Computers in Biology and Medicine 113 (2019) 103398

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

CCMapper: An adaptive NLP-based free-text chief complaint mapping algorithm

Mohammad Samie Tootooni[a,b], Kalyan S. Pasupathy[c,d], Heather A. Heaton[e], Casey M. Clements[e], Mustafa Y. Sir[c,d,*]

[a] Department of Health Informatics and Data Science, Loyola University Chicago, Maywood, IL, USA
[b] Center for Health Outcomes and Informatics Research, Loyola University Chicago, Maywood, IL, USA
[c] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[d] Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, USA
[e] Department of Emergency Medicine, Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Prior Presentations: This work was presented in part at INFORMS Annual Meeting – Phoenix, AZ, November 4–7, 2018.

Keywords:
Natural language processing
Free-text chief complaints
Emergency department
Mapping algorithm
Heuristic
Iterative enhancement
Human consensus-based validation

ABSTRACT

Objective: Chief complaint (CC) is among the earliest health information recorded at the beginning of a patient's visit to an emergency department (ED). We propose a heuristic methodology for automatically mapping the free-text data into a structured list of CCs.

Methods: A comprehensive structured list categorizing CCs was developed by experienced Emergency Medicine (EM) physicians. Using this list, we developed a natural language processing-based algorithm, referred to as Chief Complaint Mapper (CCMapper), for automatically mapping a CC into the most appropriate category (ies). We trained and validated CCMapper using free-text CC data from the Mayo Clinic ED in Rochester, MN. We developed a consensus-based validation approach to handle both indifferences and disagreements between the two EM physicians who manually mapped a random sample of free-text CCs into categories within the structured list.

Results: The *kappa* statistic demonstrated a high level of agreement ($\kappa = 0.958$) between the two physicians with less than 2% human error. CCMapper achieved a total sensitivity of 94.2% with a specificity of 99.8% and F-score of 94.7% on the validation set. The sensitivity of CCMapper when mapping free-text data with multiple CCs was 82.3% with a specificity of 99.1% and total F-score of 82.3%.

Conclusion: Due to its simplicity, high performance, and capability of incorporating new free-text CC data, CCMapper can be readily adopted by other EDs to support clinical decision making. CCMapper can facilitate the development of predictive models for the type and timing of important events in ED (e.g., ICU admission).

1. Introduction

In the earliest stage of a patient's Emergency Department (ED) visit, the chief complaint (CC) is recorded in the electronic health record (EHR) system. This free-text information is a crucial element for clinical decision-making, which affects several facets of a patient's ED visit such as nursing focus, initial medical assessment, prioritizing the treatment, determination of patient flow through the ED, and generation of different hypotheses for the final diagnosis [1–3]. From a health system standpoint, early information about the arriving ED patients can lead to improvements in ED resource allocation and prediction of downstream demand (e.g., admission rates) [4].

Natural Language Processing (NLP) approaches have been used to transform free-text healthcare information (e.g., clinical notes and chief complaints) into structured data [5–8]. While many NLP-based approaches for extracting and aggregating free-text CC data used basic statistics such as frequency [9] and variance [10] of the recorded chief complaints, others recommended employing a pre-organized structured list for the categorization of chief complaints [11–19]. For instance, Safwenberg et al. designed a table of 33 categories determined by four physicians [18]. The Canadian ED Information System (CEDIS) provides a comprehensive structured list of CCs [20], consisting of 165 categories, each associated with a set of ICD-10 codes. A systematic review of different approaches to categorizing CCs was conducted by Malmstrom et al. [14]. They also proposed an adjusted list of 89 categories based on the CEDIS CC categorization. Similar to these studies,

* Corresponding author. Mayo Clinic Department of Health Sciences Research, 200 1st St SW, Rochester, MN, 55905, USA.
E-mail addresses: MTootooni@luc.edu (M.S. Tootooni), Pasupathy.Kalyan@mayo.edu (K.S. Pasupathy), Heaton.Heather@mayo.edu (H.A. Heaton), Clements.Casey@mayo.edu (C.M. Clements), sir.mustafa@mayo.edu (M.Y. Sir).

https://doi.org/10.1016/j.compbiomed.2019.103398
Received 15 May 2019; Received in revised form 13 August 2019; Accepted 19 August 2019
Available online 21 August 2019
0010-4825/

Table 1
Descriptive statistics for the ED patient cohort used in this study.

Variable		Mean (Median/count)	(Q1, Q3)
Waiting Time Before Triage (minutes)		34.4 (5)	(2, 36)
Total Time in ED (minutes)		284.2 (221)	(143, 322)
Age (Years)		46.2 (48)	(25, 67)
Adult (Age ≥ 18)		82.7% (N = 128,082)	
Gender = Female		51.3% (N = 79,406)	
First Visit		66.6% (N = 103,090)	
Resuscitated		1.4% (N = 2167)	
Arrival* =	Walk in	48.8% (N = 75,470)	
	Wheelchair	20.9% (N = 32,407)	
	ALS Surface Ambulance	18.1% (N = 27,981)	
	Carried	5.2% (N = 8093)	
	BLS Surface Ambulance	3.1% (N = 4815)	
	Private Vehicle	1.5% (N = 2394)	
ESI level** =	1	0.7% (N = 1027)	
	2	21.1% (N = 32,731)	
	3	58.9% (N = 91,160)	
	4	18.5% (N = 28,684)	
	5	0.7% (N = 1031)	
Disposition =	Hospitalized	30.4% (N = 47,011)	
	Discharged Home	64.2% (N = 99,294)	
	Left Without Being Seen	1.9% (N = 2980)	

* Other arrival types are: law enforcement agency, Mayo One aircraft, external helicopter, medical van, and commercial vehicle.
** There were 70 patients directly admitted without any specified ESI level.

we use a structured list with 137 categories (see Table A1), which was developed by two EM physicians based on the systems of the human body.

A recorded free-text CC is assigned to a category either manually by EM physicians [18,21–23] or other ED professionals [11,24], or by using an automated mapping algorithm [17,25–28]. Manual categorization of chief complaints is a time- and resource-consuming task [25,29]. Aronsky et al. discussed possible errors associated with the manual categorization and proposed strategies to mitigate such errors [11]. Automated CC categorization, on the other hand, are more suitable for a wide range of ED applications.

We propose a keyword-based [25,30] mapping algorithm, referred to as Chief Complaint Mapper (or CCMapper, in short) to automatically map a free-text CC into a structured list, developed by two Mayo Clinic ED physicians. CCMapper includes an effective low-cost word preprocessing component, that does not need spell-checking and abbreviation substitution; is easy to comprehend, trackable, and adaptable, unlike the existing sophisticated machine learning methods; and can easily be updated by augmenting the bag-of-words when supplied with new free-text CC data from other EDs (see Refs. [11,17] for similar updating methods). CCMapper integrates existing methods into a novel and highly effective heuristic algorithm with outstanding accuracy, even when handling CCs containing multiple categories.

Another contribution of this research is the consensus-based validation approach that can consider both disagreements and indifferences in mapping decisions among physicians. This approach allows more than one ground truth to be considered for each mapping action. Therefore, it can demonstrate how the proposed approach can improve human judgment in practice. Furthermore, the consensus-based validation approach ensures fairness in consideration of missed- and wrong assignments, especially when there are multiple category assignments for a given CC. We believe that the provided error equations can be utilized to provide a baseline for evaluating the performance of similar mapping algorithms applied other free-text data in EHR.

2. Methods

CCMapper is a keyword-based algorithm, which requires a structured list of chief complaint categories along with a set of initial bags-of-words to map the free-text CC. In this section, we present the prerequisites for the development of the CCMapper algorithm and define, the algorithmic steps used in the mapping process. We then describe an iterative enhancement approach to improve the mapping results as more free-text CC data become available. Finally, we introduce the consensus-based validation approach to evaluate the performance of CCMapper.

2.1. Description of available data

We used data from the Mayo Clinic ED in Rochester, MN, which is an academic Level I trauma center with approximately 78,000 patient visits (83% adult) per year. We included all registered ED patients over two years without any exclusion. Table 1 provides statistics describing the patient cohort used in the study. A large portion of these patients was high-acuity (81% with Emergency Severity Index, or ESI, level 1, 2, & 3) and had medically complex conditions; 35% of adult patients and 13% of pediatric patients were admitted.

We removed empty and non-English-letter (e.g., “?” or “ …“) entries from the dataset (approximately 4% of all entries removed). We used the first twenty-one months of the free-text CC records as the training set and the last three months of data for validation purposes. Fig. 1 shows the breakdown of data into the training and validation sets.

