# Identifying risk factors for heart disease in clinical texts

Hankyu Jang
University of Iowa
hankyu-jang@uiowa.edu

Vasiliy Ulin
University of Iowa
vasiliy-ulin@uiowa.edu

Alexandra White
University of Iowa
alexandra-white@uiowa.edu

## ABSTRACT

Heart disease is continuously the leading cause of death in the United States[1]. In this paper we aim to find a relationship between heart disease and risk factors in patient records. Informatics for Integrating Biology and the Bedside (i2b2)[2] is funded by the National Institutes of Health (NIH) to bring open-source data sets on a wide range of medical documents. In this paper the data set analyzed was i2b2 clinical text. The goal of this paper is to identify risk factors in patient records which have been indicated with heart disease. We analyzed 1304 XML files from the i2b2 training set. With this data we were able to extract eight risk factors to later scan for indicators. An analysis was conducted to find the relationship between the risk factors and the indicators of heart disease. The purpose is to find a correlation between indicators and risk factors with heart disease from patient records.

## KEYWORDS

i2b2, risk factor prediction, heart disease

**ACM Reference Format:**
Hankyu Jang, Vasiliy Ulin, and Alexandra White. . Identifying risk factors for heart disease in clinical texts. In . ACM, New York, NY, USA, 4 pages.

## 1 INTRODUCTION

For the purpose of this paper, heart disease is defined as a wide range of diseases that affect the heart[3]. The data set used to conduct the analysis is clinical text from patients that may experience a type of heart disease. The training data set which we used to generate an automated pipeline to annotate the documents includes records from 300 patients. From these records we further analyzed the eight risk factors from heart disease. These risk factors are: diabetes, coronary artery disease (CAD), hyperlipidemia, hypertension, obesity, family history, smoking status, and medication. From these eight risk factors the data provides insight to the indicators of the risk factors.

Throughout the research process there were a variety of challenges that we faced when identifying the risk factors from the clinical text. There were differences in the clinical text where it would either mention the risk factor directly or allude to the indication of a risk factor. This difference in text formatting lead us to create two different types of methods for extracting risk factors that are 1) annotating based on direct mention and 2) tagging based on rule-based approaches.

[1]See https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm
[2]www.i2b2.org/about/index.html
[3]https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118



We set out to build a system that addresses the i2b2 goals mentioned in the paper, "Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2"[7]. In this paper, the authors review a shared task of identifying risk factors for coronary artery disease from clinical text. There were twenty teams that participated in this project, the systems of the top six teams produced F1 scores over .90. This paper was important part of the research process because it allowed us to compare systems that are trying to accomplish a similar result. The results of the twenty teams' systems were quantified and ranked, which allowed us to see which systems produced the most preferred results. In order to build a system it is important to look at previous work in order to create a system with the least amount of limitations.

The system we created can be discussed in two major categories, mention tagging and rule-based tagging. Once the annotated text was accounted in the data structure the mention tagging portion was put in place. This section of the system identified any clinical text with a direct mention of a risk factor. The rule-based tagging portion of our system was more thorough in the sense that it followed a set of rules in order to annotate the text. This part of our system was developed to catch any text that was not previously annotated by mapping or mention tagging.

## 2 RELATED WORKS

Several teams have used different approaches to identify risk factors for heart disease from clinical texts. Roberts et al. [6], ranked first, created a system that used mention level classification. First, they used the annotations as a starting point for the mentions. After this process they expanded their system to annotate text by analyzing section headers, negated words, modality words, and output from ConText. In their design there was also a rule-based aspect. They would use rules to locate trigger words in lexicon, and then further analyze these trigger words. Finally, the last aspect of this rank one system was to account for time and use temporal attributes. This system was very successful with a FI value of .9276.

Chen et al. [2], ranked second in the i2b2 challenge, came up with a pipeline that utilizes both the rule-based approach and machine learning approach in identifying risk factors for heart disease from clinical texts. They differentiated the tags of the clinical texts into three categories: 1) phrase-based, 2) logic-based, and 3) discourse-based and then extracted tags based on the category. Their system extracts candidates by risk factor and indicator from the raw text and then determines its attribute 'time'.

The team to rank third in the i2b2 challenge was Torii et al.[8]. This team created a system with rules, SVM, and regular expressions. They divided the project in multiple text classification tasks. For more clarification they stated, "each combination of a tag and attributeâĂŞvalue pairs was regarded as an independent target category"[8]. The second portion of their system used an SVM, support vector machine, that had the option to be overridden by a

regular expression. This team created a system that produced an F1 score of .9185.

Cormack et al. [3] was the team that was ranked fourth in the i2b2 challenge. This team developed a system that consisted of I2E interface, lexicon, rules, and existing tools. The first section of their implementation was to use an existing text mining system, I2E. They used the GUI, graphical user interface, of the I2E system to develop queries in order to annotate. There was also a section of their system that used regular expressions to account for synonyms and abbreviations. The developed system created an F1 of .9171.

