International Journal of Medical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records

Yue Wang[a,b,1], Jin Luo[b,1], Shiying Hao[b,1], Haihua Xu[d,1], Andrew Young Shin[c,1], Bo Jin[d,1], Rui Liu[b], Xiaohong Deng[f], Lijuan Wang[d], Le Zheng[b], Yifan Zhao[d], Chunqing Zhu[d], Zhongkai Hu[d], Changlin Fu[d], Yanpeng Hao[b], Yingzhen Zhao[b], Yunliang Jiang[b], Dorothy Dai[d], Devore S. Culver[e], Shaun T. Alfreds[e], Rogow Todd[e], Frank Stearns[d], Karl G. Sylvester[b,2], Eric Widen[d,2], Xuefeng B. Ling[b,*,2]

[a] State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, PR China
[b] Departments of Surgery, Stanford University, Stanford, CA 94305, USA
[c] Departments of Pediatrics, Stanford University, Stanford, CA 94305, USA
[d] HBISolutions Inc., Palo Alto, CA 94301, USA
[e] HealthInfoNet, Portland, ME 04103, USA
[f] Chongqing Key Lab of Catalysis & Functional Organic Molecules, Chongqing Technology and Business University, Chongqing, China

ARTICLE INFO

Article history:
Received 7 February 2015
Received in revised form 23 June 2015
Accepted 25 June 2015
Available online xxx

Keywords:
Congestive heart failure
Random forests
Natural language processing
Electronic Medical record
Prospective validation

ABSTRACT

Background: In order to proactively manage congestive heart failure (CHF) patients, an effective CHF case finding algorithm is required to process both structured and unstructured electronic medical records (EMR) to allow complementary and cost-efficient identification of CHF patients.
Methods and results: We set to identify CHF cases from both EMR codified and natural language processing (NLP) found cases. Using narrative clinical notes from all Maine Health Information Exchange (HIE) patients, the NLP case finding algorithm was retrospectively (July 1, 2012–June 30, 2013) developed with a random subset of HIE associated facilities, and blind-tested with the remaining facilities. The NLP based method was integrated into a live HIE population exploration system and validated prospectively (July 1, 2013–June 30, 2014). Total of 18,295 codified CHF patients were included in Maine HIE. Among the 253,803 subjects without CHF codings, our case finding algorithm prospectively identified 2411 uncodified CHF cases. The positive predictive value (PPV) is 0.914, and 70.1% of these 2411 cases were found to be with CHF histories in the clinical notes.
Conclusions: A CHF case finding algorithm was developed, tested and prospectively validated. The successful integration of the CHF case findings algorithm into the Maine HIE live system is expected to improve the Maine CHF care.

1. Introduction

The US Centers for Disease Control and Prevention (CDC) has reported that congestive heart failure (CHF) remains a principle cause of overall hospitalization and its prevalence has not changed significantly between 2000 and 2010 [1]. The estimated heart failure related mortality is approximately 287,000 people per year [2]. In aggregate, heart failure imparts an enormous yearly cost of approximately $31 billion dollars to the US healthcare system [3].

The Centers for Medicare and Medicaid Services (CMS) has proposed CHF readmission rate as a measure of healthcare quality and target for cost control [4]. Many CHF hospitalizations are considered to be preventable if patients were to receive timely and appropriate medical care [5]. Therefore, an effective real-time analytical solution to comprehensively identify CHF cases is needed to help guide targeted interventions and appropriate resource allocation [6].

A traditional method for CHF case finding is based on clinical coding [7] that largely depends on the availability of structured electronic medical record (EMR) datasets. However, this method is flawed resulting in a significant under-reporting of the targeted population [8]. One solution is to find those uncodified CHFs by

* Corresponding author at: Department of Surgery, S370 Grant Bldg., Stanford University School of Medicine, Stanford, CA 94305, USA.
E-mail address: bxling@stanford.edu (X.B. Ling).
[1] Co-first authors.
[2] Co-last authors.

http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007
1386-5056/

manual review of the narrative EMR clinical notes. However, high labor costs and latency prohibit the practicality of this approach. Therefore, processing both structured and unstructured data for CHF case finding can provide a complementary and cost-efficient way to identify patients and apply targeted care.

Over the last decade, use of Natural Language Processing (NLP) to analyze the EMR narrative texts has been largely confined to clinical research focusing on information extraction for the purpose of EMR enrichment and decision support [10–12]. Alternatively, the applications of NLP to clinical notes, e.g., identification of pneumonia [13], diabetes [14], and CHF [15,16], have shown promise as a case finding method. The reported NLP based case finding studies to date have achieved good *F*-measures [9]. However, these studies utilized relatively small sample size or focused on specific types of clinical notes. The challenge in the current study was to execute CHF case finding utilizing a statewide patient population, where the class distribution is highly imbalanced. Unstructured notes from multiple facilities across the Maine State were found with known narrative expression variability thus impacting information comprehensiveness. To deal with this challenge, the algorithm should have: (1) a comprehensive knowledge base to capture the accumulated domain knowledge from the targeted patient population; (2) a rigorous data model is needed to encompass the unstructured clinical notes of various formats across different facilities; (3) a robust and scalable analytical pipeline is needed to process the vast amount of EMR notes across statewide facilities.

In this study, we set to develop and integrate a real-time NLP-based CHF case finding algorithm into the Maine HIE care flow (Fig. 1).

