Coding Free-Text Chief Complaints from a Health Information Exchange: A Preliminary Study

Sotiris Karagounis[1], Indra Neil Sarkar, PhD, MLIS[1,2], Elizabeth S. Chen, PhD[1]
[1]Center for Biomedical Informatics, Brown University, Providence, Rhode Island
[2]Rhode Island Quality Institute, Providence, Rhode Island

Abstract

Chief complaints are important textual data that can serve to enrich diagnosis and symptom data in electronic health record (EHR) systems. In this study, a method is presented to preprocess chief complaints and assign corresponding ICD-10-CM codes using the MetaMap natural language processing (NLP) system and Unified Medical Language System (UMLS) Metathesaurus. An exploratory analysis was conducted using a set of 7,942 unique chief complaints from the statewide health information exchange containing EHR data from hospitals across Rhode Island. An evaluation of the proposed method was then performed using a set of 123,086 chief complaints with corresponding ICD-10-CM encounter diagnoses. With 87.82% of MetaMap-extracted concepts correctly assigned, the preliminary findings support the potential use of the method explored in this study for improving upon existing NLP techniques for enabling use of data captured within chief complaints to support clinical care, research, and public health surveillance.

Introduction

Electronic Health Records (EHRs) are composed of many different types of data that can be categorized as structured (e.g., diagnoses) and unstructured data (e.g., clinical notes). When performing analyses that involve EHRs, structured data provide the flexibility to manipulate the data for various uses such as statistical analyses or machine learning. An issue arises when data such as a diagnosis or an adverse reaction to a drug of a patient are only noted in narrative form, making embedded information harder to incorporate in analysis or use. A commonly used approach to address this challenge is to extract the information from unstructured data using techniques such as natural language processing (NLP)[1].

Chief complaints consist of patient reported symptoms, conditions, or diagnoses during their admission into a clinical care setting (e.g., the emergency department[2] [ED]). As one of the first instances of recorded data when patients enter a clinical setting, the free text can contain valuable information that could enrich EHRs. Extracting structured symptom data from chief complaints can help make ED visits more efficient, monitor disease outbreaks[3], and further improve personalized medicine[4]. Chief complaints tend to be inconsistent and challenging to organize or analyze. The free text format gives the flexibility to the writer of the chief complaint to be as expressive or concise as needed for a patient's symptoms on a case-by-case basis. However, the flexibility given by the format also comes at the cost of ambiguity of meaning of the complaint, errors and potential lack of information. The ambiguity of meaning can arise from the use of words or abbreviations in the text that can be interpreted in multiple ways, which makes assigning a correct concept to a certain phrase more challenging. Errors in chief complaints pertain to spelling mistakes, incorrect punctuation or incorrect symptoms described. Lastly, some chief complaints simply lack the necessary information to be able to extract concepts. An example of such a phrase would be "LAB WORK", where the phrase gives the piece of information that lab work was done after the patient was admitted; however, it lacks the context of the admittee's symptoms and thus would not be mapped to a medical concept for further analysis. Some of these characteristics of chief complaint data can be remedied through NLP tools and pre-processing techniques.

Extensive research has been done in the field of using structured EHR data for applications and the usefulness of such data in medicine cannot be understated. Structured symptom data extracted from chief complaints could better support clinical, quality, research and public health needs. Using natural language processing (NLP), particularly information extraction techniques, to gather structured data from free text is a difficult task and an active area of research[4,5]. Existing methodologies and tools for information extraction can be used in a plethora of different EHR settings including chief complaints to extract medical concepts. Several studies[3,5-8] have demonstrated the potential for information extraction from chief complaints, using either a heuristics-based or a supervised learning approach. A recent study[5] resulted in an algorithm, CCMapper, which used a bag of words approach on data collected from the Mayo Clinic ED. The algorithm performed with a sensitivity of 94.2% and a specificity 99.8% in mapping the chief

complaints from the specific hospital to symptom categories created by expert knowledge. The approach involves manually assigning words with categories and when free text is inputted into the algorithm, it chooses the category which has the most mappings to words in the text.