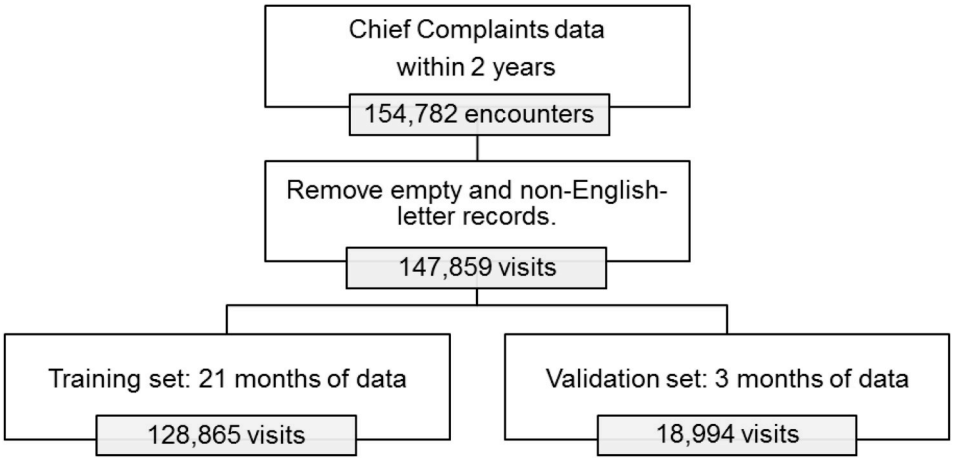

Fig. 1. The breakdown of the free-text chief complaint data collected over a 24-month period.

2.2. Prerequisites for CCMapper

2.2.1. Establishing an expert-based structured list for chief complaints

Two board-certified emergency medicine (EM) physicians designed a structured list for chief complaints by creating an initial list of categories based on the systems of the human body (see Table A.1). They then added more general categories representing complaints from certain events, e.g., environmental, trauma. For each category, they listed common issues such as "pain" or "injury" as well as system-specific issues (e.g., the relation between "nausea and vomiting" and the gastrointestinal system or between "shunt problem" and neurology). The physicians also provided synonyms for some complaints based on their experience, like chest pain/discomfort/heaviness. Lastly, they included a miscellaneous category to capture a variety of non-specific complaints that are common in practice but vague in etiology.

2.2.2. Assembling a bag-of-words corresponding to each category in the structured list

A chief complaint can be verbalized differently in the free-text format. For instance, a person, who visits ED with abdominal pain, can describe their conditions as *"belly pain,"* "*stomachache,"* or other similar ways. The nurse may record it as *{"abd pain"}* or *{"pain abdomen"}*. A *bag-of*-words is a collection of possible recorded words related to a specific category. The words in a bag-of-words do not need to be meaningful or spelled correctly. Conway et al. discuss the importance of capturing similar chief complaints that have *surface variation* (e.g., synonyms and acronyms) for ensuring the accuracy of keyword-based NLP algorithms [25]. They emphasize that capturing truncations, misspellings, and typographic errors is essential for a successful mapping algorithm, especially when spell-checking is not considered in the preprocessing phase. Ergo, in our bag-of-words design, we included words representing synonyms (e.g., *renal* vs. *kidney*), similar symptoms (e.g. *URI Symptoms such as sneeze, runny nose, etc.*), medical terms (e.g. melena vs. tarry stool), different abbreviations (e.g., *abdominal* vs. *abd.*), different forms (e.g., *hands* vs. *hand*), common language effect (e.g., tube vs. catheter), explanation instead of problem (e.g., *can't breathe* vs. *shortness of breath*), common misspelling (e.g., *hypertention* vs. *hypertension*), and detailed vs. general representation (e.g. *finger* vs. *upper extremity*). As a result, each category in the structured list given in Table A.1 was assigned to a bag-of-words, which included all medical terms related to that category along with their possible surface variations.

2.3. Algorithmic steps of CCMapper

CCMapper is composed of two phases: 1) pre-processing the incoming free-text CC, and 2) assignment of the pre-processed CC to one or more categories in the structured list. Fig. 2 shows a flowchart of various steps of the CCMapper algorithm.

2.3.1. Phase I: pre-processing

The pre-processing phase consists of a series of simple modifications to the free-text data mostly to remove *non-value-added* information. For instance, numbers and signs, as well as capital cases in the free-text chief complaint data, often do not have meaningful information. Therefore, CCMapper automatically replaces these with appropriate blank space and small cases. There is no restriction on the length of words to be processed. Lastly, words proceeded by negating adjectives (e.g., no, without) are removed. Fig. 3 exemplifies the pre-processing algorithm:

2.3.2. Phase II: mapping

The first step in the mapping phase is to check if a chief complaint contains any shortcut-word. Shortcut-words (e.g. {"syncope", "lip", "tongue", "vaginal"}) are a list of predetermined words exclusively associated with a category in the structured list. In case the chief complaint contains one of the shortcut-words, the corresponding category is assigned immediately; and there is no need to look for similarities with the assigned category. For instance, consider the CC {"*fall/LBP/Rt leg Inj.*"}. After the preprocessing phase, it gets converted to the list {"fall", "lbp", "rt", "leg", "inj"}. The list contains "fall," which is a shortcut-word for the category "*19.1 trauma-fall.*" CCMapper assigns CC to the category and removes the shortcut-word "fall" from the list. Next, it scans the remaining list of words {"lbp", "rt", "leg", "inj"} and checks if there are any similarities between the list and any of the bag-of-words. The CC is assigned to the category whose bag-of-words has the highest number of similar words. For the above example, the CC is assigned to the category "*9.*1 LE-*Injury*,",[1] whose bag-of-word has two similar words ({"lbp"} and {"leg"}) with the list {"lbp", "rt", "leg", "inj"}. After removing these similar words, (the remaining list is now {"rt", "inj"}), the same process is repeated, and the CC is assigned to the category "*7.1 Back- Back pain*" since the bag-of-words associated with this category contains both words with the remaining list. Therefore, a total of three categories are assigned to the free-text CC {"*fall/LBP/Rt leg Inj.*"}. As a side note, after the mapping process, CCMapper does not rank the assignments and the order of the assignments is not considered in the evaluation of the algorithm's performance.

2.4. Iterative enhancement of CCMapper

The performance of an NLP-based mapping algorithm highly relies on the comprehensiveness of their initial bag-of-words [26,27]. Particularly when the algorithm is trained with data from one hospital, the performance may diminish when applied to data from another hospital. The reasons for such a decline in accuracy include: (1) there may be lingual and cultural differences between different regions [30,31]; (2) common typos made by staff from different hospitals might follow different patterns; (3) common abbreviations used for specific symptoms (such as "bh" for behavioral patients or "n v" for nausea and vomiting) may vary between healthcare systems [32]; and (4) the variety of medical conditions in different regions might be also different [29,33]. Therefore, without knowledge about these nuances in the input data, it is difficult to design an accurate mapping algorithm. One approach to improve the adaptability of a mapping algorithm is to update the table of synonyms periodically to identify common unmatched phrases (iterative enhancement). Aronsky et al. use this approach to refine a set of frequent chief complaints [11]. They report common causes of variation as misspelling, multiple-complaints, writing admission diagnosis instead of chief complaint, and using abbreviations and synonyms. Thompson et al. also discuss the importance of iterative enhancement for mapping algorithms [17].

We propose a five-step approach for iterative enhancement outlined in Fig. 4. In the first step, the initial bag-of-words are manually constructed by an ED expert, who is an informatician with an extensive domain expertise in emergency medicine. After applying the CCMapper algorithm to the training dataset using this initial bag-of-words, many of the free-text CC entries may remain unmapped. The majority of these words are typically surface variations of the existing words in the initial bags-of-words. The ED expert manually updates the bag-of-words by adding the most frequent words in the unmapped free-text CCs to their corresponding bag-of-words. Besides the surface variation, there may be some medical concepts among the unmapped CCs that require the ED expert to seek assistance from an experienced ED physician to identify the correct bag-of-word for assignment. This process is repeated until no significant improvement can be made in the mapping performance. Determining how many such enhancement iterations needed is difficult. Setting a proper number of iterations requires the consideration of the characteristics of the recorded free-text CC data (e.g., the variety of complaints and the range of common typos, and abbreviations). Also, the inclusiveness of the initial bags-of-words and

[1] LE represents Lower Extremity.