2. Methods

2.1. Ethics statements

No PHI was released for the purpose of this clinical research. Because this study analyzed de-identified data, the Stanford University Institutional Review Board considered it exempt (October 16, 2014).

2.2. Data source

The health information exchange in Maine (HealthInfoNet, HIN) is an independent nonprofit organization initiated in 2009 that contains records for nearly the entire population of the state of Maine residents and is connected to the majority of health care facilities in the state. There are currently 35 hospitals, 384 federally qualified health centers, and over 400 ambulatory practices connected to the Maine HIE. HIN maintains an opt-out consent process with a patient opt-out rate of slightly over 1%; and certain behavioral health and HIV related information is excluded from the database as required by Maine law. To identify the CHF cohort from a statewide population, all categories of clinical notes from the connected facilities were included. There were 2,139,299 notes in the Maine HIE EMR database covering a period from July 1, 2012 to June 30, 2014, with more than 100 different types of clinical reports including history/physical reports, discharge summaries and emergency reports.

2.3. Experimental design (Fig. 2)

CHF cases were identified utilizing the Clinical Classification Software (CCS) single-level diagnosing group (#108 Congestive heart failure; nonhypertensive) [17], and the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes including 398.91, 428.0, 428.1, 428.20, 428.21, 428.22, 428.23, 428.30, 428.31, 428.32, 428.33, 428.40, 428.41, 428.42, 428.43, and 428.9. The CHF case finding algorithm consisted of two phases: (1) the EMR database analyses of patient encounters with the CHF ICD9 codes; (2) NLP-based case finding analyses based on our knowledge base and de-noised dictionary (see below). Total of 8 datasets as outlined in Fig. 2 were utilized throughout this study: retrospective A and A1 to A4, prospective B and B1 to B2 datasets. The NLP engine was trained with data set A1, analyzed with dataset A2, finalized with manually chart-reviewed gold standard dataset A3, and evaluated with manually chart-reviewed gold standard dataset A4 in a retrospective timeframe from July 1, 2012 to June 30, 2013. Our case finding algorithm was prospectively deployed to the HIE live system (Supplementary Table 1). The algorithm's prospective performance was gauged using another chart-reviewed gold standard dataset B1 of uncoded encounters within the prospective testing period from July 1, 2013 to June 30, 2014. The clinical notes of the NLP identified cases were further profiled to explore unique clinical patterns associated with these previously uncoded but genuine CHF cases.

2.4. Workflow to construct the gold standard dataset A3, A4 and B1

Clinical notes of samples were randomly selected and manually reviewed by two physician curators. When there was a disagreement on diagnosis that could not be resolved by the curators, the sample was excluded. The resultant datasets was used as the gold standard to validate our NLP case finding method.

2.5. NLP knowledge base

The developed knowledge base has two modules: (1) a controlled vocabulary consisting of CHF related clinical terms; and (2) the extracted rules combining vital signs and comorbidities in the clinical notes (Supplementary Fig. 1).

The clinical terms in our NLP knowledge base were derived from the following sources: (1) ICD-9-CM code string descriptions and corresponding synonyms; (2) the comprehensive clinical terminology within the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) [18]; (3) a mapping between ICD-9-CM and SNOMED CT proposed by the US National Library of Medicine (NLM) [19]; and (4) a controlled vocabulary thesaurus named Medical Subject Headings (MeSH) used by NLM for article indexing [20]. These clinical terms in the knowledge base were further tokenized, combined and filtered to derive our controlled vocabulary of single and dual tokens. If those controlled vocabularies contain stop words, e.g. "the", "a", "of", provided by the text mining (tm) package [21], they were removed. A total of 148 final NLP terms were compiled, and 52/148 were found to be significantly (Mann–Whitney test P value <0.05) associated with CHF.

The vital sign/comorbidity, including BMI, CHF standard markers, obesity, fasting blood glucose level, smoking history, and alcohol use status, can provide important cues of being with CHF. To compile the knowledge base that enabled structured information to be derived, a series of regular expressions representing the rules to enable information unification and design of different feature categories were compiled. As an example, BMI was presented directly in some notes, but could also be calculated based on height and weight. Therefore, BMI information was unified from two sources and normalized into four categories: underweight, normal, overweight and obesity according to the BMI classification of the World Health Organization (WHO) [22]. The blood pressure and fasting blood glucose levels were classified according to related standards from American Heart Association and American Diabetes Association, respectively [23,24].

Please cite this article in press as: Y. Wang, et al., NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records, Int. J. Med. Inform. (2015), http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007

