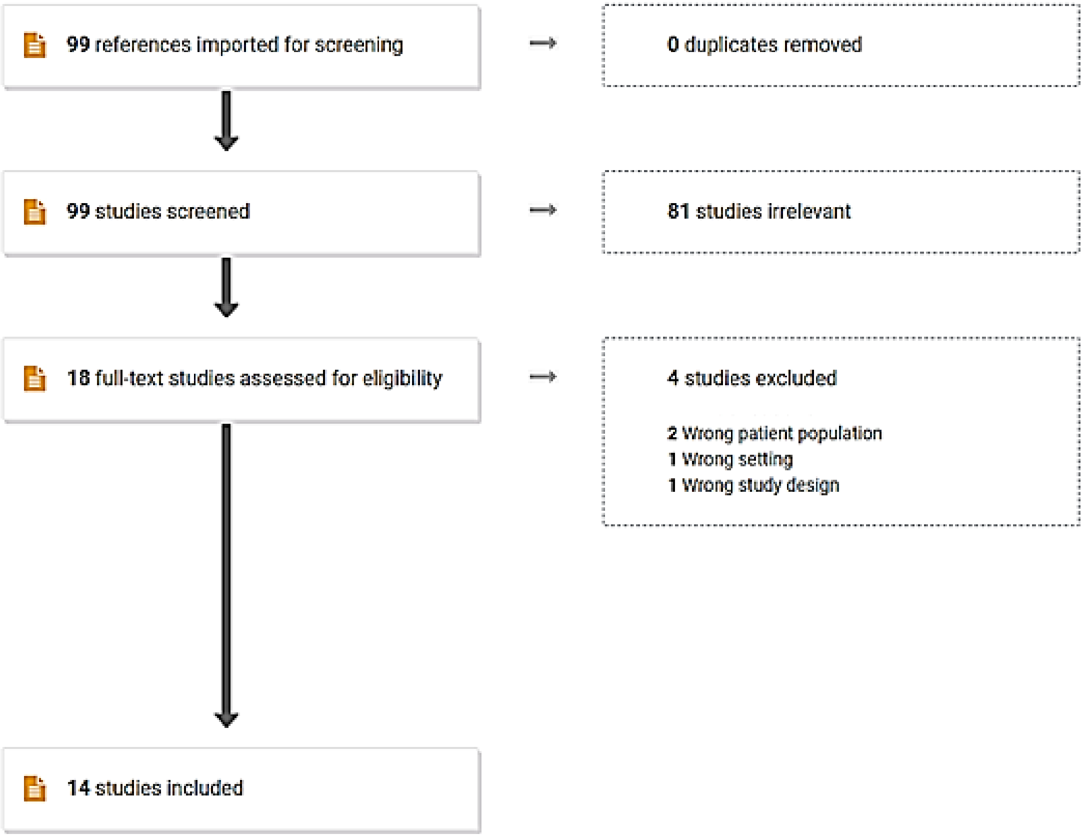
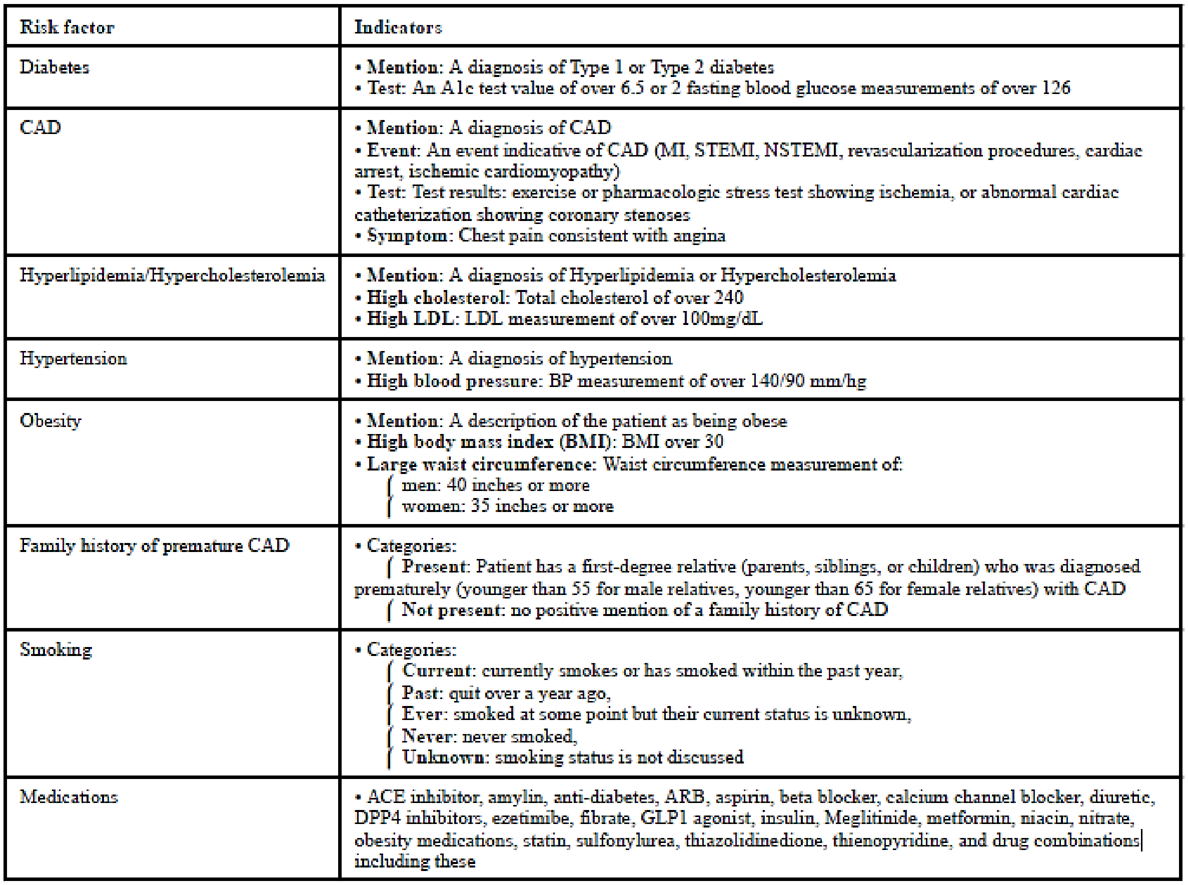