## 3 METHODS

### 3.1 Data

For the i2b2 risk factor annotation challenge, 1304 xml documents from 300 patients were provided in total. Fully annotated 790 documents with indications of risk factors for heart disease, from 180 patients, were provided as a training set. Figure 1 shows a portion of a document in the training set of a patient number 220, the first record of the patient. Remaining 514 documents from 120 patients were given without annotation, which means the xml document only contains the plain text in the <TEXT> tag and the rest was to be generated.

### 3.2 Mapping (Risk Factor - Indicator - Text)

Each xml file in the complete set contains tags of 8 risk factors and its corresponding attributes, containing text and indicator. The tag <CAD> in Figure 1 denote that 'CAD' is annotated from the direct mention in the following text 1) "known hx CAD", 2) "coronary artery disease", 3) " PLAVIX", or 4)"CAD". From these tags in the training documents, we generated mapping of 'risk factor' - 'indicator' - 'text' from all unique combination. Precisely, we generated the mapping using a Python dictionary where the 1) keys are risk factors for heart disease and 2) values are dictionaries with keys as indicator and values as a set of texts.

Table 1 depicts the overview of the mapping. From 488 unique texts, diabetes was annotated in documents. Among those, 191 texts directly mentioned diabetes, for instance, "Type 2 DM", "DIABETES", and "Diabetes type 2." Two hundred forty-four texts indicated diabetes by high A1c value of over 6.5, such as "HgA1c 6.6", "HgbA1c was in October and was 7.7", and "HGBA1C 9.30 (H)." Likewise, 53 texts indicated diabetes by high glucose measurement of over 126, such as "BG's of 400's", "GLU-POC 309", and "FS are in the 160-220 range."

**Table 1: Mapping of risk factor - indicator - text**

| Risk Factor | Indicator | Text |
|---|---|---|
| Diabetes | | 488 |
| | A1C | 244 |
| | mention | 191 |
| | glucose | 53 |
| CAD | | 921 |
| | mention | 113 |
| | event | 508 |
| | test | 199 |
| | symptom | 101 |
| Hyperlipidemia | | 129 |
| | mention | 61 |
| | high LDL | 51 |
| | high chol. | 17 |
| Hypertension | | 784 |
| | mention | 69 |
| | high bp | 715 |
| Obese | | 56 |
| | BMI | 22 |
| | mention | 34 |
| Family history | | 40 |
| | unknown | 15 |
| | current | 7 |
| | never | 16 |
| | past | 2 |
| Smoker | | 524 |
| | past | 261 |
| | never | 170 |
| | current | 78 |
| | ever | 15 |
| | unknown | 52 |
| Medication | | 1088 |
| | ACE inhibitor | 104 |
| | statin | 116 |
| | aspirin | 92 |
| | thienopyridine | 18 |
| | diuretic | 48 |
| | nitrate | 92 |
| | calcium channel blocker | 56 |
| | beta blocker | 156 |
| | ARB | 52 |
| | metformin | 43 |
| | sulfonylureas | 47 |
| | insulin | 223 |
| | fibrate | 15 |
| | ezetimibe | 5 |
| | niacin | 5 |
| | thiazolidinedione | 12 |
| | DPP4 inhibitors | 3 |
| | anti diabetes | 1 |

### 3.3 Mention Tagging

For any of the risk factors, if there are the exact match of texts in the mapping that directly mention any of the risk factors, the document is annotated with that risk factor. This matching rule using mentions holds for: diabetes, CAD, Hyperlipidemia, Hypertension, Obese, that have "mention" as the indicator and the medication.