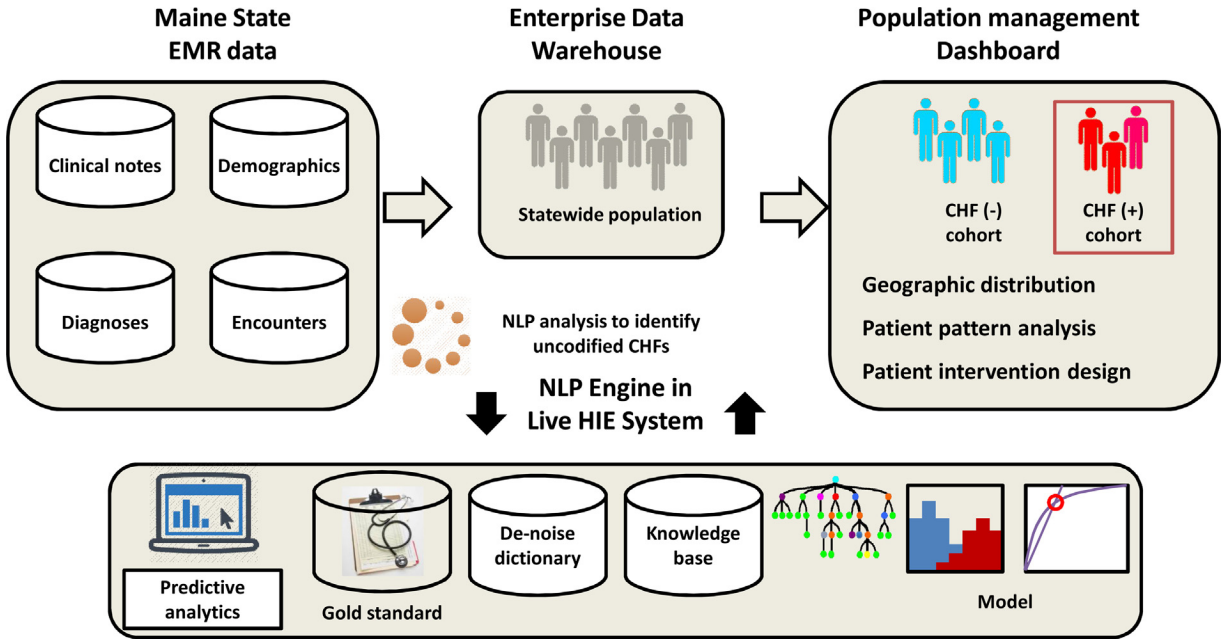

Fig. 1. A schematic of the NLP based algorithm integrated to the Maine HIE workflow to allow statewide CHF case finding and monitoring.

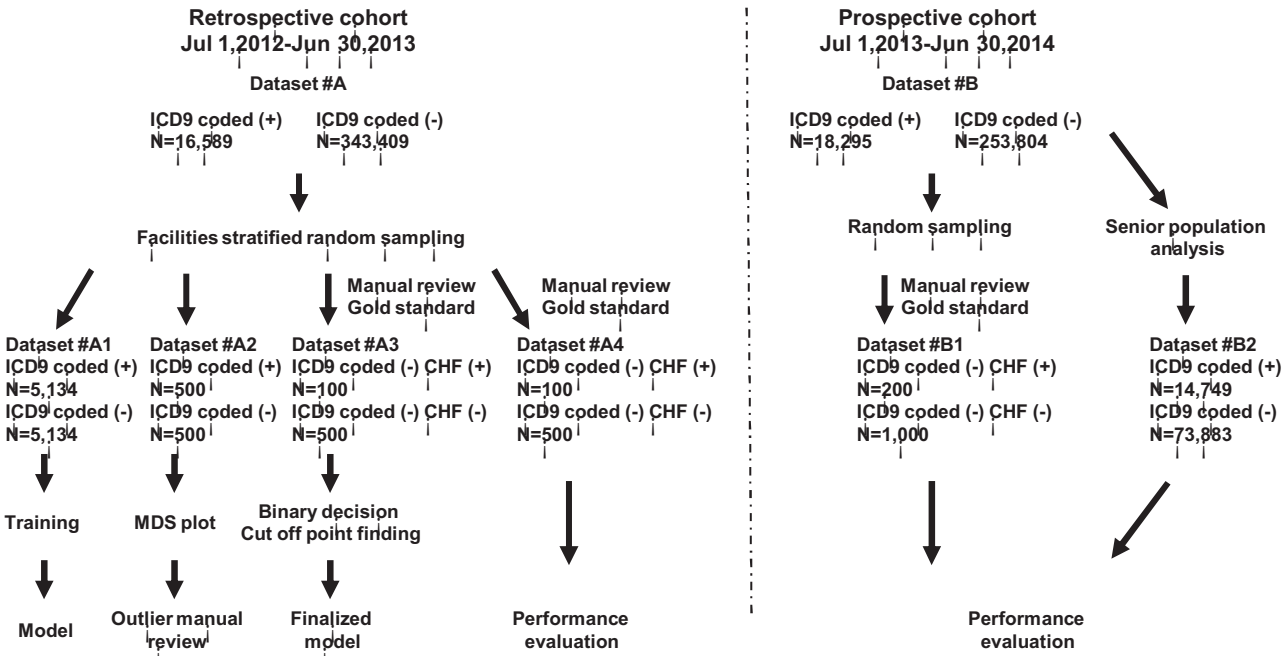

Fig. 2. Experimental design to train, validate and live system production of the NLP based CHF case finding algorithm. There were total of 8 datasets constructed through the retrospective and prospective analyses. Three manual-reviewed gold standard datasets, two in retrospective and one in prospective analyses, were utilized to gauge the model performance.

2.6. Clinical note preprocessing (Supplementary Fig. 2)

The clinical notes from different facilities were either in plain text or HTML format. To standardize, all HTML data were converted into plain text using XPath based matching methods implemented with the XML package [25].

To improve the NLP specificity, annotated terms associated with negation and family history were removed from notes. One method was to remove the words in a fixed interval centering on a negation or family history word [26]. This method, however, required manually defined parameters and is not adaptive. The following three steps were executed: (1) The text was collapsed into disjoint segments, which can be paragraphs, sentences or lines. If a paragraph (sentence) satisfied a criterion of being a segment, it would be regarded as one segment without any further decomposition. Otherwise, the paragraph (sentence) was divided into sentences (lines). The criterion was developed based on the segment length and the number of newline characters. The part-of-speech was annotated and referred for sentence boundary detection against the confusion between period and decimal point using openNLP [27]. (2) The de-noise dictionary of negation and family history was developed iteratively on the notes from retrospective cohort with NegEx lexicon and family member vocabulary as seeds [28,29]. (3) In each clinical note the segments containing any entries in the

Please cite this article in press as: Y. Wang, et al., NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records, Int. J. Med. Inform. (2015), http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007

de-noise dictionary were removed. The remaining segments were re-combined as a new clinical note, which was expected to have no negation and family history in its content.