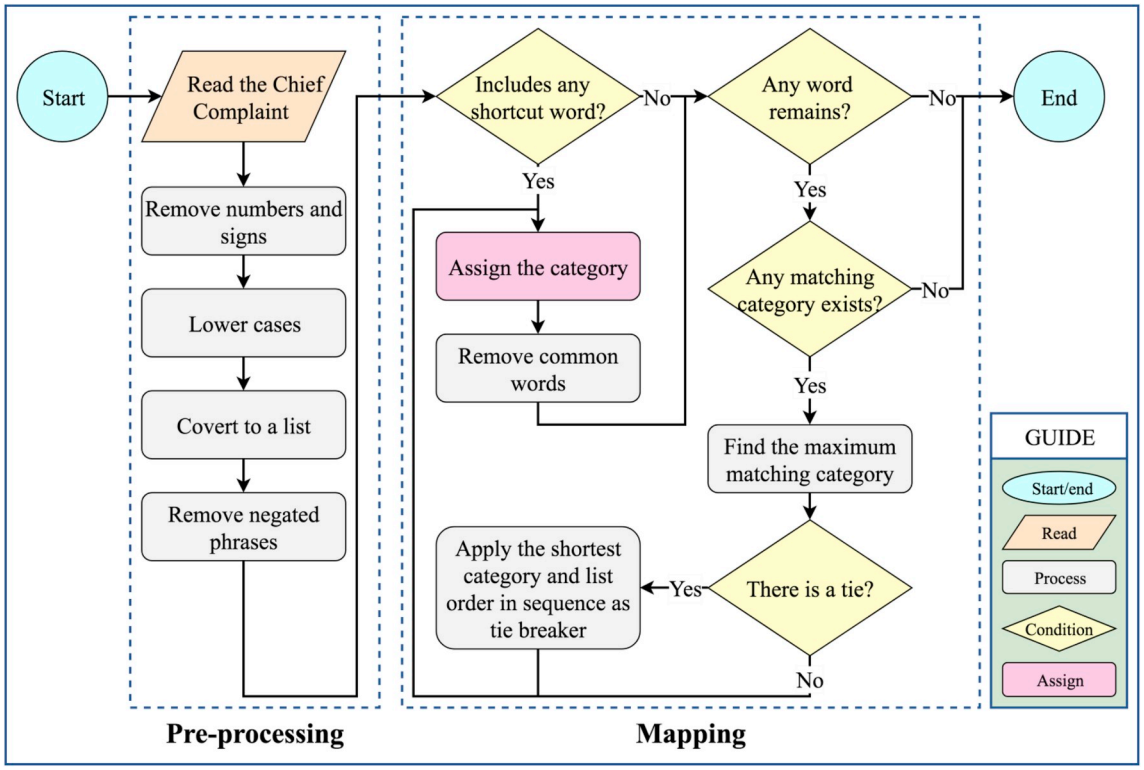

Fig. 2. A flowchart of the proposed CCMapper algorithm. The rectangles represent operations; the diamonds show binary checkpoints; the parallelogram indicates free-text data input; and the circles mark the start and end of the algorithm. The directed arrows indicate the information flow.

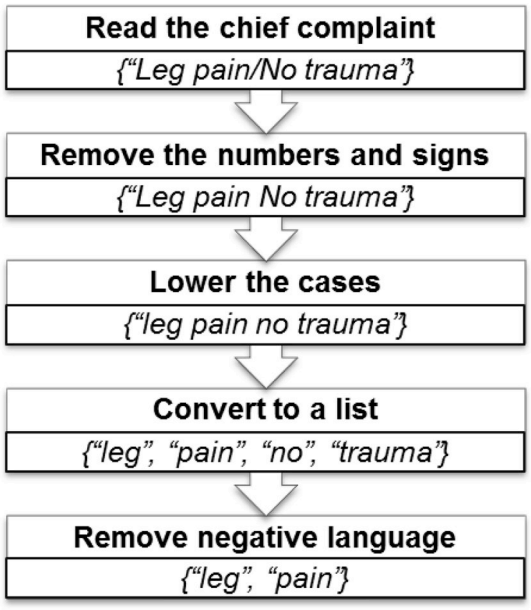

Fig. 3. Illustration of the pre-processing phase of the CCMapper algorithm.

the number of words added to them in each iteration is essential to consider when determining a proper number of iterations. Considering these facts, the iterative enhancement process in this study was terminated after eight iterations resulting in a final set of bag-of-words with a total of 1422 words.

2.5. Validation

Evaluating the performance of CCMapper requires labeled data (i.e., a set of free-text CC data mapped to the categories in the structured list), which requires significant time commitment by the EM physicians. To minimize the burden on the two EM physicians, who agreed to participate in the validation process, we randomly sampled 300 free-text CC entries from the validation dataset. We then asked the two EM physicians to study the provided structured list of chief complaints and manually assign the sampled free-text CCs to the most related category (s). The remaining portion of the validation set remained untouched.

Upon completion of this task, we asked each physician to compare their assignments with those made by CCMapper for each sampled free-text CC, and, for the cases where the assignments were not in agreement, report whether their or CCMapper's assignment was more accurate. Finally, we asked the two physicians to discuss the cases, for which their final assignments differed from each other, and reach a consensus. For those cases that the physicians did not reach a single set of assigned categories or they were indifferent about assigning a particular category, we considered multiple assignment sets as gold-standard. This rigorous validation process provides insights about human judgment error and variations in labeling the free-text CC data by ED experts. Using the consensus-based expert opinion, we evaluated the performance of CCMapper. This validation process is illustrated in Fig. 5 and the results are analyzed in Section 3.

Below, we define necessary notation and key performance metrics for the validation process.

2.5.1. Notation

- I is the set of free-text CC entries randomly sampled from the validation dataset, indexed by i, with $|I| = n$.
- P is the set physicians, who participate in the validation process, indexed by p.
- J is the set of categories in the structured list:
  - $J^{\mathrm{CCM}}(i)$ is a subset of J representing categories that CCMapper assigned to the free-text CC i.
  - $J^{p}(i)$ and $\tilde{J}^{p}(i)$ are subsets of J, representing categories that

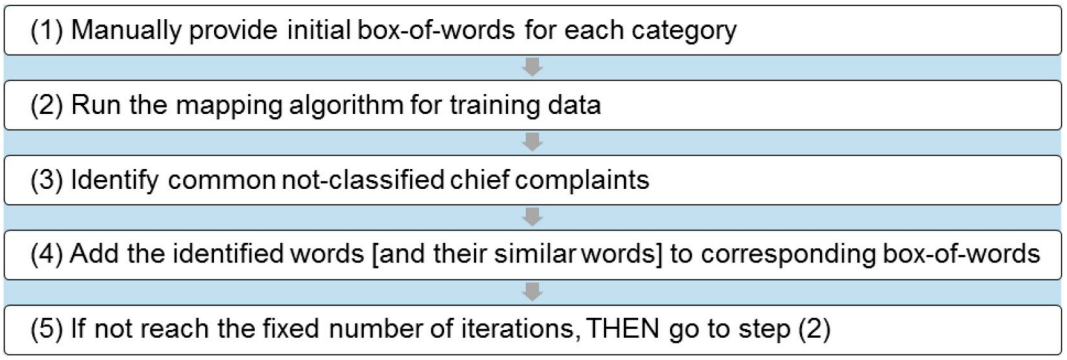

Fig. 4. The steps of the iterative enhancement approach.

physician p assigned to the free text CC i, before and after seeing the CCMapper assignments, respectively.