Figure 1. PRISMA checklist

Study/Year Country	NLP technique (key components)	Micro Precision	Micro Recall	Micro F1
Roberts/2015 USA	Machine Learning (fine grained annotations, SVM, rules)	89.51	96.25	92.76
Chen/2015 China	Hybrid (phrase, logic and discourse based tags)	91.06	94.36	92.68
Torii/2015 USA	Hybrid (tags, SVM, rules)	89.72	94.09	91.85
Cormack/2015 UK and USA	Rule based (text mining; Information Extraction Platform)	89.75	93.75	91.71
Yang/2015 UK	Hybrid (Conditional Random Field algortihm, rules)	88.47	94.88	91.56
Shivade/2015 USA	Rule based (NLP engine with UMLS concepts, Regular Expressions and rules)	89.07	92.61	90.81
Chang/2015 Taiwan	Hybrid (Context aware section classifiers, Conditional random field algorithm)	85.94	93.87	89.73
Karystianis/2015 UK	Rule based (Vocabularies, rules)	85.57	90.07	87.76
Khalifa/2015 USA	Hybrid (Textractor and caTAKES open source NLP platform)	85.52	89.51	87.47
Urbain/2015 USA	Rule based (entity recognition, Bayesian statistics, and rule logic)	80.00	88.70	84.10
Stubbs/2015 USA	Machine Learning	Not reported	Not reported	Not reported
Jonnagaddala/2015 Australia	Rule based	Not reported	Not reported	Not reported

Study/Year Country	Technique	Sensitivity	Specificity	PPV	NPV
Liao/2014 USA	Structured data (ICD codes) plus NLP (using Health Information Text Extraction (HITEX); a rule-based pipeline)	87	96.3	90	94.5

Table 1. Micro-precision, Micro-recall and Micro-F1 outcomes of various NLP systems identifying CAD risk indicators on unstructured clinical notes