2.7. Retrospective datasets

There were 1,257,952 clinical notes included in the retrospective cohort (July 1, 2012–June 30, 2013). All the notes were randomly partitioned, according to the patient associated clinical facilities, into two subsets: one for training, multidimensional scaling (MDS) plot and cutoff point finding (N = 228,220, 20 facilities), and the other for blind testing (N = 130,630, 15 facilities). In the subsets, the notes of the same patient were merged as one note, resulting in patient level subsets. Within the training, MDS plot and cutoff point finding subset, patients (N = 5134) with codified CHF diagnoses, and an equal number of uncodified patients (N = 5134), were randomly sampled to construct the training sub-cohort for model training (Fig. 2, dataset A1). Patients (N = 500) with codified CHF diagnoses and uncodified patients (N = 500) were randomly sampled as MDS plot sub-cohort (Fig. 2, dataset A2). In the remaining uncodified patients, a gold standard dataset was constructed through manual chart review by randomly selecting 100 positive (CHF cases) and 500 negative (non CHF cases) to build the cutoff-point finding sub-cohort (Fig. 2, dataset A3). 100 positive and 500 negative patients were randomly sampled from the blind testing subset to construct the blind testing sub-cohort (Fig. 2, dataset A4).

2.7. Model development

All the features (both from the structured database and unstructured clinical notes) related to a patient were concatenated into a feature vector denoted as f. The identification of CHF was stated as a maximum posterior probability (MAP) estimate problem:

$$\hat{\text{CHF}} = \underset{\text{CHF}}{\text{agrmax}} P(\text{CHF}|f)$$

where CHF was a binary random variable indicating whether the sample belonged to CHF cases (CHF = 1). To take diagnosis codes into consideration, the binary variable ICD was introduced to indicate whether a patient was codified (ICD = 1). By inserting ICD into the posterior and then applying the Bayesian rule, we have the decomposition:

$$P(\text{CHF}|f) = P(\text{CHF}|\text{ICD}=1,f)P(\text{ICD}=1|f) + P(\text{CHF}|\text{ICD}=0,f) \times P(\text{ICD}=0|f)$$

Since the assignment of diagnosis code was independent to the extracted feature, the model was simplified to:

$$P(\text{CHF}|f) = P(\text{CHF}|\text{ICD}=1,f)P(\text{ICD}=1) + P(\text{CHF}|\text{ICD}=0,f) \times P(\text{ICD}=0)$$

The first term on the right side determines the probability of CHF for a codified patient while the second term for an uncodified patient. As the coding information was known, we had two branches to obtain the posterior.

$$P(\text{CHF}|f) = \begin{cases} P(\text{CHF}|\text{ICD}=1,f) & \text{codified patient} \\ P(\text{CHF}|\text{ICD}=0,\text{f}) & \text{uncodified patient} \end{cases}$$

Our hypothesis was that the great majority of the patients without CHF codes were non-CHF cases while CHF codified patients were most likely CHF cases. This led to our class labeling method: (1) when a patient was codified, he/she should be assumed as a CHF case;

$$P(\text{CHF=1}|f) = P(\text{CHF=1}|\text{ICD}=1,f) = 1$$

(2) when a patient was not codified, a model T should be built to estimate the probability.

$$P(\text{CHF=1}|f) = P(\text{CHF=1}|\text{ICD}=0,f) = \text{T}(f)$$

For a codified patient, the inference of CHF case only required a database query, while for uncodified patient, we applied a random forest model [30,31] as T(f):

$$\text{T}(f) = \frac{1}{N}\sum_n t_n(f)$$

where t_n was the nth decision tree in the random forest.

At the perspective of hierarchical tree, the model of posterior can be seen as a binary determined tree at the top, of which the one child had a random forest at the bottom. The testing subjects were partitioned according to their associated ICD9 codes at the top level of the top determined tree. Decisions to label the uncodified patients were achieved at the bottom random forest using clinical parameters ranked by random forest importance scoring metric [30]. Therefore, our final model encapsulated a combination of human prior knowledge and the machine learned knowledge.

We hypothesized that uncodified but genuine CHF cases can be found based on the similar patterns within the clinical notes of both codified and uncodified but genuine CHFs. The model was trained using the training sub-cohort (Fig. 2, dataset A1). The false positives found during the training process were reviewed manually, so that (1) genuine CHF cases missing the CHF ICD-9 codes could be revealed, and (2) confusing words and phrases related to family history, and negation were identified. The two steps, including training and chart review, were executed iteratively to improve our knowledge base and fine-tune the model.

The multidimensional scaling (MDS) plots were constructed to visualize the analysis results (Fig. 2, dataset A2) of uncodified CHF cases (through chart review) and codified CHF case distributions.