- $K(i) = \{J_k^{\mathrm{CON}}(i): k \in K(i)\}$is a collection of sets of assigned categories reached by consensus, where each $J_k^{\mathrm{CON}}(i)$ is considered as an acceptable gold-standard set of assigned categories for chief complaint i.
- A set of assigned categories may contain errors either due to missing assignments or wrongly placed assignments with respect to a gold-

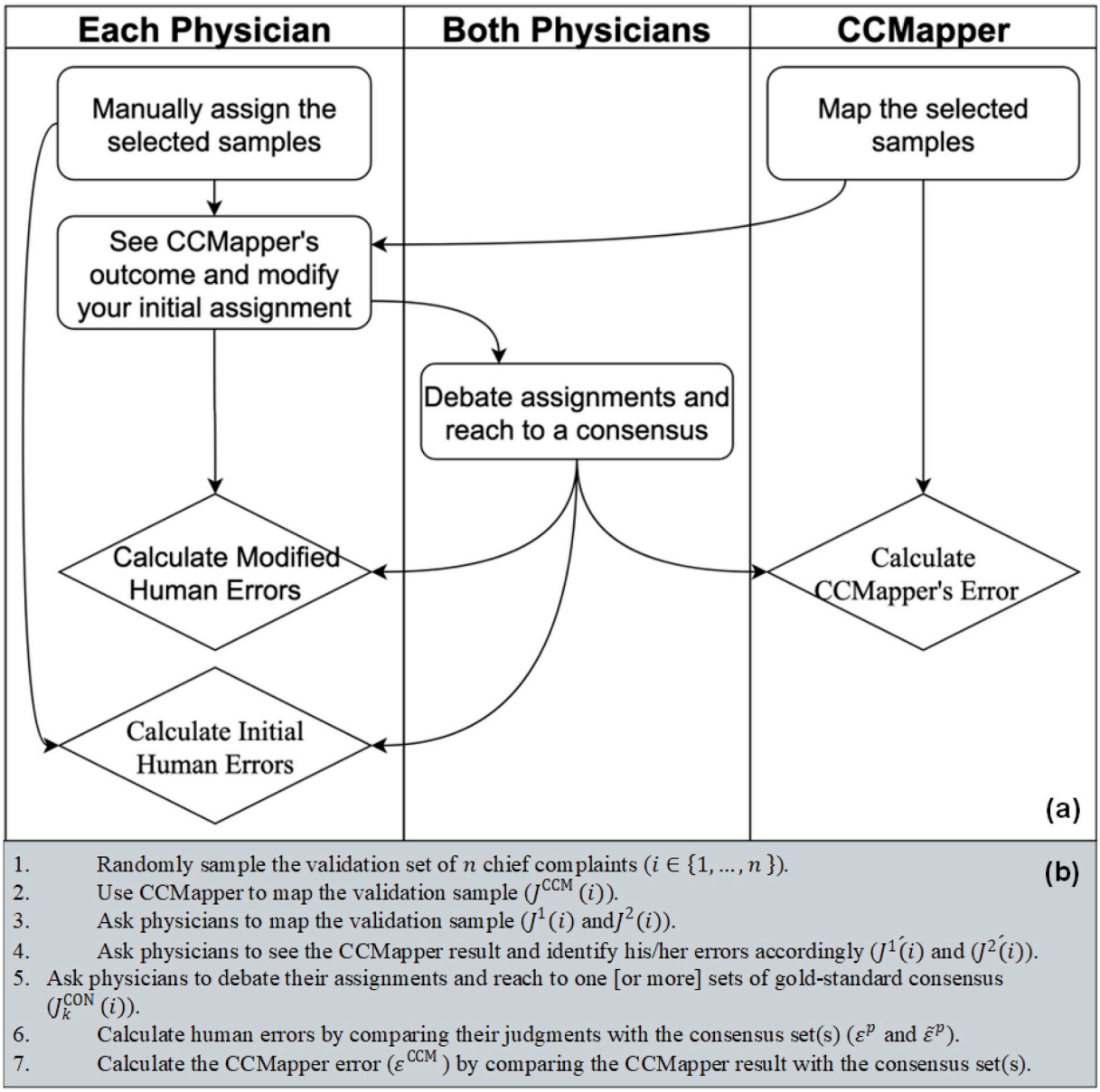

Fig. 5. The consensus-based validation approach. (a) Cross-functional flow diagram of the validation process. (b) The description of each step from the flow diagram.

standard set of assigned categories. We consider three types of error measures to evaluate the performance of CCMapper:

° Average human error percentage before seeing CCMapper assignment:

$$\varepsilon^p = \frac{1}{n}\sum_{i\in I}\min_{k\in K(i)}\left(\frac{|J^p(i) - J_k^{\mathrm{CON}}(i)|}{\max(|J^p(i)|,\ |J_k^{\mathrm{CON}}(i)|)} + \frac{\max(|J_k^{\mathrm{CON}}(i)| - |J^p(i)|,\ 0)}{|J_k^{\mathrm{CON}}(i)|}\right), \forall\ p \in P \quad \text{Eq. (1)}$$

where $|A - B|$ represents the cardinality of the difference of set A from set $B$and the first and second terms in the parenthesis represent the percentage of wrongly placed assignments and missed assignments with respect to a gold-standard set of assigned categories. $J_k^{\mathrm{CON}}(i)$.

° Average human error percentage after seeing CCMapper assignment:

$$\tilde{\varepsilon}^p = \frac{1}{n}\sum_{i\in I}\min_{k\in K(i)}\left(\frac{|\tilde{J}^p(i) - J_k^{\mathrm{CON}}(i)|}{\max(|\tilde{J}^p(i)|,\ |J_k^{\mathrm{CON}}(i)|)} + \frac{\max(|J_k^{\mathrm{CON}}(i)| - |\tilde{J}^p(i)|,\ 0)}{|J_k^{\mathrm{CON}}(i)|}\right), \forall\ p \in P \quad \text{Eq. (2)}$$

° Average error percentage of CCMapper:

$$\varepsilon^{\mathrm{CCM}} = \frac{1}{n}\sum_{i\in I}\min_{k\in K(i)}\left(\frac{|J^{\mathrm{CCM}}(i) - J_k^{\mathrm{CON}}(i)|}{\max(|J^{\mathrm{CCM}}(i)|,\ |J_k^{\mathrm{CON}}(i)|)} + \frac{\max(|J_k^{\mathrm{CON}}(i)| - |J^{\mathrm{CCM}}(i)|,\ 0)}{|J_k^{\mathrm{CON}}(i)|}\right), \forall\ p \in P \quad \text{Eq. (3)}$$

Eq. (1) and Eq. (2) are measures of physician judgment error with respect to the goal-standard set before and after seeing the set of categories assigned by CCMapper to free-text CC data, respectively. Similarly, Eq. (3) represents average percent error in CCMapper assignments. In Appendix B, we provide an example of how these error measures are calculated.

3. Results

We implemented the CCMapper algorithm outlined in Fig. 2 in R version 3.4.2 and ran all numerical experiments on a desktop computer equipped with Intel® Core™ i5-6500 CPU @ 3.20 GHz CPU and 8 GB RAM.

CCMapper assigned 91% and 7.4% of the free-text CC entries in the training dataset to a single category and multiple categories, respectively. CCMapper was unable to assign 1.4% of the free-text CC entries in the training set to any category.

Table 2 shows the breakdown of the free-text CC entries in the training dataset based on the number of assigned categories.

The structured list in Table A1 has 137 categories organized under 19 main groups. The most frequently assigned group and category in the training dataset was Gastrointestinal (GI) (15.04% of all encounters) and Abdominal Pain within the GI group (9.08% of all encounters), respectively. The least frequently assigned group was Endocrine with only 0.41% encounters. Table 3 compares the 19 groups with respect to the number of free-text CC entries in the training dataset assigned to each group.

Table 2
The fraction of the free-text CC entries in training dataset assigned to single vs. multiple categories.

Number of Assigned Categories	Number of entries (n)	%
One	117,544	91.21%
Two	9095	7.06%
Three	482	0.37%
Four	19	0.01%
Total Assigned	127,140	98.66%
Unassigned	1725	1.34%
Total	128,865	100.0%

Table 3
The count and percent of assignments to the chief complaint groups in the structured list given in Table A1.

	Chief Complaint Groups	Number of Assignments (n)	%
1	Gastrointestinal (GI)	20,338	15.0%
2	Neuro	14,860	11.0%
3	Miscellaneous (Misc.)	13,800	10.2%
4	Respiratory (Resp.)	11,499	8.5%
5	Chest	11,167	8.3%
6	Lower Extremity (LE)	10,989	8.1%
7	Upper Extremity (UE)	9293	6.9%
8	Behavioral (BH)	8304	6.1%
9	Genitourinary (GU)	6720	5.0%
10	Trauma	5595	4.1%
11	Ear, Nose, and Throat (ENT)	5065	3.7%
12	Head	4680	3.5%
13	Back	3995	3.0%
14	Eye	2762	2.0%
15	Dermatology (Derm.)	2696	2.0%
16	Neck	1236	0.9%
17	Environment	880	0.7%
18	Allergy	793	0.6%
19	Endocrine	549	0.4%

We evaluated the performance of the CCMapper algorithm using a sample of 300 chief complaints (I), which were not included in the training data. We asked two EM physicians (P) to manually assign each sampled free-text CC entry into one or more categories in the structured list ($J^p(i),\ p \in P,\ i \in I$) without consulting with each other. Then, we asked them to individually compare their assignments with the assignments made by CCMapper ($J^{\mathrm{CCM}}(i)$), determine for which assignments they agreed or disagreed with CCMapper, and modify their initial assignments accordingly ($\tilde{J}^p(i)$). The modified assignments of the two physicians ($\tilde{J}^1(.)$ and $\tilde{J}^2(.)$) were different in 2.94% of the sampled free-text CC entries. These cases included both human judgment error and the cases where the two physicians were indifferent about the different assignment the other colleague made.