### 3.4 Rule-Based Tagging

In addition to annotating the document with direct mention of the risk factor, we designed a rule-based tagging system. The system has a set of rules to annotate the document, based on the annotation guideline given as a pdf file from the i2b2 challenge. All the rules below are applied in each sentence of the plain text.

```
<?xml version='1.0' encoding='UTF-8'?>
<root>
  <TEXT><![CDATA[
Record date: 2067-05-03
Narrative History
   55 yo woman who presents for f/u
   Seen in Cardiac rehab locally last week and BP 170/80.  They called us and we increased her HCTZ to 25 mg from 12.5 mg.  States her BP's
were fine there since - 130-140/70-80.
   Saw Dr Oakley 4/5/67 - she was happy with results of ETT at Clarkfield.  To f/u 7/67.  No CP's since last admit.
(skip some lines…)
]]></TEXT>
  <TAGS>
    <MEDICATION id="DOC0" time="during DCT" type1="ACE inhibitor" type2="">
      <MEDICATION id="M0" start="1339" end="1346" text="ZESTRIL" time="during DCT" type1="ACE inhibitor" type2="" comment=""/>
      <MEDICATION id="M1" start="1339" end="1347" text="ZESTRIL " time="during DCT" type1="ACE inhibitor" type2="" comment=""/>
      <MEDICATION id="M2" start="1339" end="1359" text="ZESTRIL (LISINOPRIL)" time="during DCT" type1="ACE inhibitor" type2="" comment=""/>
    </MEDICATION>
(skip some lines…)
    <CAD id="DOC20" time="after DCT" indicator="mention">
      <CAD id="C0" start="838" end="850" text="known hx CAD" time="after DCT" indicator="mention" comment=""/>
      <CAD id="C1" start="977" end="1000" text="coronary artery disease" time="after DCT" indicator="mention" comment=""/>
      <CAD id="C2" start="1173" end="1180" text=" PLAVIX" time="after DCT" indicator="mention" comment=""/>
      <CAD id="C3" start="847" end="850" text="CAD" time="after DCT" indicator="mention" comment=""/>
      <CAD id="C4" start="977" end="1001" text="coronary artery disease " time="after DCT" indicator="mention" comment=""/>
      <CAD id="C5" start="1930" end="1933" text="CAD" time="after DCT" indicator="mention" comment=""/>
    </CAD>
(skip some lines…)
  </TAGS>
</root>
```

**Figure 1: A portion of a clinical text (220-01.xml)**

If a patient has a high A1c or high glucose measurement, diabetes is annotated in the document. To detect high A1c test value, we first look up for text [“A1C", “A1c", “a1c", "a1C"] in each sentence of the document. If any of the text is detected, then we extracted the first number that is either: 1) separated by spaces or 2) with decimal points that appear from the location of the match of the term. For instance, the system would extract 6.8 from the following text: “A1c 6.8 and glucose after eating 200.” If any number returned from the above procedure is within 6.5 (given in the annotation guidelines) and 15 (our choice), we annotated diabetes with indicator A1C. The reason for setting an upper bound for the A1C has two folds: 1) patient with highest A1C value in the training set was 14 and 2) to minimize false positives because our approach has a possibility of extracting a number that is not an A1c value.

To detect high glucose measurements, we used a similar approach. We looked up for text [“GLU", “BG", “FS", “FG", “glu", “Glu", “finger", “Finger"] in each sentence of the document which appear frequently in the texts in out mapping with indicator ‘glucose’. If any of the text is detected, then we extracted an integer in the sentence that is separated by spaces. If the number returned from the above procedure is within 126 (given in the annotation guidelines) and 500 (our choice), we annotated diabetes with indicator glucose. The reason for choosing this upper bound is the same as that of A1C explained above.

For the remaining risk factors, we applied similar rules as we have applied in diabetes. Risk factors with an indicator that needed to be compared with some number, such as: Hyperlipidemia, Hypertension, and Obesity, we followed the direction from the annotation guidelines.

To annotate Family history of premature CAD, we set a rule to look up for male first-degree relatives by searching for words in [“Father", “father", “Brother", “brother", “Son", “son"]. If a sentence contained male first-degree relatives and a word “CAD", we extracted a number and checked if the number is less than 55. Similar logic was applied to female first-degree relatives, and in this case, check if the number is less than 65. These numbers that we compared with to generate rules are based on the annotation guidelines.

Annotating smoker status was troublesome since the document should be annotated with one of the five categories, such as [“current", “past", “ever", “never", “unknown"]. Hence, we first annotated the document based on the condition in the annotation guideline. If there were no annotations made, we annotate the document as smoking: “unknown." Otherwise, if there is more than one annotation made for the smoker, we removed the annotations in the following order: 1) “ever", 2) “never", 3) “current." The idea behind this logic is that a patient could have “ever" smoked and is “current" smoker or “past" smoker. Unless there is evidence of the patient being “past" smoker, the patient is assumed to be current smoker. The patient may have “never" been a smoker but could have been annotated as “ever" smoker due to the mention of “smoke" related words in the clinical text. We set a rule of annotating the smoker based on these observations.

## 4 RESULTS

### 4.1 Training Set Evaluation

In order to use the evaluation script that was provided from the i2b2 challenge, we re-named the tags in the gold set with time attribute as “before DCT". Table 2 shows the result of running the evaluation script on the result of the annotation on the training set using the previously mentioned tagging method.

Our evaluation was done using a provided script which produced the following results: Recall (R), Precision (P) and F1 scores for both macro-precision and micro-precision. F1 being the main focus and the primary result we were interested in. We could further use the results shown in table 2 to identify the results with highest F1 score.