Approaches, such as the bag of words method differ from concept-based approaches for information extraction, as they rely on training data or the construction of specific keyword mappings to structured data. MetaMap[9] is a NLP tool created at the National Library of Medicine to map biomedical texts to medical concepts. The medical concepts used by MetaMap can be found in the Unified Medical Language System (UMLS) Metathesaurus[10], a large dataset that maps medical concepts to various medical vocabularies including diagnosis codes (e.g., ICD-10-CM). Since the MetaMap tool has pre-existing mappings, identifying concept mappings is streamlined[11] and there is no need for training or example data. The main challenge with working with the MetaMap tool involves attempting to pre-process messy text data such that it most accurately maps a phrase to UMLS concepts.

The overall goal of this paper was to demonstrate the use of a method to map chief complaint free text data to structured coded data. The method involves manipulating the text data, using MetaMap to map the phrases to UMLS concepts and hence using the UMLS Metathesaurus to map the concepts to ICD codes and categories.

Methods

Figure 1 depicts the approach for extracting diagnosis codes from free-text chief complaints in four main steps: (1) preprocessing the data from its raw form, (2) running the processed phrases through MetaMap, (3) post-processing the outputted concepts and (4) evaluating the outcomes. Each step of the pipeline will be discussed in detail in their respective sections. Code for processing and evaluation can be found in the supplemental GitHub repository: https://github.com/bcbi/chief-complaint-coder.

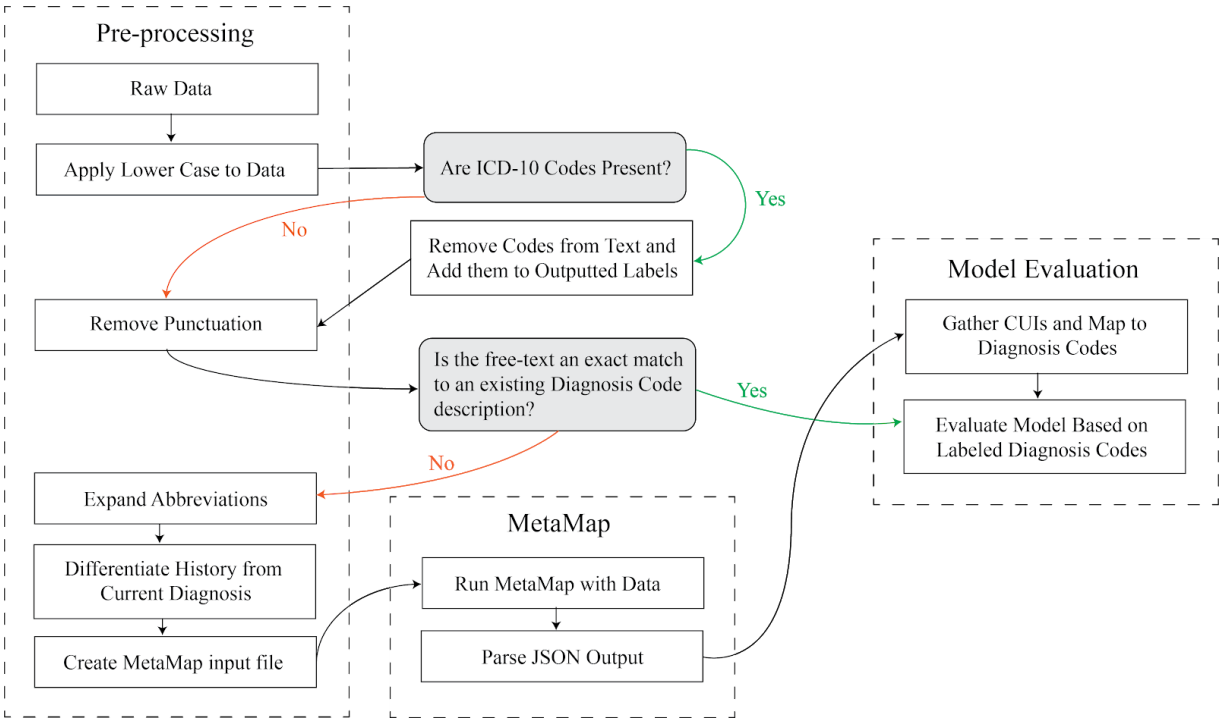

Figure 1. The above diagram shows the mapping process from chief complaint text phrases to ICD-10-CM codes.