To evaluate the level of agreement between the two physicians after seeing the CCMapper assignments, we calculated the Kappa statistic [34]:

$$\kappa = \frac{P_a - P_e}{1 - P_e}, \quad \text{Eq. (4)}$$

where P_a is the relative observed agreement between $\tilde{J}^1(\cdot)$ and $\tilde{J}^2(\cdot)$, and P_e is the probability of an agreement happening by chance. There was a total of 318 ($\sum_{i\in I}|\tilde{J}^1(i)|$) and 310 ($\sum_{i\in I}|\tilde{J}^2(i)|$) assignments made by physician 1 and 2, respectively. The total number of distinct assignments made by both physicians was 319 ($\sum_{i\in I}|J\tilde{J}^1(i) \cup \tilde{J}^2(i)|$). The two physicians had disagreements in 13 out of 319 assignments. Accordingly, the relative observed agreement between physicians was calculated as$P_a = (319 - 13)/319 = 0.959$. The probability of an agreement happening by chance was calculated as $P_e = \sum_{j\in J}\left(\prod_{p\in P}\frac{\sum_{i\in I}|\tilde{J}^p(i)\cap\{j\}|}{319}\right) = 0.031$, where $\sum_{i\in I}|\tilde{J}^p(i)\cap\{j\}|$ is the total number of times that category j is assigned by physician p. The resulting Kappa statistic was calculated as $\kappa = 95.5\%$, implying that the two physicians had a high level of agreement even before establishing a consensus.

The second step of the validation process was carried out to establish a collection of sets of assigned categories reached by consensus for each sampled free-text CC entry. First, the two physicians were asked to arrive at a consensus for the 13 assignments, over which they had disagreements. For more than 50% of these assignments (1.5% of all assignments), the two physicians believed that both of their opinions, although not identical, were valid. Therefore, for each sampled chief

complaint, a consensus was represented by a collection of assigned category sets as opposed to a single set of assigned categories to accommodate such situations. Such cases can provide useful insights about the future improvement and modification of the structured list. A short discussion about these cases is given in Section 4.

Having established consensus-driven collections of assigned category sets as ground truth, we calculated the three types of errors given in Equations Eq. (1)-Eq. (3). The average human error percentage before seeing the CCMapper assignment (ε^{p}) was 2.0% and 1.4% for the physician 1 and 2, respectively. After seeing the CCMapper assignment, the average human percentage ($\tilde{\varepsilon}^{p}$) reduced to 1.0% for physician 1 (50% reduction) and 0.4% for physician 2 (71% reduction). This implies that NLP-based algorithms such as CCMapper can effectively aid human decision-making and reduce human errors.

CCMapper itself showed an average error rate (ε^{CCM}) of 5.56%. Although this error rate is higher than human error, it is still relatively low compared to existing chief complaint mapping algorithms. In addition, considering the possibility of updating the bag-of-word sets through the iterative enhancement approach, the error rate can be continuously improved with the avail of additional free-text chief complaint data. Table C.1 lists all free-text chief complaints that CCMapper failed to map perfectly compared gold standard (consensus set).

We also considered other measures to evaluate the performance of the CCMapper algorithm. Table 4 presents the sensitivity (with 95% Wilson score confidence intervals with continuity correction [35]), precision, specificity, and F-score of CCMapper within each group. The groups with a sample size smaller than six were combined into a single group labeled as “Other.” As can be seen from these results, CCMapper demonstrated excellent performance with total sensitivity of 94.2%, total precision of 95.3%, total specificity of 99.8%, and finally total F-score of 94.7%. CCMapper also performed well when only mapping multiple-complaint cases. Its sensitivity for multiple-complaint cases was 82.3%, with a precision of 82.3%, specificity of 99.1%, and the total F-score of 82.3%.

4. Discussion

Conway et al. highlighted the capability of a mapping algorithm to assign multiple syndromes to a single chief complaint as an essential feature [25]. Many studies mentioned the lack of such capability as a limitation. For example, some studies considered only one complaint (first or “main” complaint) [18,36]; while others considered up to two complaints [10]. A few studies offered ways to prioritize complaints [14,36]. This limitation often results in losing valuable information from free-text chief complaints. CCMapper neither prioritizes nor limits the number of assigned categories to a chief complaint (see Table 2), enabling the retainment of information from the free-text chief complaint data to the fullest extent possible.

An exemplary work of keyword-based mapping systems was developed by Wagholikar et al., who used free-text CC data from 794 ED encounters from six hospitals [19]. They employed a standard structured list provided in Systematized Nomenclature of Medicine – Clinical Terminology (SNOMED CT) [37] and a mapping tool called Snapper [28]. Their method was manually validated by two ED nurses in four main categories, including chest pain, abdominal pain, dyspnea, and trauma. Although their results were promising for some of these categories, the authors acknowledged their algorithm's limited comprehensiveness and weak performance in mapping multiple-complaint encounters. Thompson et al. also proposed an adaptive keyword-based mapping system [17]. They provided a structured list with 228 categories and collected an inventory of CCs (a table of synonyms with more than 2500 phrases) for each category. Their algorithm was able to map 87.5% of chief complaints into their structured list. To the best of our knowledge, there is no existing CC mapping algorithm with such high fidelity rate as CCMapper, evaluated by real-world data, and systematically validated with experts' opinion.

The iterative enhancement approach described in Section 2 improves both the comprehensiveness and the adaptability of CCMapper. Prior studies typically used a generic word-bank dictionary to map free-text chief complaints. One drawback of using generic dictionaries and word-banks is *locality error*. We define locality error as common variations in expressing feelings and symptoms between different regions [19] and common terms used in local practice for clinical data recording. Through the iterative enhancement approach, CCMapper can quickly adapt to the local practice of clinical data recording. CCMapper's ability to identify “missed” chief complaints and update the bag-of-words allows adapting to the dynamically changing medical conditions in the population [17].

As shown in Fig. 1, the first 21 months of this study was considered

Table 4
CCMapper's performance within each group of the structured list given in Table A1, considering the consensus sets established by the two physicians as gold-standard.

	Group	Sensitivity		Precision	Specificity	F-score
		Average	95% CI*			
1	Head	100%	[59.8, 100]	84.2%	99.5%	91.4%
2	Chest	92.8%	[70.9, 99.6]	93.5%	99.5%	93.2%
3	Respiratory (Resp.)	96.9%	[82.0, 100]	100%	100%	98.4%
4	Gastrointestinal (GI)	100%	[89.4, 100]	97.6%	99.6%	98.8%
5	Genitourinary (GU)	62.5%	[25.9, 87.3]	71.4%	99.3%	66.7%
6	Back	100%	[64.3, 100]	100%	100%	100%
7	Upper Extremity (UE)	98.2%	[76.4, 100]	97.4%	99.8%	97.8%
8	Lower Extremity (LE)	94.8%	[72.3, 100]	96.5%	99.8%	95.7%
9	Neuro	97.8%	[86.8, 100]	100%	100%	98.9%
10	Dermatology (Derm.)	84.6%	[42.6, 100]	100%	100%	91.7%
11	Eye	100%	[61.4, 100]	100%	100%	100%
12	Ear, Nose, and Throat (ENT)	92.3%	[62.1, 100]	82.8%	99.1%	87.3%
13	Miscellaneous (Misc.)	90.7%	[75.7, 96.9]	95.8%	99.4%	93.2%
14	Behavioral (BH)	91.7%	[59.8, 100]	100%	100%	95.7%
15	Trauma	100%	[65.1, 100]	100%	100%	100%
16	Other**	100%	[61.4, 100]	100%	100%	100%
	Cumulative	**94.2%**	**[90.8, 96.3]**	**95.3%**	**99.8%**	**94.7%**

* The 95% Confidence interval is based on Wilson Score with continuity corrections.
** Each of Neck, Allergy, Environment, and Endocrine groups in the validation set have less than six samples. These groups were combined into a single group labeled as “Others”.

as training set and a randomly selected sample from the last 3 months of collected CCs was manually labeled and used for validation. One of our concerns was to evaluate the capability of CCMapper in mapping the incoming chief complaints in the future which persuaded us to prefer this way of validation over selecting the validation set randomly from 24 months of the study period. Therefore, the three months period for validation data was selected outside of the training period to test the algorithm performance in cases where the data may slightly change during the time. In addition, to minimize the effect of season-based outbreaks, we considered the last three months of the year which start from late summer to early winter.