2.8. Patient classification cutoff point determination

The receiver operating characteristic (ROC) curve analysis [32] was used to decide the optimal binary decision cutoff point (Fig. 2, dataset A3). Given that our algorithm assigned a classification probability to each subject, we set to find an optimal cutoff point to achieve the maximum classification sensitivity with a predefined positive predictive value (PPV) level of 90%. To achieve 90% PPV, the classification specificity can be calculated through a linear formula, thus forming a straight line overlaid on the ROC curve. The combination of sensitivity and specificity in the region above the line warranted the performance >90% PPV. Thus the cutoff point was set at the first intersection between the line and the ROC curve from the top to bottom.

2.9. Retrospective blind test

The model was blind tested on blind testing sub-cohort (Fig. 2, dataset A4), in which the retrospective patients were associated with care facilities independent from other sub-cohorts (Fig. 2, dataset A1–A3). The blind testing results demonstrated the model performance on an independent patient cohort, indicating that the knowledge from some hospitals could be leveraged to allow the prediction in others [33].

Please cite this article in press as: Y. Wang, et al., NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records, Int. J. Med. Inform. (2015), http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007

2.10. Prospective validation

Our NLP-based CHF case finding algorithm was integrated to the live HIE population exploration dashboard system. Therefore, a prospective validation of our NLP method was feasible.

The clinical notes (N = 881,347, dataset B) from patients between July 1, 2013 and June 30, 2014 were analyzed to find additional CHF cases. The algorithm performance was validated using an independent gold standard dataset of patients randomly selected from the prospective cohort. This gold dataset included chart-reviewed clinical notes of 1200 uncodified patients (200 positive and 1000 negative CHF subjects, Fig. 2, dataset B1). The prospective classification performance was evaluated using PPV, sensitivity, specificity, NPV (negative predictive value) and ROC AUC (area under curve). The senior population of age 65+ was analyzed (Fig. 2, dataset B2).

The prospective outcomes were visualized in groups of true positive, false positive, false negative and true negative. To understand the unique patterns associated with the false positive and true positive samples, these patients' notes were analyzed.

2.11. Learning transfer

In the retrospective study, we had two gold standard datasets, one for cutoff point finding and the other for blind-testing purposes. The cutoff point finding dataset was constructed by the notes from the same care facilities as the training dataset. The blind-testing dataset, on the other hand, was constructed by the notes from other care facilities independent from the ones used in the training and cutoff point finding dataset. The blind test demonstrated results similar to that in either the training (the notes of which was collected from different facilities from testing) or the prospective validation (the notes of which was collected from all facilities), indicating that the model learned from one group of facilities can be transferred to others.

3. Results

3.1. CHF discriminant variables

A total of 32CHF discriminant features were selected in the final model, including demographics (2), vital signs (2), comorbidity (4), clinical history (1), and NLP extracted clinical terms (23) (Fig. 3). The top four features, "heart failure", "congestive heart", "congestive heart failure" and "chf", were directly related to the CHF disease. "Age" ranked fourth, consistent with the notion that CHF hospitalization rate correlates significantly with patient aging [1]. The remaining discriminant risk factors included "glucose" and "blood pressure" (a proxy to the diabetes and hypertension comorbidities), smoking and alcohol histories, as well as BMI and "obesity", reflecting patient current physiological status in line with previous findings.

3.2. CHF case-finding algorithm development

With the CHF codified (positive) and uncodified (negative) patients' clinical note features, a binary classifier was developed for CHF case finding. A MDS plot was constructed to visualize the classification performance (Fig. 4), and 6/500 uncodified patients were classified as CHF cases. Specifically, the ratio (6/500, 1.2%) was the number of uncodified CHF cases, identified by the algorithm, over the total number of patients without CHF ICD9 codes when we sampled the training database. Thus, ratio of 1.2% could be very small due to the two possible reasons: (1) most of the true positive CHF cases had been codified; (2) CHF population prevalence is low. A close examination of other diagnosis (#2 and #5) and past histories (#1, #3, #4 and #6) in their clinical notes, however, revealed that these "false positive" patients were genuine CHF cases. Such results indicated: (1) our model effectively separated codified CHF patients from uncodified ones; (2) our initial hypothesis, that uncodified CHF cases can be identified by NLP profiling of the clinical notes, was validated.

3.3. Patient classification cutoff point determination

The decision tree based classification scores were evaluated to achieve the maximum sensitivity while the PPV was higher than 90% (Fig. 5). With this cutoff value of the classification probabilities set as 0.864, the continuous classification scoring outputs were used to reach a binary decision to determine the genuine CHF cases.

3.4. Inclusion of CHF standard markers

Due to the high availability of blood pressure information in the clinical notes, our study subjects were categorized as normal blood pressure, pre-hypertension, and hypertension. The addition of these features improved our algorithm performance (F-measure) from 0.729 to 0.753 in the prospective analysis.

In addition, other CHF makers including brain natriuretic peptide (BNP) [34], coronary artery disease (CAD), ejection fraction (EF) [35], pulmonary edema, pleural effusion, dyspnea on exertion, edema, exercise intolerance, paroxysmal nocturnal dyspnea or elevated jugular venous pressure are very sparse in either the codified EMR data or the clinical notes, limiting their contributions. The inclusion of additional CHF markers did not significantly improve our CHF case finding algorithm. (Supplementary Fig. 3).

3.5. Retrospective blind testing

As shown in Fig. 6A and B summarizing the retrospective blind testing, our NLP analytics achieved a PPV of 0.920 (69/75), sensitivity of 0.690 (69/100), specificity of 0.988 (494/500) and NPV of 0.941 (494/525). The ROC AUC score was 0.886. The F-measure was 0.789 (It was computed using PPV and sensitivity which were equivalent to precision and recall). These results demonstrated the learning can be transferred to different hospitals.