Data Source

Founded in 2001, the Rhode Island Quality Institute (RIQI) is a 501(c)(3) nonprofit center that serves as the Regional Health Information Organization for the State of Rhode Island. RIQI operates CurrentCare, the statewide Health Information Exchange (HIE), which collects, aggregates, normalizes, stores, and makes accessible longitudinal electronic health record (EHR) data from >500 sources for more than 50% (n>550,000) of the state's population. For this study, chief complaints of Rhode Island patients enrolled in CurrentCare were obtained. The

advantage of using data provided through the HIE was that it allowed the data to be as close to real-world conditions as possible. The types of chief complaints vary substantially, as the data are gathered from multiple hospitals with different standards on how chief complaints are entered. The variability of the free text from multiple hospitals leads to the data's generality, as it prevents evaluation metrics and pre-processing methods from not generalizing to data from other hospitals.

Two deidentified datasets containing samples of chief complaints (for one year and one month) were generated and used in this study. The first dataset (n = 7,942) contains the unique chief complaints along with frequencies spanning a full year from January 1, 2018 to December 31, 2018. This "exploratory dataset" was primarily used to do a semantic type analysis and better understand the distribution of concepts in the text data. The second dataset (n = 123,086) spans the month of December 2018 and provides chief complaints with their assigned ICD-10-CM codes during the patient's admittance. Each chief complaint in the second dataset may have multiple codes, as shown in Table 1. This "evaluation dataset" was primarily used to assess the performance of the method.

The different standards of hospitals[12] as mentioned previously leads to the chief complaints taking many forms. Several chief complaints have short descriptions describing symptoms, which have the potential to be mapped to UMLS concepts. Other complaints contain a list of diagnosis codes or diagnosis code descriptions, which can be extracted and used as an output label. Several examples of chief complaint phrases in the dataset can be found in Table 1.

Table 1. Example chief complaints with coded diagnoses from the Evaluation Dataset.

ID	Chief Complaint	ICD-10-CM Code	ICD-10-CM Description
1	"LAB WORK"	D64.9	"Anemia, unspecified"
1	"LAB WORK"	R53.83	"Fatigue, unspecified type"
2	"rlq pain"	R10.31	"Right lower quadrant pain"
3	"Hypertension, unspecified type"	I10	"Hypertension, unspecified type"
4	"R50.9"	R50.9	"Fever, unspecified"

The various types of chief complaints influence decisions in the pre-processing and post-processing stages of the method. Each case mentioned above involves different techniques to clarify the concepts from the phrase and thus can be optimally mapped to UMLS concepts through MetaMap.

Pre-Processing

The preprocessing phase, which was implemented using the Julia general purpose programming language, attempts to sort through phrases that have additional labels, phrases that can be directly assigned a label and finally to create the MetaMap input file from the remaining phrases.

To begin with, all the phrases are set to lowercase allowing for proper text comparisons. Then using a regular expression, all ICD-10-CM codes are extracted from the phrase. If the phrase includes ICD-10-CM codes, then they are stored in the set of output labels and removed from the string. An example of this transformation can be seen below, where a tuple represents a phrase and the corresponding pool of predicted labels:

`(Phrase: "`R50.9, rlq pain`",Labels = []) -> (Phrase: "`, rlq pain`",Labels = [`R50.9`])`

The remaining portion of the string without the codes is left in the pool of strings that had no ICD codes in them. This is exemplified above, where ", rlq pain" should be assigned a different ICD code.

The next step in preprocessing is to remove punctuation with a regular expression. The ICD code matching is done before this step to make sure the matching is done correctly. Then a second comparison happens to see whether the

free text has an exact description match with ICD code description. If so, the phrase is assigned its corresponding ICD label and then sent to the pool of phrases ready for evaluation, as there is no other definitive information that the phrase can contain outside of the code description.

Since, some chief complaints have been matched to ICD-10-CM codes and descriptions, the next step in processing the data follows expanding abbreviations. Abbreviations can cause issues for MetaMap, as they add ambiguity to the phrase and make it difficult to map to a concept. Hence, a mapping of 102 prominent abbreviations was created based on the full year dataset to map the medical abbreviation to an expanded form. An example of the transformation that occurs can be seen below:

`(Phrase: "hx hf",Labels = []) -> (Phrase: "history of heart failure", Labels = [])`

The above example is used as it is characteristic of both mapping a medical abbreviation to its expanded concept, as well as, attempting to structure the phrase such that MetaMap can map the phrase to a concept that encompasses the temporal aspect of the symptoms. The distinction between a history of a condition and current condition is important, as the current conditions lead to an ICD-10-CM code of the patient in their current state; while historical symptoms should be kept in a different pool as they pose other uses. The process of manipulating the string is a heuristic that finds whether "hx" or "history" exists in a string and edits the string for the proposition "of " in order to make the part of the phrase semantically clear for mapping. Lastly, with a preprocessed pool of phrases, a text file is created containing IDs and phrases in order to input the text data to MetaMap.