5. Limitations

In the validation process, while the two physicians were discussing each sampled chief complaint to reach a consensus, they identified some cases as "indifference," indicating that both physicians believed their own set of assigned categories to be valid even though it differed from that of their colleague. For instance, consider the free-text chief complaint {"*speech, visual changes resolved"}*. Both physicians assigned it to the category "*10.7 Neuro- Rule out CVA*." One of the physicians also assigned it to "*12.5 Eye- Blurry vision/Vision changes NOS*" while the other physician did not. However, they did not disagree with each other judgment in the consensus. A similar scenario occurred for the chief complaint {"*Unresponsive/ETOH*"}; where both physicians assigned it to "*16.2 BH- Alcohol intoxication/abuse/detox/withdrawal*," and only one of them also assigned it to "*14.20 Misc- Consciousness issues*." As reported in Section 3, the two physicians were indifferent about their differing assignments in 1.5% of the sampled cases. Comparing this number with the human error rate of $< 2\%$ reveals that a significant portion of variations in human judgment is due to the structured list itself. Identifying such cases, determining the sources of confusion, and modifying the structured list by adjusting existing categories or adding new ones could improve CCMapper's performance.

We asked two EM physicians to map 300 chief complaints, randomly sampled from approximately 19,000 chief complaints of the validation set. While the sampled 300 chief complaints may be adequate to assess the total performance fidelity of the algorithm, they may not be sufficient to individually evaluate the performance of the algorithm in each group within the structured list. Although designing an experiment with a larger sample size for each group within the structured list could address this issue, manual labeling is a time-consuming and laborious process for physicians. We addressed this concern by reporting the confidence interval in addition to the true positive rate (TPR) for each group. Considering the Bernoulli nature of TPR and limited sample size of each group, we calculated the confidence interval using Wilson score with continuity correction instead of the simple asymptotic normal approximation-based method [38,39]. Finally, to avoid making an unsubstantiated statement about the algorithm's performance, we did not report the CCMapper's performance for groups having an extremely small sample size. Instead, we combined these groups into a single on labeled as "Other."

Although the proposed methodology is designed to accommodate any free-text chief complaints from any institution, it is trained and validated using a dataset from a single institution. Further work is needed to validate the performance of the CCMapper algorithm in different ED settings such as rural vs. urban and pediatric vs. adult.

We have observed a high level of agreement between the two experienced EM physicians participated in the study and reported the human error based on their judgments. However, in practice, the collection of chief complaint data is typically performed by other health professionals such as nurses. Therefore, including nurses in the validation process would improve the validity of the results.

We made a conscious choice to split the training and validation set based on time periods instead of randomly splitting the dataset. Our rationale was to test the capability of CCMapper in mapping the incoming chief complaints in the future, which might be influenced by dynamically changing surface variation as well as seasonality. In this way, we were able to test the algorithm performance while the chief complaint data may slightly change over time. Further testing using random sampling and data across multiple institutions can be beneficial.

5.1. Potential performance improvement

All missed and wrong assignments by CCMapper among the 300 CCs manually mapped by the consensus of two experienced EM physicians are listed in Table C.1. As can be seen, major missed and wrong assignments are due to the incompleteness of the bag-of-words and word removal and tie-breaking processes, as opposed to errors initiated during the pre-processing phases. In order to map a free-text chief complaint, CCMapper searches for a bag-of-word with the maximum number of similar words with the chief complaint. One limitation of the current approach is that the bag-of-words only contain mono-grams and are neutral to the order of the words in the free-text chief complaint. Therefore, incorporating bi-grams and tri-gams to the bag-of-words could potentially improve the algorithm performance.

CCMapper is a keyword-based algorithm and its performance relies on the inclusiveness of the bag-of-words. Spelling errors, different abbreviations, and other variations in the free-text chief complaint data that are not included in the bag-of-words negatively affect the performance of the algorithm. As a future research direction, we plan to incorporate spell-checking and auto-correction as part of the pre-processing phase.

In pre-processing phase, the negating phrase is identified only based on their proximity to negating adjectives. For instance, consider the free-text CC {"*laceration/no pain in leg*"}. During the pre-processing of the phrase "*no pain in leg,*" the word "*pain*" will be removed while the word "*leg*" will be kept. This approach sometimes helps the algorithm to find the right category (for instance, since the word "leg" was not removed, CCMapper will correctly assign category "9.2 LE-Laceration/Abrasion" to the aforementioned CC). However, a more thorough negating phrase removal algorithm capable of identifying longer and more complex negated phrases could potentially improve the performance of the algorithm.

CCMapper has lower sensitivity in cases involving multiple complaints. We believe one main reason for this is the word removal step. For instance, consider the chief complaint {"rt foot pain/sweling"}. CCMapper will assign this chief complaint to the category "9.1 LE. Pain" and will remove common words with the corresponding box-of-word, resulting in the remaining set of words {"rt sweling"}. This set, however, no longer includes "foot," which misleads the algorithm in subsequent assignments. A more intelligent word removal strategy is needed to reduce such errors in the assignments. This is left as a future research direction.

6. Conclusions

In this study, we propose a novel NLP-based heuristic algorithm, referred to as CCMapper, to assign free-text chief complaints into one or more categories within a pre-defined structured list. CCMapper's performance is demonstrated by a robust validation process based on manual assignments performed by two experienced EM physicians. CCMapper relies on an intuitive rule-based approach and can assign a free-text chief complaint to a category in negligible amount time with outstanding accuracy. By design, CCMapper can easily adapt to various local settings and dynamically changing medical conditions in the patient population through an iterative enhancement process. This process decreases the number of missed and wrong assignments and continuously updates the bag-of-words with each new chief complaint recorded.

The proposed CCMapper algorithm enables the usage of free-text

chief complaint data in real-time clinical and operational decision making. The chief complaint is among the earliest health information collected on a patient visiting an ED. Assigning unstructured free-text form into a category within a structured list may make it possible to develop highly accurate machine learning models to predict important outcomes such as disposition (i.e., discharge vs. admission) at the early stages of the care delivery process.

Author contributions

KSP and MYS conceptualized the study and provided feedback on methods and application context. MST contributed in conceptualization, designed the method, performed the data analysis, and drafted the initial manuscript. HAH and CMC (emergency medicine physicians) provided clinical perspectives and participated in manual annotation. All authors contributed to manuscript revision and approved the final manuscript as submitted.

Conflicts of interest

MST, KSP, HAH, CMC, and MYS report no conflict of interest.

Acknowledgments

This work is funded in part by the Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery.

Appendix A. Structured list for chief complaints

Table A. 1
Structured list for chief complaints

Group	Category
1. Head	1.1 Facial/Head/Jaw pain
	1.2 Facial/Head bleeding/laceration/injury
	1.3 Facial rash
	1.4 Facial swelling
	1.5 Lip laceration/sore/inflammation/swelling
	1.6 Scalp lumps
2. Neck	2.1 Pain
	2.2 Laceration/Wound
	2.3 Swelling
	2.4 Injury
3. Chest	3.1 Chest pain/discomfort/heaviness
	3.2 Palpitations/Irregular
	3.3 Breast pain
	3.4 Chest lump
	3.5 Rib pain/swelling/injury/redness
	3.6 Edema
	3.7 Acid firing
4. Respiratory	4.1 Cough
	4.2 Difficulty breathing/Shortness of breath/Respiratory difficulty/Distress
	4.3 Spitting up blood
	4.4 URI symptoms
	4.5 Airway foreign body
	4.6 Wheezing
5. Gastrointestinal	5.1 Abdominal pain
	5.2 Nausea, Vomiting
	5.3 Diarrhea
	5.4 Wound
	5.5 GI bleed
	5.6 Constipation
	5.7 Food bolus/FB In throat
	5.8 Redness
	5.9 Rectal pain
	5.10 Jaundice
	5.11 Feeding tube issue
	5.12 Bowel obstruction
	5.13 Ileostomy
	5.14 Fistula problem
	5.15 Hernia problem/evaluation
6. Genitourinary	6.1 Flank pain/Kidney pain
	6.2 Vaginal bleeding/discharge/irritation/pain/sores/swelling
	6.3 Catheter issue
	6.4 STI/Yeast infection
	6.5 Dysuria/Hematuria/Urinary/UTI symptoms
	6.6 Injury/Problem
	6.7 Groin pain
	6.8 Groin sores/wound/redness
	6.9 Pelvic pain
	6.10 Penis/Testicular swelling/pain
	6.11 Pregnancy test
	6.12 Urinary retention/Unable to void
7. Back	7.1 Back pain
	7.2 FB in back