Risk factor	Indicators
Diabetes	• **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
CAD	• **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
Hyperlipidemia/Hypercholesterolemia	• **Mention**: A diagnosis of Hyperlipidemia or Hypercholesterolemia • **High cholesterol**: Total cholesterol of over 240 • **High LDL**: LDL measurement of over 100mg/dL
Hypertension	• **Mention**: A diagnosis of hypertension • **High blood pressure**: BP measurement of over 140/90 mm/hg
Obesity	• **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
Family history of premature CAD	• Categories: { **Present**: 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 { **Not present**: no positive mention of a family history of CAD
Smoking	• Categories: { **Current**: currently smokes or has smoked within the past year, { **Past**: quit over a year ago, { **Ever**: smoked at some point but their current status is unknown, { **Never**: never smoked, { **Unknown**: smoking status is not discussed
Medications	• ACE inhibitor, amylin, anti-diabetes, ARB, aspirin, beta blocker, calcium channel blocker, diuretic, DPP4 inhibitors, ezetimibe, fibrate, GLP1 agonist, insulin, Meglitinide, metformin, niacin, nitrate, obesity medications, statin, sulfonylurea, thiazolidinedione, thienopyridine, and drug combinations including these

Table 2. Indicators for CAD risk factors (1)

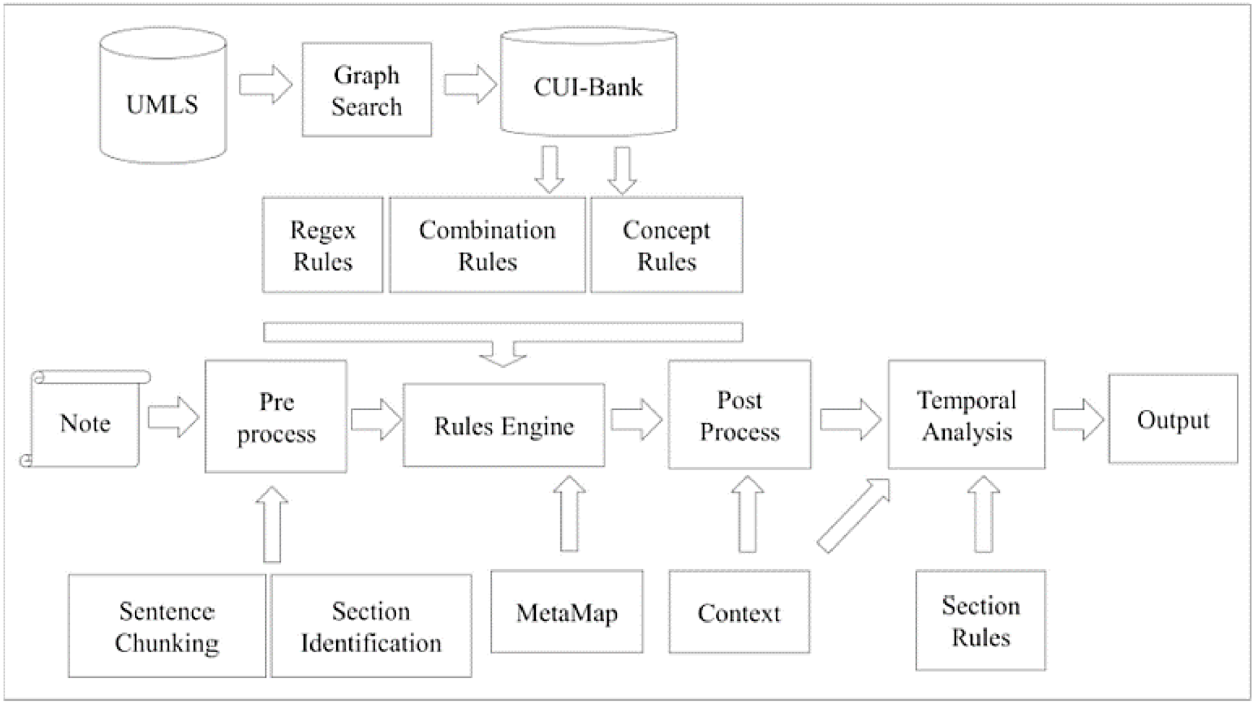


Figure 3. Rule based system architecture (2)