Risk factor with the highest F1 score we identified is Family History with an F1 score of 96 percent in both macro-/micro-precision. Various other categories of risk factors for a heart disease produced fairly high F1 scores, among them, are: Medications at 72.3 percent macro F1 and 80.7 percent micro F1, Hypertension at 70.6 percent macro F1 and 88.3 percent micro F1 score. The overall performance of our system on the training set had an F1 score of around 76 percent.

As we compare the precision and recall of the risk factors, there was a general trend of recall being higher than precision. One way to interpret this phenomenon is that our system tends to annotate a document with more risk factors than the actual existence of the risk factors. Since the system is annotating more risk factors than the actual number of risk factors, it tends to annotate a real risk factor correctly. However, due to the high number of false positives, the precision score tend to get lower.

**Table 2: Risk factor annotation results on training documents**

| Category | Macro | | | Micro | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| Diabetes | 0.583 | 0.665 | 0.622 | 0.664 | 0.969 | 0.788 |
| CAD | 0.258 | 0.396 | 0.312 | 0.295 | 0.881 | 0.443 |
| Hyperlipidemia | 0.422 | 0.429 | 0.425 | 0.630 | 0.940 | 0.754 |
| Hypertension | 0.706 | 0.706 | 0.706 | 0.851 | 0.918 | 0.883 |
| Obesity | 0.187 | 0.188 | 0.188 | 0.771 | 0.994 | 0.869 |
| Family History | 0.961 | 0.961 | 0.961 | 0.961 | 0.961 | 0.961 |
| Smoker | 0.665 | 0.665 | 0.665 | 0.734 | 0.681 | 0.707 |
| Medication | 0.614 | 0.880 | 0.723 | 0.679 | 0.996 | 0.807 |
| All | 0.633 | 0.932 | 0.754 | 0.648 | 0.935 | 0.765 |

## 4.2 Test Set Evaluation

Our system generated tags for the test set, however, we were not able to evaluate results since the test documents did not have ground truth annotations.

# 5 LIMITATIONS

There are several limitations in our system. In our approach to mention tagging, we did not consider negation information that resides in the text. This may have led to increasing false positives, meaning that our system may annotate risk factors that may contain the direct mention but with some presence of negative words that negate the meaning of the sentence. To handle this phenomenon, a possible solution would be incorporating negative word list from NegEx [1] to prevent the above false positives.

The evaluation result of training set shows that our system did the worst in annotating medication. The reason for the poor performance for this category is because we treated the indicator as mentions and looked up for direct matches only. If we had utilized disease-drug dictionaries, maybe from DrugBank [5], to extract drug mentions in documents, it would have enhanced the performance of our system on the medication category.

Our system is focused on rule-based tagging where the rules are from the given annotation guideline and our decision choices. Our lack of medical knowledge when choosing the upper limit threshold of the number to be extracted may have led to increase in false negatives. The thresholds we used could be adjusted with more realistic numbers, and the rules we made in the system could be modified for the better result.

Last but not least, we used the clinical text as is, without making any modifications. If we have pre-processed the documents to extract information such as negation words or section headers using a tool such as Context [4] and incorporated the result in our system, it would have increased the performance of our system.

# 6 CONCLUSION

We developed a pipeline that automatically annotates risk factors for heart disease from clinical texts. Our system annotates risk factors based on direct mentions or rule-based tagging system. Although the performance of our system is not comparable to that of the i2b2 top ranking teams, the system works in some level in annotating risk factors for heart disease in clinical texts.

# REFERENCES


[1] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics* 34, 5 (2001), 301–310.

[2] Qingcai Chen, Haodi Li, Buzhou Tang, Xiaolong Wang, Xin Liu, Zengjian Liu, Shu Liu, Weida Wang, Qiwen Deng, Suisong Zhu, et al. 2015. An automatic system to identify heart disease risk factors in clinical texts over time. *Journal of biomedical informatics* 58 (2015), S158–S163.

[3] James Cormack, Chinmoy Nath, David Milward, Kalpana Raja, and Siddhartha R Jonnalagadda. 2015. Agile text mining for the 2014 i2b2/UTHealth Cardiac risk factors challenge. *Journal of biomedical informatics* 58 (2015), S120–S127.

[4] Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. 2009. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics* 42, 5 (2009), 839–851.

[5] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al. 2013. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* 42, D1 (2013), D1091–D1097.

[6] Kirk Roberts, Sonya E Shooshan, Laritza Rodriguez, Swapna Abhyankar, Halil Kilicoglu, and Dina Demner-Fushman. 2015. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. *Journal of biomedical informatics* 58 (2015), S111–S119.

[7] Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of biomedical informatics* 58 (2015), S67–S77.

[8] Manabu Torii, Jung-wei Fan, Wei-li Yang, Theodore Lee, Matthew T Wiley, Daniel S Zisook, and Yang Huang. 2015. Risk factor detection for heart disease by applying text analytics in electronic medical records. *Journal of biomedical informatics* 58 (2015), S164–S170.