(*continued on next page*)

Table A. 1 (*continued*)

Group	Category
8. Upper Extremity	8.1 Pain
	8.2 Laceration/Abrasion
	8.3 Injury
	8.4 Infection/Redness/Wound check/lump
	8.5 Swelling
9. Lower Extremity	9.1 Pain
	9.2 Laceration/Abrasion
	9.3 Injury
	9.4 Swelling
	9.5 Hematoma
	9.6 Infection/Redness/Wound check/lump
10. Neuro	10.1 Headache/Migraine
	10.2 Weakness/Fatigue/Tired
	10.3 Altered mental status/Confusion
	10.4 Dizziness
	10.5 Numbness/Tingling
	10.6 Lightheaded/Near syncope
	10.7 Rule out CVA
	10.8 Seizure/Shaking/Tremors
	10.9 Syncope
	10.10 Shunt problem
	10.11 Vertigo
11. Dermatology	11.1 Rash/Hives
	11.2 Laceration/Abrasion/Bite/Lesion NOS[2]
	11.3 Abscess
	11.4 Itching
	11.5 Scabies
	11.6 Blister
	11.7 Skin Issue
	11.8 Sunburn/Burn(S)
	11.9 Bruising
12. Eye	12.1 Eye Pain
	12.2 Floaters
	12.3 Orbits
	12.4 Eye Swelling
	12.5 Blurry vision/Vision changes NOS
	12.6 Eye drainage/redness
	12.7 Eye NOS
	12.8 Eye Foreign body
13. Ear, Nose, and Throat	13.1 Dental pain
	13.2 Ear pain
	13.3 Difficulty swallowing
	13.4 Epistaxis
	13.5 Mouth injury/sores
	13.6 Nose injury/laceration
	13.7 Sore throat
	13.8 Throat swelling/tightness
	13.9 Tongue pain
14. Miscellaneous	14.1 Pain
	14.2 Post-operative/procedure Complication/Dressing change/Wound check NOS
	14.3 Infection/Fever/Chills NOS
	14.4 Abnormal labs/imaging
	14.5 Assault
	14.6 Body aches/Body sensitivity
	14.7 Chest congestion
	14.8 Blood pressure
	14.9 Fussy
	14.10 Medication refill/reaction
	14.11 Multiple complaints*
	14.12 Needle stick
	14.13 Referral
	14.14 Dialysis
	14.15 Restless legs
	14.16 Safety evaluation
	14.17 Vascular problem
	14.18 Foreign body NOS
	14.19 Dehydration
	14.20 Consciousness issues
	14.21 Scheduled recheck
	14.22 Decreased oral intake
	14.23 wt loss
15. Allergy	15.1 Allergic reaction
	15.2 Lymph node

(*continued on next page*)

Table A. 1 (*continued*)

Group	Category
16. Behavioral	16.1 Accidental overdose
	16.2 Alcohol intoxication/abuse/detox/withdrawal
	16.3 Anxiety/Depression/SI
	16.4 Hallucinations
	16.5 Insomnia
	16.6 Evaluation
17. Environmental	17.1 Bat bite/exposure
	17.2 Bee sting/Insect bite
	17.3 Dog/Cat/Animal bite/scratch
18. Endocrine	18.1 Blood sugar
19. Trauma	19.1 Fall
	19.2 Motor Vehicle Accident (MVA)

[2] Not Otherwise Specified.

Appendix B. An example for the validation process

Below, we provide an example of how the validation process works. We describe each step of the validation process in detail to clarify how different error measures are calculated. Ph.1 and Ph.2 represent the two physicians who participated in the validation process.

1- Assignments of a free-text chief complaint into multiple categories:

- Ph.1 assigned the chief complaint to {Cat.1, Cat.4, Cat.5, Cat.6}
- Ph.2 assigned the chief complaint to {Cat.3, Cat.4}
- CCMapper assigned the chief complaint to {Cat.1, Cat.2, Cat.4}

2- Assignments change AFTER seeing CCMapper outcome:

- Ph.1 made no change.: {Cat.1, Cat.4, Cat.5, Cat.6}
- Ph.2 changed his opinion from Cat.3 to Cat.2.: {Cat.2, Cat.4}

3- The two physicians discuss about their assignment and reaches to a consensus:

- Consensus sets: {Cat.2, Cat.4, Cat.6} & {Cat.2, Cat.4, Cat.5, Cat.6}
  - Ph.1 accepted that Cat.1 is a unfitting assignment.
  - Ph.2 accepted that he missed Cat.6.
  - The two physicians had no concern about Cat.5 to get assigned.

4- Error Calculation

- The second consensus set is used to calculate Ph.1's assignment error. She had originally one unfitting, and no missed assignment: 1/4 + 0/4 =25%. She has the same error rate after seeing the CCMapper outcome.
- The first consensus set is used to calculate Ph.2's assignment error. He had originally one unfitting, and one missed assignment:1/3 +1/3 =67%. His error after seeing the algorithm result became 0/3 + 1/3 = 33%.
- The first consensus set is used to calculate CCMapper's assignment error. It had no missed and one unfitted assignment: 0/3 +1/3 =33%.

Appendix C. List of missed and wrong assignments by CCMapper

Table C. 1
List of missed and wrong assignments by CCMapper among the 300 CCs manually mapped by the consensus of two experienced EM physicians.

Free-text CC	CCMapper Assignment			Reason Behind Missed/Wrong Assignment
	Correct Assignment	Missed Assignment	Wrong Assignment	
Left Sided weakness		10.7 Rule out CVA	10.2 Weakness/Fatigue/Tired	Shared word in bag-of-word, tie-breaking*
facial numbness	10.5 numbness/tingling		1.1 facial/head/jaw pain	Word-removal Process**
voiding problems		6.12 Urinary retention/Unable to void		Incomplete Bag-of-word***
Chest Pressure/Tightness		3.1.1 Chest pain/discomfort/heaviness		Incomplete Bag-of-word
biopsy site bleeding		14.2 Post-operative/procedure Complication/Dressing change/Wound check nos	13.5 mouth injury/sores	Shared word in bag-of-word, tie-breaking
Dehydrated		14.19 Dehydration		Incomplete Bag-of-word
sternal pain post EGD	14.2 Post-operative/procedure Complication/Dressing change/Wound check nos		14.1 pain	Word-removal Process
Genito Urinary Problem		6.6 Injury/Problem	6.2 Vaginal bleeding/discharge/irritation/pain/sores/swelling	Incomplete bag-of-word
post-op bleeding	14.2 post-operative/procedure complication/dressing change/wound check NOS		13.5 mouth injury/sores	Word-removal Process
left cheek lesion		11.2 Laceration/Abrasion/Bite/Lesion nos	1.2 facial/head bleeding/laceration/injury	Shared word in bag-of-word, tie-breaking
Right sided weakness		10.7 Rule out CVA	Neuro - 10.2 Weakness/fatigue/tired	Shared word in bag-of-word, tie-breaking
Lower Extremity Weakness	10.2 Weakness/Fatigue/Tired		8.1 Pain	Word-removal Process
dypnea		4.2 Difficulty breathing/Shortness of breath/Respiratory difficulty/Distress		Incomplete bag-of-word
Sepsis Eval		14.3 Infection/Fever/Chills nos	3.1.1 Chest pain/discomfort/heaviness	Incomplete bag-of-word & Shared word in bag-of-word, tie-breaking
Fall hip radius fx	19.1 Fall	8.3 Injury 9.3 Injury	9.1 Pain	Incomplete bag-of-word; &Word-removal Process
fecal incontinence		5.3 Diarrhea		Incomplete bag-of-word
Ingestion		16.1 Accidental overdose	5.7 Food bolus/FB in throat	Incomplete bag-of-word
human		11.2 Laceration/Abrasion/Bite/Lesion nos	Environment - 17.1 Bat bite/exposure	Incomplete bag-of-word & Shared word in bag-of-word, tie-breaking

* **Shared word in bag-of-word, tie-breaking**: is when one or more word in the free-text CC exists in more than one category and the current tie-breaker chooses a wrong category.
** **Word-removal Process**: is when a word in the free-text CC is related to more than one category but after assigning the first category it is *removed* and causes missing the other related categories. Also, due to not having multi-grams in bag-of-words, sometimes a whole phrase is related to a category but only one-word of a phrase exists in the category, and the remaining word(s) causes additional mapping(s) to wrong category(ies).
*** **Incomplete Bag-of-word**: is when one or more word in the free-text CC does not exist in the correct category's bag-of-word and causes missing category.