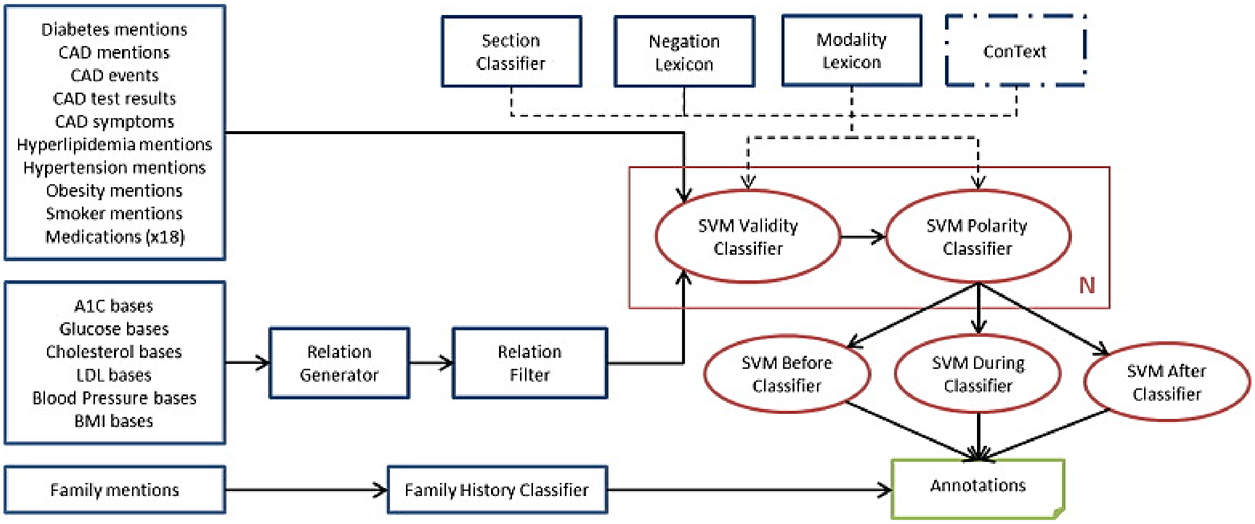
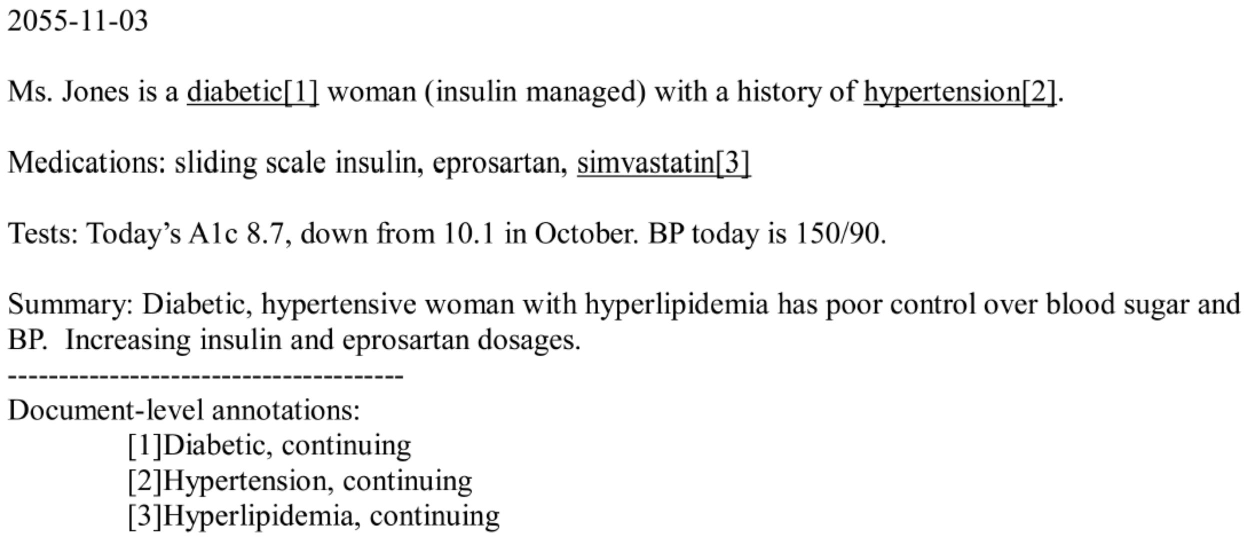


Figure 4. Machine learning system architecture (3)



2055-11-03

Ms. Jones is a diabetic[1] woman (insulin managed) with a history of hypertension[2].