3.6. Case finding algorithm performance evaluation

Since the algorithm was deployed to allow real time CHF case finding, a prospective validation was feasible. We evaluated the method performance (Table 1) using retrospective MDS plot sub-cohort (Fig. 2, dataset A2), prospective cohort (Fig. 2, dataset B), and manual chart-reviewed cohort (Fig. 2, dataset B1), and prospective senior population (Fig. 2, dataset B2). In MDS plot sub-cohort, there were 500 codified and 500 uncodified patients. Among the uncodified patients, 6 patients were found by NLP, resulting in a percentage of 1.2%. In the total prospective population (dataset B), there were 18,295 codified CHF patients, and 253,804 patients without CHF codings. Our NLP case finding algorithm asserted 2411 uncodified CHF cases, resulting in percentages of NLP found uncodified CHF patients of 0.95%.

In the total prospective senior population, there were 14,749 codified CHF patients, and 73,883 patients without CHF codings. Our NLP analysis (Table 1) found additional 1814 uncodified CHF cases, which is 2.4% of the 73,883 patients with no CHF codings. Our results concluded that 80.6% codified and 75.2% NLP-revealed uncodified CHF cases were senior patients. The prospective performance, summarized in Fig. 6A and B, of the algorithm was

Please cite this article in press as: Y. Wang, et al., NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records, Int. J. Med. Inform. (2015), http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007

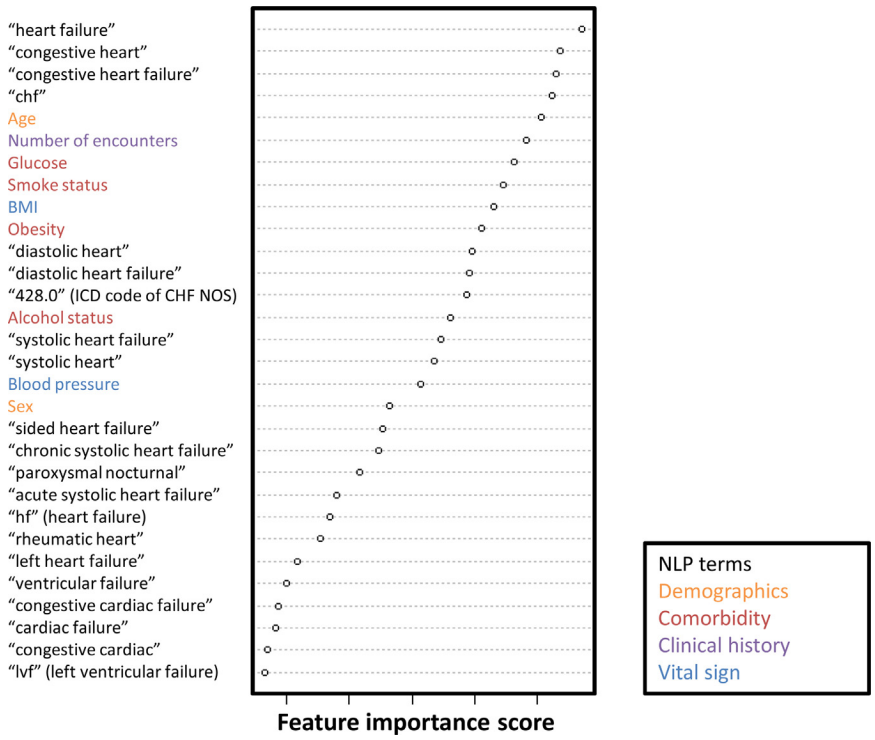

Fig. 3. Importance of the CHF NLP model clinical variables in descending order. The importance was defined as the mean decrease in algorithm accuracy scaled by standard deviation after randomly permuting the variable values, thus, a higher mean decrease in accuracy suggesting greater importance for the variable. The model was derived with retrospective data set A1 (as shown in Fig. 1).