References

[1] H.C. Sox, M.A. Blatt, M.C. Higgins, K.I. Marton, Medical Decision Making, ACP Press 2007.
[2] E.N. Association, Making the Right Decision: a Triage Curriculum, Des Plaines, Author, IL, 2001.
[3] D.A. Travers, S.W. Haas, Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department, J. Biomed. Inform. 36 (2003) 260–270.
[4] J.P. Ruger, L.M. Lewis, C.J. Richter, Identifying high-risk patients for triage and resource allocation in the ED, Am. J. Emerg. Med. 25 (2007) 794–798.
[5] N. Sager, M. Lyman, C. Bucknall, N. Nhan, L.J. Tick, Natural language processing and the representation of clinical data, J. Am. Med. Inform. Assoc. 1 (1994) 142–160.
[6] U. Hahn, K. Schnattinger, M. Romacker, Automatic knowledge acquisition from medical texts, Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association, 1996, p. 383.
[7] C. Friedman, G. Hripcsak, Natural language processing and its future in medicine, Acad. Med. 74 (1999) 890–895.
[8] R.H. Baud, C. Lovis, P. Ruch, A.-M. Rassinoux, A toolset for medical text processing, Stud. Health Technol. Inform. 77 (2000) 456–461.
[9] D. Tandberg, C. Qualls, Time series forecasts of emergency department patient volume, length of stay, and acuity, Ann. Emerg. Med. 23 (1994) 299–306.
[10] S. Poole, S. Grannis, N.H. Shah, Predicting emergency department visits, AMIA Summits Transl. Sci. Proc. 2016 (2016) 438–445.
[11] D. Aronsky, D. Kendall, K. Merkley, B.C. James, P.J. Haug, A comprehensive set of coded chief complaints for the emergency department, Acad. Emerg. Med. 8 (2001) 980–989.
[12] D.P. Hansen, M.L. Kemp, S.R. Mills, M.A. Mercer, P.A. Frosdick, M.J. Lawley, Developing a national emergency department data reference set based on SNOMED CT, Med. J. Aust. 194 (2011) S8–S10.
[13] M.E. Matheny, F. FitzHenry, T. Speroff, J.K. Green, M.L. Griffith, E.E. Vasilevskis, E.M. Fielstein, P.L. Elkin, S.H. Brown, Detection of infectious symptoms from VA emergency department and primary care clinical documentation, Int. J. Med. Inform. 81 (2012) 143–156.
[14] T. Malmström, O. Huuskonen, P. Torkki, R. Malmström, Structured classification for ED presenting complaints – from free text field-based approach to ICPC-2 ED application, Scand. J. Trauma Resusc. Emerg. Med. 20 (2012) 76.
[15] C.A. Sniegoski, Automated syndromic classification of chief complaint records, Johns Hopkins APL Tech. Dig. 25 (2004) 68–75.
[16] W.W. Chapman, J.N. Dowling, M.M. Wagner, Classification of emergency department chief complaints into 7 syndromes: a retrospective analysis of 527,228 patients, Ann. Emerg. Med. 46 (2005) 445–455.
[17] D.A. Thompson, D. Eitel, C. Fernandes, J.M. Pines, J. Amsterdam, S.J. Davidson, Coded chief complaints—automated analysis of free-text complaints, Acad. Emerg. Med. 13 (2006) 774–782.
[18] U. Safwenberg, A. Terént, L. Lind, Differences in long-term mortality for different emergency department presenting complaints, Acad. Emerg. Med. 15 (2008) 9–16.
[19] A.S. Wagholikar, M.J. Lawley, D.P. Hansen, K. Chu, Identifying symptom groups from emergency department presenting complaint free text using SNOMED CT, AMIA Ann. Symp. Porc. 2011 (2011) 1446–1453.
[20] E. Grafstein, M.J. Bullard, D. Warren, B. Unger, Revision of the Canadian emergency department information system (CEDIS) presenting complaint list version 1.1, CJEM, Can. J. Emerg. Med. 10 (2008) 151–161.
[21] J. Wiswell, K. Tsao, M.F. Bellolio, E.P. Hess, D. Cabrera, “Sick” or “not-sick”:

accuracy of System 1 diagnostic reasoning for the prediction of disposition and acuity in patients presenting to an academic ED, Am. J. Emerg. Med. 31 (2013) 1448–1452.
[22] D. Cabrera, J.F. Thomas, J.L. Wiswell, J.M. Walston, J.R. Anderson, E.P. Hess, M.F. Bellolio, Accuracy of 'my gut feeling:' comparing system 1 to system 2 decision-making for acuity prediction, disposition and diagnosis in an academic emergency department, West. J. Emerg. Med. 16 (2015) 653–657.
[23] M. Mockel, J. Searle, R. Muller, A. Slagman, H. Storchmann, P. Oestereich, W. Wyrwich, A. Ale-Abaei, J.O. Vollert, M. Koch, R. Somasundaram, Chief complaints in medical emergencies: do they relate to underlying disease and outcome? The Charité Emergency Medicine Study (CHARITEM), Eur. J. Emerg. Med. 20 (2013) 103–108.
[24] M.R. Vaghasiya, M. Murphy, D. O'Flynn, A. Shetty, The emergency department prediction of disposition (EPOD) study, Australas. Emerg. Nurs. J. 17 (2014) 161–166.
[25] M. Conway, J.N. Dowling, W.W. Chapman, Using chief complaints for syndromic surveillance: a review of chief complaint based classifiers in North America, J. Biomed. Inform. 46 (2013) 734–743.
[26] K.B. Cohen, L. Hunter, Getting started in text mining, PLoS Comput. Biol. 4 (2008) e20.
[27] J.H. Martin, D. Jurafsky, Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Pearson/Prentice Hall2009.
[28] D. Vickers, M. Lawley, Mapping existing medical terminologies to SNOMED CT: an investigation of the novice user's experience, HIC 2009, Proceedings; Frontiers of Health Informatics-Redefining Healthcare, National Convention Centre Canberra, 19-21 August 2009, Health Informatics Society of Australia (HISA), 2009, p. 46.
[29] C. Sniegoski, Methodology for categorizing emergency department chief complaint records for syndromic surveillance, Proceedings of the 131st Annual Meeting of the American Public Health Association, 2003.
[30] H.-M. Lu, H. Chen, D. Zeng, C.-C. King, F.-Y. Shih, T.-S. Wu, J.-Y. Hsiao, Multilingual chief complaint classification for syndromic surveillance: an experiment with Chinese chief complaints, Int. J. Med. Inform. 78 (2009) 308–320.
[31] H. Mowafi, D. Dworkis, M. Bisanzo, B. Hansoti, P. Seidenberg, Z. Obermeyer, M. Hauswald, T.A. Reynolds, Making recording and analysis of chief complaint a priority for global emergency care research in low-income countries, Acad. Emerg. Med. 20 (2013) 1241–1245.
[32] H. Yu, G. Hripcsak, C. Friedman, Mapping abbreviations to full forms in biomedical articles, J. Am. Med. Inform. Assoc. 9 (2002) 262–272.
[33] G.T. Perkoff, M. Anderson, Relationship between demographic characteristics, patient's chief complaint, and medical care destination in an emergency room, Med. Care 8 (1970) 309–323.
[34] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.
[35] R.G. Newcombe, Two-sided confidence intervals for the single proportion: comparison of seven methods, Stat. Med. 17 (1998) 857–872.
[36] R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulldorff, D. Weiss, Syndromic surveillance in public health practice, New York City, Emerg. Infect. Dis. 10 (2004) 858–864.
[37] M. Bramley, R. Richards, M. Cordell, C. Richardson, M. Guo, R. Richards, NEHTA terminology analysts, health information management, J. Health Inf. Manag. Assoc. Aust. 38 (2009) 59–63.
[38] D. Wenzel, A. Zapf, Difference of two dependent sensitivities and specificities: comparison of various approaches, Biom. J. 55 (2013) 705–718.
[39] S. Erdoğan, O.T. Gülhan, Alternative confidence interval methods used in the diagnostic accuracy studies, Comput. Math. Methods Med. (2016) 2016.