Medications: sliding scale insulin, eprosartan, simvastatin[3]

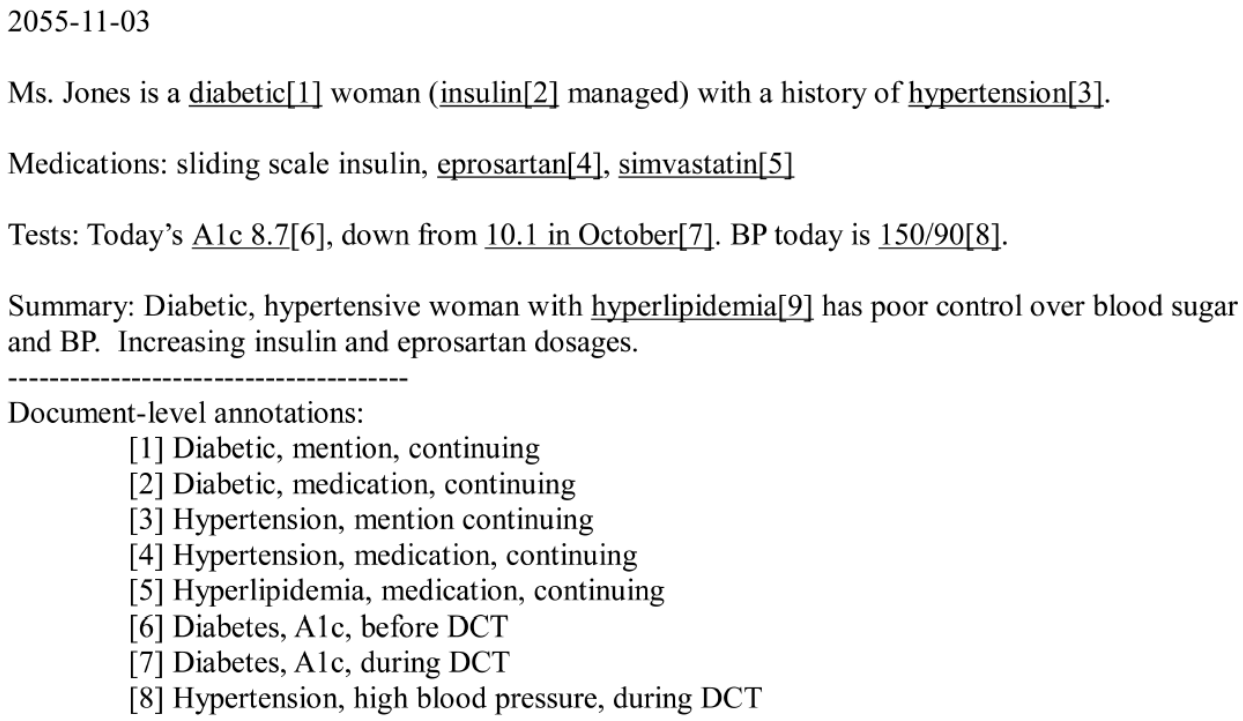
Tests: Today's A1c 8.7, down from 10.1 in October. BP today is 150/90.

Summary: Diabetic, hypertensive woman with hyperlipidemia has poor control over blood sugar and BP. Increasing insulin and eprosartan dosages.

Document-level annotations:

[1]Diabetic, continuing
[2]Hypertension, continuing
[3]Hyperlipidemia, continuing

Figure 5. An example of light annotation. Here, one risk factor is annotated per document(1).



2055-11-03

Ms. Jones is a diabetic[1] woman (insulin[2] managed) with a history of hypertension[3].

Medications: sliding scale insulin, eprosartan[4], simvastatin[5]