Mapping Phrases to Concepts

Mapping parts of phrases to ICD codes can be facilitated through the National Library of Medicine's tool, MetaMap. The tool processes texts and identifies parts of phrases that are similar to UMLS concepts. This method of assigning ICD-10-CM codes or categories to phrases was chosen over other supervised learning techniques, as the nature of the problem of extracting diagnoses would lead to too many different labels for the amount of data at hand. Hence, the approach of using a heuristic tool, such as MetaMap, was preferable as it could map to specific codes or categories without needing copious data.

MetaMap is highly configurable and hence several options were used for post-processing. The command and arguments used for this study were:

```
metamap --JSONn --sldiID input.txt
```

`JSONn` specifies an output of an unformatted JSON file. Output files tend to be very large as the tool produces many candidate mappings for each phrase. The approach used for this study was only designed for use on smaller chief complaint datasets. `sldiID` allows for multiple phrases to be inputted via text file.

The processed JSON files were then subject to postprocessing written in Julia. The postprocessor extracted information that MetaMap produced from the JSON file. The first step in the process is extracting several key features of the phrases such as ID, the phrase itself and the score that MetaMap gave to the mapping. Then the top scoring candidates are chosen and stored with their corresponding concept name, Concept Unique Identifier (CUI), semantic type and a boolean of whether the phrase was negated. These data are sufficient to attempt to extract ICD-10-CM data from the phrases, as well as, perform a semantic type analysis on the data. The phrases and their corresponding data are then stored into Julia structs for further extendibility.

CUIs are unique identifiers for concepts in the UMLS Metathesaurus. When MetaMap identifies a concept, which matches the symptoms of a diagnosis code, it adds the corresponding CUI to the candidate mapping set. Mappings from UMLS CUIs to ICD-10-CM codes were obtained using the UMLS Metathesaurus. Since a phrase can describe many symptoms, there can be multiple mapped CUIs that correspond to diagnosis codes.

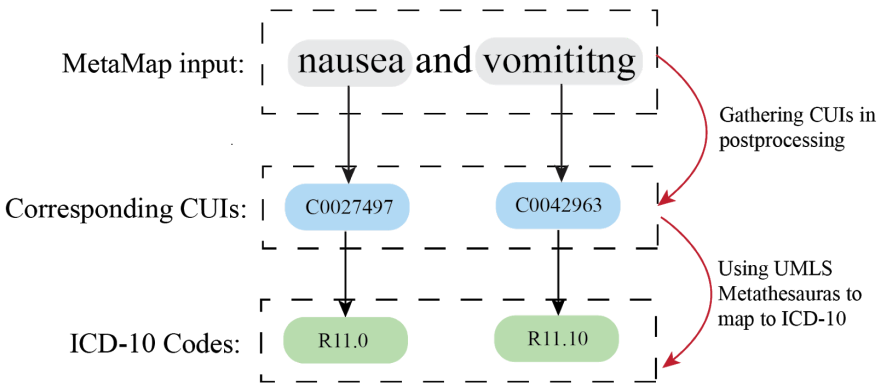

Figure 2. Example of a chief complaint processed by MetaMap and mapped to multiple diagnosis codes using the UMLS Metathesaurus.

Evaluation

Since the chief complaints in the evaluation dataset can have multiple labeled ICD-10-CM codes, as well as multiple predicted labels, a process to evaluate the method needed to be created. Since many chief complaints simply do not have enough information to capture all the labeled diagnoses, the main focus of the evaluation is whether the method can map to at least one labeled code.

Table 2. Example where the chief complaint phrase does not encompass all the information from the labeled diagnosis codes.

ID	Chief Complaint	ICD-10-CM Code	ICD-10-CM Description
255	"rlq pain"	R10.31	"Right lower quadrant pain"
255	"rlq pain"	R50.9	"Fever, unspecified"
255	"rlq pain"	R11.0	"Nausea"

Symptoms of nausea and fever are a usual occurrence with right lower quadrant pain; however, there is not enough information in the example phrase above to infer all of the labels. The method will be considered to have performed well on the phrase if it is able to map to the diagnosis code that the phrase encompasses: "Right lower quadrant pain".