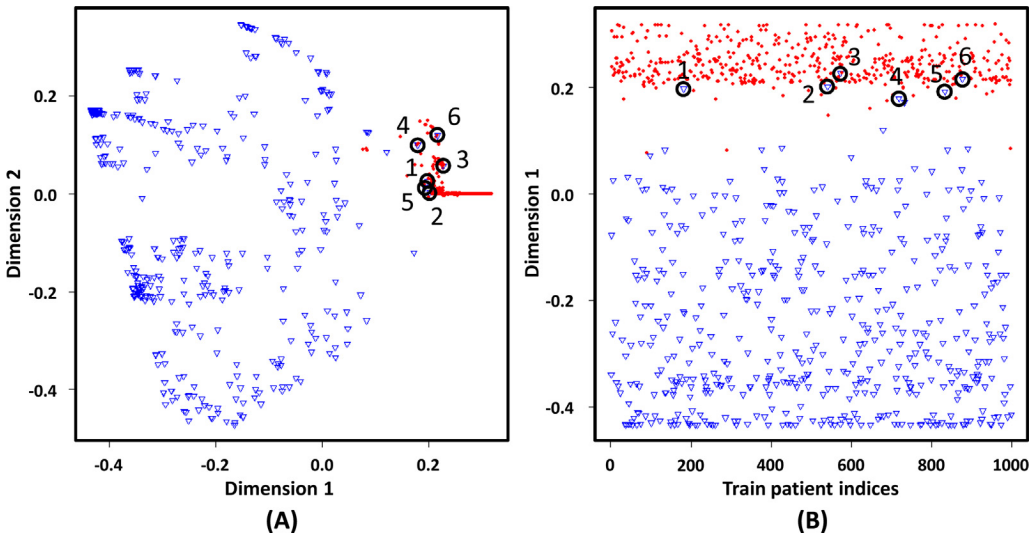

Fig. 4. The multidimensional scaling (MDS) plots of the training result. This analysis is to detect meaningful underlying dimensions, e.g., #1 and #2, that allow the explanation of the observed similarities (distances) between the investigated subjects. The axes of the MDS plots represent no real sizes and thus were marked as Dimension 1 and Dimension 2 without units. The red dots and blue triangles, indicating codified and uncodified patients, were clearly separated. The "false positives" were circled in the plot. Manual chart review confirmed true positive CHF cases. The plot was derived with retrospective data set A2 (as shown in Fig. 1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
NLP results from the test/validation analyses and the live system.

Dataset	ICD9 coded (+)	ICD9 coded (−)	NLP found uncodified CHFs	% NLP found uncodified CHFs*
Retrospective dataset A2	500	500	6	1.2%
Prospective dataset B	18,295	253,804	2,411	0.95%
Prospective dataset B2 Senior (Age 65+)	14,749	73,883	1,814	2.4%

* *t*-Test revealed no significant difference between the analysis results of dataset A2 and live system data set B.

Please cite this article in press as: Y. Wang, et al., NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records, Int. J. Med. Inform. (2015), http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007

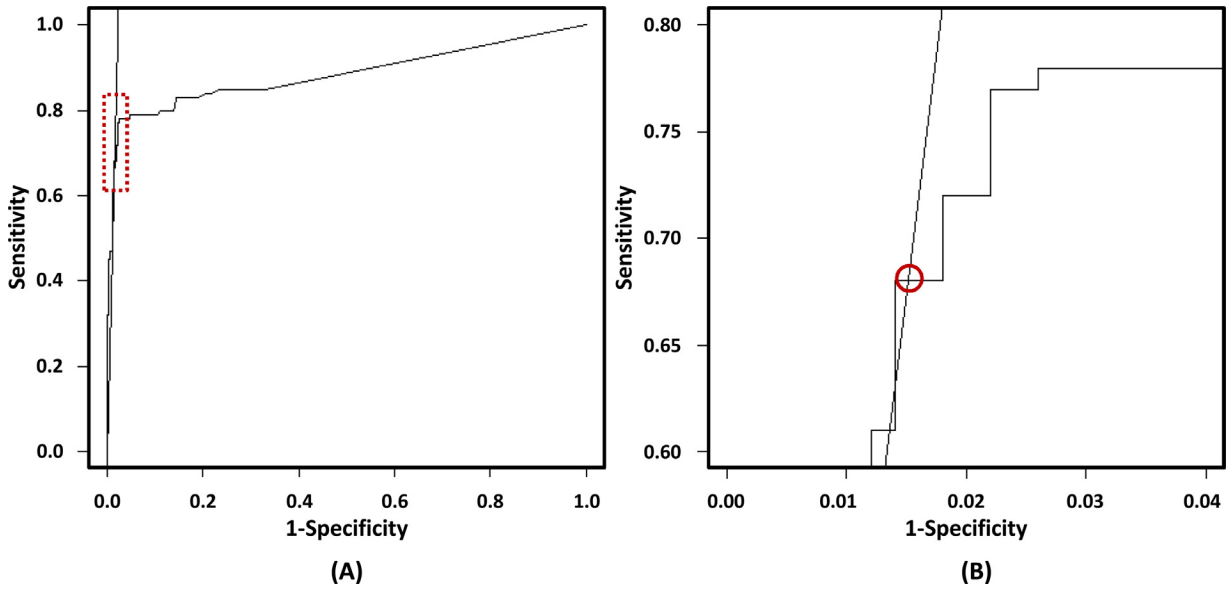

Fig. 5. Cutoff point determination to achieve optimal binary decisions. The model was finalized with retrospective gold standard data set A3 (as shown in Fig. 2). (A) The ROC curve and the line determined by the prevalence and 90% PPV were intersected at the cutoff point. (B) Their intersection (dashed rectangle) was zoomed and is indicated by the circle.

(A)

	Retrospective (ICD9 coded (-) dataset A4)		Prospective (ICD9 coded (-) dataset B1)	
	CHF (−)	CHF (+)	CHF (−)	CHF (+)
Identified as CHF (−)	494	31	988	72
Identified as CHF (+)	6	69	12	128

(B)

Cohort	PPV	Sensitivity	Specificity	NPV
Retrospective	0.920	0.690	0.988	0.941
95% CI	0.836 – 0.963	0.594 – 0.772	0.974 – 0.994	0. 917 – 0.958
Prospective	0.914	0.640	0.988	0.932
95% CI	0.856 – 0.950	0.571 – 0.703	0.979 – 0.993	0.915 – 0.946

Fig. 6. The CHF case finding performance analyses. (A) The contingency tables on blind test sub-cohort and prospective cohort. (B) The PPV, sensitivity, specificity and NPV values on the retrospective blind test sub-cohort and prospective cohort. The model was evaluated with retrospective gold standard data set A4 (as shown in Fig. 1).

validated by chart review of a randomly selected uncodified sub-cohort (Fig. 2, dataset B1). The PPV was 0.914 (128/140), which was within the 95% confidence interval of the retrospective blind-testing PPV (0.836–0.963). The sensitivity was 0.640 (128/200). The specificity, NPV and AUC were 0.988, 0.932 and 0.919, respectively. The prospective *F*-measure was 0.753.