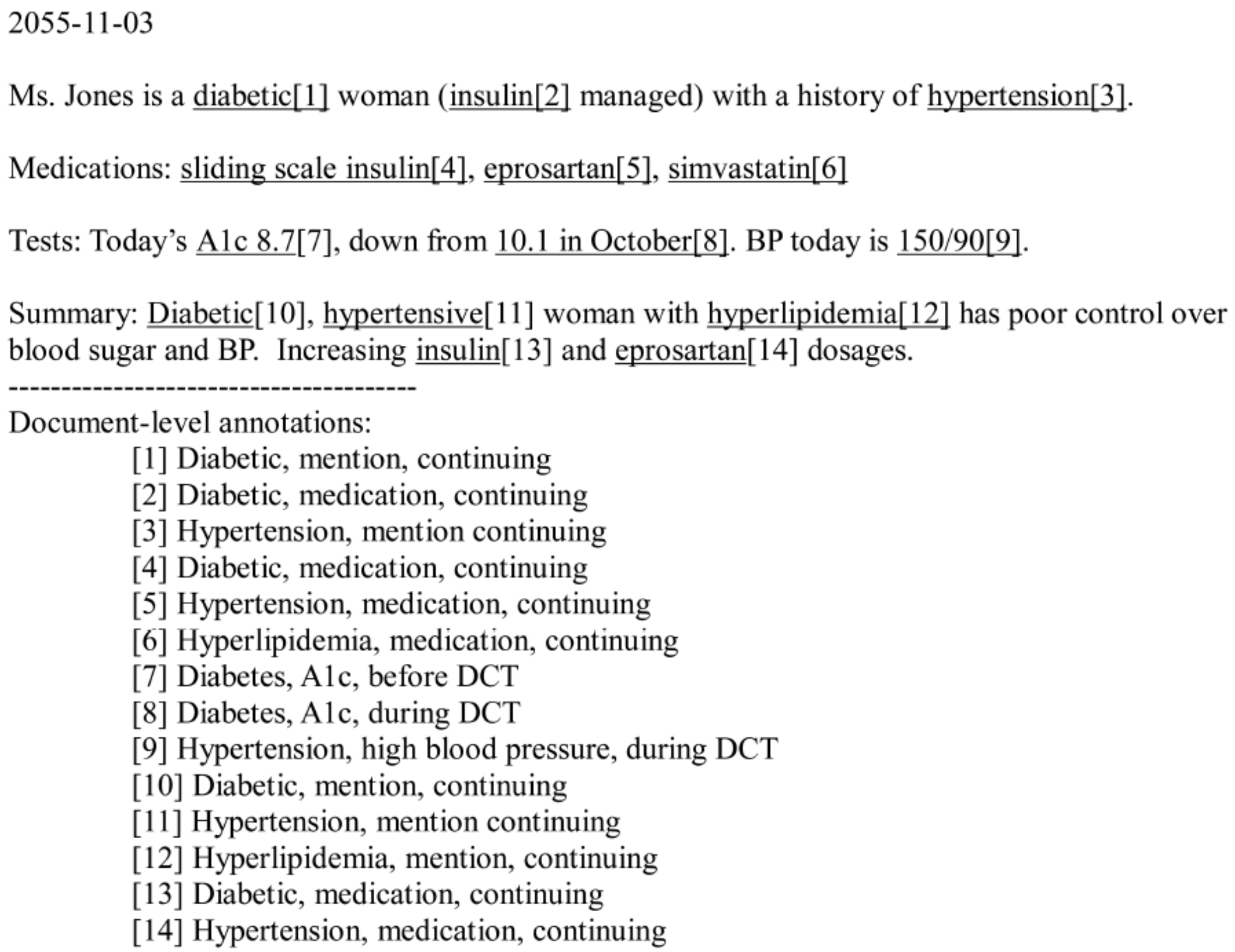
Tests: Today's A1c 8.7[6], down from 10.1 in October[7]. BP today is 150/90[8].

Summary: Diabetic, hypertensive woman with hyperlipidemia[9] has poor control over blood sugar and BP. Increasing insulin and eprosartan dosages.

Document-level annotations:

[1] Diabetic, mention, continuing
[2] Diabetic, medication, continuing
[3] Hypertension, mention continuing
[4] Hypertension, medication, continuing
[5] Hyperlipidemia, medication, continuing
[6] Diabetes, A1c, before DCT
[7] Diabetes, A1c, during DCT
[8] Hypertension, high blood pressure, during DCT

Figure 6. An example of moderate annotation. Here, one new indicator per risk factor is annotated in a document(1).



2055-11-03

Ms. Jones is a diabetic[1] woman (insulin[2] managed) with a history of hypertension[3].

Medications: sliding scale insulin[4], eprosartan[5], simvastatin[6]

Tests: Today's A1c 8.7[7], down from 10.1 in October[8]. BP today is 150/90[9].

Summary: Diabetic[10], hypertensive[11] woman with hyperlipidemia[12] has poor control over blood sugar and BP. Increasing insulin[13] and eprosartan[14] dosages.

Document-level annotations:

[1] Diabetic, mention, continuing
[2] Diabetic, medication, continuing
[3] Hypertension, mention continuing
[4] Diabetic, medication, continuing
[5] Hypertension, medication, continuing
[6] Hyperlipidemia, medication, continuing
[7] Diabetes, A1c, before DCT
[8] Diabetes, A1c, during DCT
[9] Hypertension, high blood pressure, during DCT
[10] Diabetic, mention, continuing
[11] Hypertension, mention continuing
[12] Hyperlipidemia, mention, continuing
[13] Diabetic, medication, continuing
[14] Hypertension, medication, continuing

Figure 7. An example of exhaustive annotation. Here, every indicator for a risk factor in the document is annotated, even if redundant(1).