Hence, a scoring function was constructed as follows:

$$score(p) = \begin{cases} 1, & \text{at least one map}(p) \in L(p) \\ 0, & \text{otherwise} \end{cases}$$

Where p is a given phrase, map(p) is the set of matched ICD-10-CM codes and $L(p)$ is the set of labeled ICD-10-CM codes corresponding to the phrase p.

An important note about the dataset used for evaluation is that the ICD-10-CM labels tend to be noisy. ICD codes can be too general or too specific for the information conveyed in a phrase. This could be attributed to the different coding standards of the various hospitals from which the data was gathered. For example if "postprocedural fever" is given as a phrase and the label for the given phrase is "R50.9", "Fever, unspecified", it could be difficult to find the correct mapping to the label's ICD code, as MetaMap may have mapped the phrase to the more specific ICD code for "R50.82", "Postprocedural fever." The opposite applies if the phrase does not contain enough information, while the labeled ICD code is more specific.

Two versions of the method were evaluated. One data pipeline attempted to predict the ICD code in full, which for the reason mentioned above poses a difficult task provided due to the noisy labels, as well as the potential lack of information conveyed in a phrase. Another proposed pipeline attempted to find the ICD category (e.g., "R50") rather than a specific code, as it would offer better mappings, while also being more granular than classifications buckets in other proposed methods. The method of evaluation outlined above differs from other studies, as it attempts to encompass the performance when there are many matchings for each phrase. Other studies[5,7] have used more standard approaches, such as precision and recall, which are viable when there is only one correct medical concept per phrase.

Results

Two datasets used in this study were used with the developed approach. The first dataset, labeled the exploratory dataset, contains the frequencies of chief complaints over a period of a year and was used to perform a semantic group analysis to understand the distribution of concepts in the chief complaints. The second dataset, labeled as the evaluation dataset, has assigned ICD codes for a month of chief complaints and was used to evaluate the accuracy of the method for mapping medical concepts to both ICD-10-CM codes and categories.

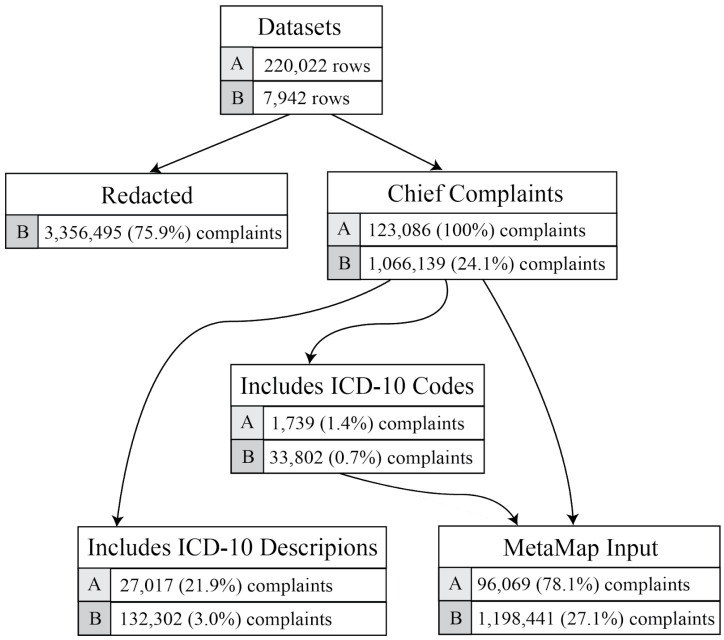

Figure 4. Distribution of Chief Complaints in Dataset A (Evaluation) and B (Exploratory).

Figure 4 shows the breakdown of the number of chief complaints while running the method. The evaluation dataset (Dataset A) contains 220,022 rows, where each row represents a unique mapping from a chief complaint to a label. There are a total of 123,086 unique complaints in dataset A. The exploratory dataset (Dataset B) contains 7,942 rows, where each row contains a chief complaint and the number of times the complaint was recorded over a period of a year. The number of chief complaints for dataset B in each branch in the above diagram was calculated by adding the respective chief complaint frequencies. The majority of labels were redacted, as the data contained are potentially sensitive or identifying. The number of redacted entries in dataset A was not provided.

A semantic group analysis was performed on the exploratory dataset so as to better understand the distribution of chief complaints. Mapped concepts are assigned UMLS semantic types by MetaMap[13], which are general categories that UMLS concepts have been grouped by. Since there are many semantic types, semantic groupings also exist to group the types into broader categories.

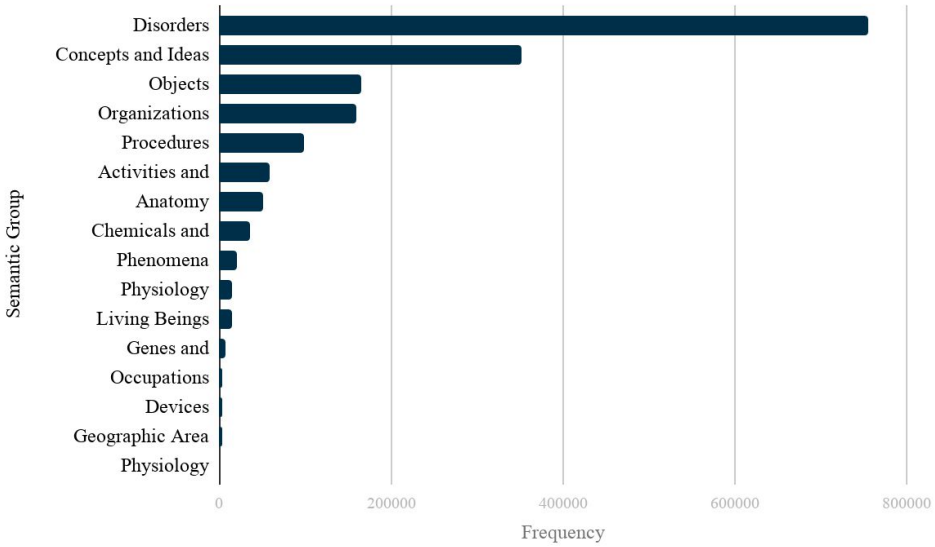

Figure 5. Frequency of UMLS Semantic Types in the Exploratory Dataset.

Figure 5 shows the frequency of recorded UMLS semantic groups in concepts that were mapped by MetaMap from the exploratory dataset. Disorders seem to be the most prevalent concepts in the chief complaint phrases. Since symptom or condition information would fall under the Disorder semantic group, it is important for chief complaints to contain such concepts in order to find ICD-10-CM mappings. The above distribution can also be indicative of the cases where MetaMap has not mapped the phrase to correct concepts. Both Geographic Area and Genes and Molecular Sequences are not concepts expected to be found in chief complaints. Indeed, the phrases that had mapping to these semantic groups had less well known abbreviations that the pre-processor did not expand and hence MetaMap incorrectly assigned non-related concepts. An example of such an abbreviation that MetaMap incorrectly assigns a concept is the chief complaint "st". The abbreviation should be expanded to "sore throat"; however, MetaMap assigns the concept "Santa Lucia", which is a geographic location. Since these unrelated semantic groups are infrequent, as seen in the figure above, the edge cases where MetaMap did not assign the related concepts when evaluating the method is also infrequent.

Results from the evaluation dataset were found using the evaluation method discussed previously. A total of 59.39% ICD-10-CM mappings were found out of the phrases in the evaluation set. Phrases that were not mapped to any CUIs did not contain enough information within them to lead to a successful mapping to diagnosis code.

Table 3. Evaluation of the method on the labeled dataset.

Metric	**ICD-10 Code**	**ICD-10 Category**
% Acc. of total	24.95%	52.16%
% Acc. of mappings	42.02%	**87.82%**

As seen in the table above, while the method underperforms when attempting to predict a full ICD code, due to both label noise and generality of information in free text, the model that maps to ICD-10 categories performs well.

Table 4. Examples where the method successfully mapped the concepts in the phrases to ICD-10-CM categories.

Phrase	Predicted Output	Labels
"streptococcal sore throat"	"J02"	"J02.0"
"severely overweight"	"E66"	"E66.3", "E03.8"
"acute venous embolism"	"I82"	"I82.4Z9"

To illustrate instances of phrases where the method failed to find a correct mapping, several cases from the can be seen in Table 5. The first phrase shows an instance where the predicted output would match the phrase; however, the label is more specific. In the second phrase, MetaMap mapped the phrase to the ICD code of "Congenital syphilis" rather than A64, "Unspecified sexually transmitted disease", which should have been identifiable given the phrase. The third phrase mapped to "Other joint disorder, not elsewhere classified", which is partially correct. The labels for the phrase are specific and the information for such codes is not encompassed in the phrase itself. The last phrase did not get mapped to any ICD-10-CM code, as it seems the additional complexity of the phrase did not allow MetaMap to identify the diagnosis.

Table 5. Examples where the method incorrectly mapped the concepts in the phrases to ICD-10-CM categories.

Phrase	Predicted Output	Labels
"hypothyroidism"	"E03"	"E05.90"
"sextually transmitted diseases"	"A50"	"Z30.90", "A64", "Z00.00"
"right shoulder pain"	"M25"	"S43.51XA", "S46.811A", "Z98.890", "X50.1XXA", "Y93.89", "Y92.234"
"acute right-sided low back pain with sciatica presence"	-	"M54.4"

Discussion

Results from other researched methods of information extraction from chief complaints involved using supervised or heuristic NLP techniques and have also proven to be successful reporting high specificities. Other studies[3,5] used a keyword or bag of word based approach. However, the goals of these other methods of information extraction differ from the present study, as they attempt to classify chief complaints with classification buckets designed by individuals with expert-knowledge. The method outlined in this paper attempts to extract ICD codes and categories that allow for more granularity in predicted labels and thus may prove to be more informative in a clinical, research, or public health setting.

The type of data used to evaluate the proposed method also differs from other researched methods, as the data used for this study can be more easily generalized to real-world chief complaints. Using chief complaints from a specific hospital risks biasing results, as the model could overfit to the hospital's written chief complaint standards and may not perform well on data received from other sources. Future work will involve characterizing chief complaints based on source or hospital as well as examining patterns of mapping based on patient or disease characteristics.

Due to the nature of the problem that was the focus of this study, where there can be multiple predicted outputs and multiple labels for every chief complaint, consideration must be given to the process of evaluating the method. While the scoring and averaging method outlined in this paper gives an intuition on performance, further research should be done on how to best weigh different mapping scenarios. For example, another scoring method would be to

weigh the score of multiple correct mappings higher than for chief complaints where only one correct mapping is found. Additionally, other metrics such as precision and recall could have possible multiple mapping implementations with further research. Definitions of false positives and false negatives could be created in order to create a multiple mapping confusion matrix for future evaluation of the method. The evaluation of the model could also be performed on other unstructured data, such as clinical notes, which could provide a better idea of the method's ability to extract information in different scenarios. Future evaluation methods should also involve an analysis of how many phrases were coded in order to better understand the capabilities of the NLP algorithm or system. Lastly, different mappings from ICD codes to more general diagnosis categories such as those from the Clinical Classifications Software[14] (CCS) could be helpful for evaluating the mappings that are close in meaning, but are classified incorrectly in the current evaluation scheme. With a better understanding of method performance and evaluation, more concrete benchmarks on method improvement can be made.

There are various improvements that can be made to attain better results with the proposed method of the study. The example in Table 2 outlines a case where the method can be improved upon. Right lower quadrant pain usually coincides with symptoms of pain and fever. A certain diagnosis code being associated with other symptoms tends to be a frequent occurrence. Hence, a model to consider to be used in conjunction with the proposed method is that of recommending frequent diagnosis codes that coincide with a given mapped diagnosis code. The output of these recommended diagnosis codes could supplement the mapped ICD-10-CM codes in order to have a higher chance of having a match between mapped code and label. Additionally, these recommended ICD-10-CM codes could provide a more complete set of predicted labels that would have not been inferred by the free text due to the potential generality of a chief complaint phrase.

Further exploration into the configurations that MetaMap provides could lead to better mappings. For example, attempting to use a word sense disambiguation server for MetaMap may prove to help create more accurate mappings for ambiguous chief complaint phrases. Results could also be further improved by the use of other existing NLP tools for biomedical or clinical text[15]. A study[16] has shown that the performance of Apache's clinical Text Analysis and Knowledge Extraction System (cTAKES)[17] is on par with that of MetaMap and could prove to perform better for the task of extracting structured data from chief complaints. Next steps include comparing the performance of MetaMap for mapping of chief complaints to cTAKES, Clinical Language Annotation, Modeling, and Processing Toolkit[18] (CLAMP), and other NLP tools for biomedical and clinical text.

Additional advanced methods could also potentially improve results. Semi-supervised learning and language models with the use of neural networks seem to be promising and warrant additional research in their application of information extraction from chief complaints. An ensemble model that uses both supervised and unsupervised techniques could also prove to have good results. The issue that arises with extracting diagnosis codes is that there are not enough samples of phrases that could train a classifier for each ICD-10-CM code. Hence, an ensemble method could involve having a training dataset and identifying the *n* most frequent ICD-10-CM codes where there are enough samples to train classifiers. For the remaining data where there are not enough training samples for classifiers, unsupervised learning, heuristic rules or the proposed method in this study could be used to extract diagnosis codes from the subset of data. This technique could improve accuracy, as the classifiers could potentially be trained to find specific frequent ICD codes rather than overarching ICD categories of the proposed method and hence would pose to be even more useful in a hospital or public health setting.

Another improvement on the current method would be possible if expanded access were given to other EHR data for patients, using past chief complaints, encounter diagnoses, problems, and family history. These additional data could be factored into current predictions in order to increase the detail of predicted labels from ICD-10-CM categories to codes and may also supplement phrases that only provide partial information. For example, if the phrase "LAB RESULTS" is given, the method could use prior diagnoses and map the lab results to a specific ICD-10-CM code that was previously diagnosed and required lab tests. Additional granularity to the proposed method, as well as, analyses could also be possible if data on source hospitals and EHR systems used are obtained for each chief complaint. The additional data could be used to help the method become more accurate, as abbreviations and chief complaint etiquette for specific hospitals could be trained on and thus the preprocessing can cater to each type of chief complaint. Additionally, given such data, analyses can give feedback to hospitals to create a more standardized method of writing chief complaints. The effect of using these data with the proposed method would be to enrich

health records and have a better understanding of temporally when events occur in a patient's medical history in order to provide better and more efficient care.

Conclusion

In this preliminary study, a method was created to extract structured symptoms or conditions from chief complaints provided by hospitals in Rhode Island through the statewide health information exchange. With a probability given a mapping exists of 87.82%, the method can have real world applications in mapping chief complaint phrases to structured data in the form of ICD categories. Despite the noisiness of the labels, as well as the generality of the phrases, the data used in this study closely reflects the true data-distribution of chief complaints and the issues that arise with the messiness of such data. The method explored in this study can yield better results with further research. The importance of research in the area of information extraction from chief complaints, as one of the earliest forms of information in health records stored during an emergency department visit, cannot be understated.

Acknowledgments

The authors thank Sarah Eltinge and Luke Bruneaux from the Rhode Island Quality Institute for assistance in generating the CurrentCare datasets. This work was funded in part by National Institutes of Health grant U54GM115677. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. Journal of the American Medical Informatics Association. 2011.
2. Chen ES, Melton GB, Burdick TE, Rosenau PT, Sarkar IN. Characterizing the use and contents of free-text family history comments in the Electronic Health Record. AMIA Annu Symp Proc. 2012;
3. Conway M, Dowling JN, Chapman WW. Using chief complaints for syndromic surveillance: A review of chief complaint based classifiers in North America. Journal of Biomedical Informatics. 2013.
4. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. Vol. 88, Journal of Biomedical Informatics. 2018. p. 11–9.
5. Tootooni MS, Pasupathy KS, Heaton HA, Clements CM, Sir MY. CCMapper: An adaptive NLP-based free-text chief complaint mapping algorithm. Comput Biol Med. 2019;
6. Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artif Intell Med. 2005;
7. Brown P, Halász S, Goodall C, Cochrane DG, Milano P, Allegra JR. The ngram chief complaint classifier: A novel method of automatically creating chief complaint classifiers based on international classification of diseases groupings. J Biomed Inform. 2010;
8. Thompson DA, Eitel D, Fernandes CMB, Pines JM, Amsterdam J, Davidson SJ. Coded Chief Complaints-Automated Analysis of Free-text Complaints. Acad Emerg Med. 2006;
9. https://metamap.nlm.nih.gov/
10. https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html
11. Aronson AR, Lang FM. An overview of MetaMap: Historical perspective and recent advances. J Am Med Informatics Assoc. 2010;
12. Terry W, Ostrowsky B, Huang A. Should we be worried? Investigation of signals generated by an electronic syndromic surveillance system--Westchester County, New York. MMWR Morb Mortal Wkly Rep. 2004;
13. Aronson AR. Metamap: Mapping text to the umls metathesaurus. Bethesda MD NLM NIH DHHS. 2006;
14. https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp
15. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of Biomedical Informatics. 2017.
16. 1. Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Med Inform Decis Mak. 2018;
17. https://ctakes.apache.org/
18. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Informatics Assoc. 2018;