Close examination (Fig. 7) of the prospective uncodified patient validations categorized the NLP false positives ($N = 12$): risks of CHF ($N = 2$, Supplementary Fig. 4 labeled #1 patient), ambiguity of pronouns ($N = 3$, Supplementary Fig. 4 labeled #2 patient), mistyped information ($N = 2$, Supplementary Fig. 4 labeled #3 patient) and undetermined CHF ($N = 5$, Supplementary Fig. 4 labeled #4 patient). Integration of correction plans of these error sources into our knowledge base will iteratively improve our NLP analytics to deliver solutions with enhanced performance.

4. Discussion

In this study, we retrospectively developed and prospectively validated a NLP based CHF case finding algorithm. Clinical notes from a specific set of randomly chosen hospitals were profiled and the learned results were successfully reproduced in other facilities. In the retrospective training set, the algorithm found 6CHF cases from 500 uncodified patients (Table 1). In prospective cohort, the algorithm found 2411 CHF cases from 253,804 cases without CHF codings. The two results (6/500 and 2411/253,804, or 1.20% and 0.95%) were not significantly different (*t*-test *P*-value 0.61). The high *P*-value, together with the result obtained from the gold standard datasets (that performance in prospective gold standard dataset was within the 95% CI of that in the retrospective gold standard dataset), demonstrated that the proposed case finding

Please cite this article in press as: Y. Wang, et al., NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records, Int. J. Med. Inform. (2015), http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007

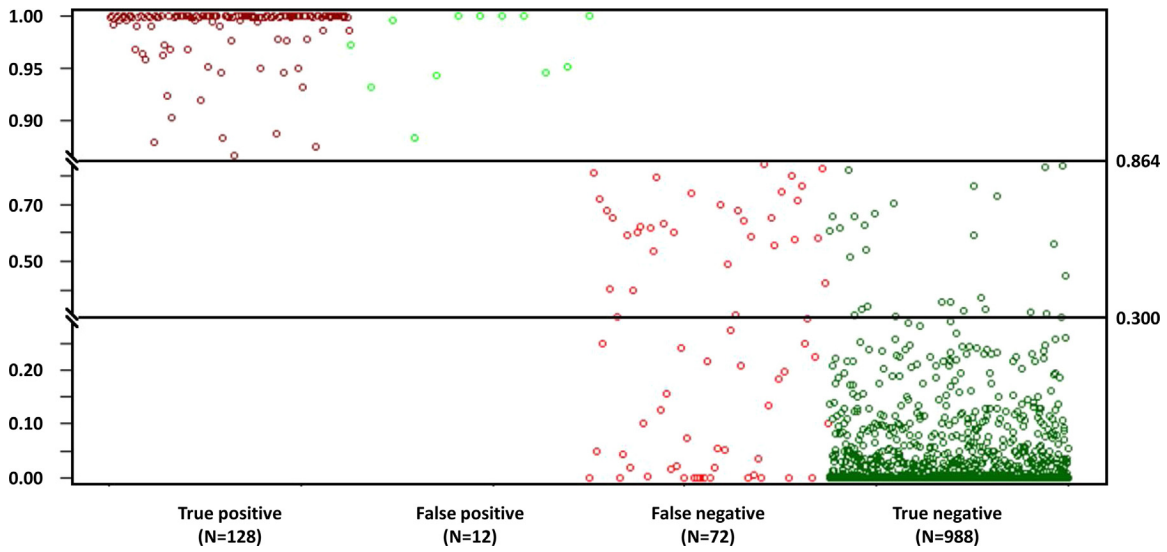

Fig. 7. Visualization of the model outcome. The random forest classification probability (*y* axis) of true positive, false positive, false negative, and true negative were demonstrated in separate regions from left to right, marked with dark red, red, green and dark green, respectively. The *x* axis is the sample indices. The analysis results were visualized with prospective gold standard data set B1 (as shown in Fig. 2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

algorithm yielded the consistent performances in both retrospective and prospective cohorts. These findings strongly endorsed our hypothesis that uncodified but genuine CHF cases can be revealed based on the similar narrative text patterns within the clinical notes of both codified and uncodified but genuine CHFs.

Compared with previous studies: (1) A naïve Bayes model using full text features with a *F*-measure of 0.666[15], (2) A naïve Bayes model using both full text features and clinical terms with a *F*-measure of 0.637[16], and (3) A perceptron model using full text features and clinical terms with a *F*-measure of 0.666, our algorithm outperformed these previous results with a *F*-measure of 0.753. This success was attributable to our comprehensive knowledge base developed using all hospital EMR contents of the State of Maine as well as the use of standard clinical terminologies and synonyms that provided the capacity to reduce the disturbance from negation and family history. The retrospective analysis and prospective integration of the NLP process to the HIE workflow facilitated the close to real time iterative optimization in further support of the observed superior performance.

Our CHF case-finding algorithm demonstrated that the vast amount of narrative clinical note information could be effectively utilized to construct disease cohorts, supplementing the codified EMR dataset. More importantly, the developed case finding framework, of knowledge base as well as preprocessing and statistical modeling analyses, can be repeated as a generalized NLP solution to construct any targeted disease cohort for population exploration. Specifically, by constructing other disease NLP knowledge bases in similar fashion as this study, our approach can be applied to case findings of other disease groups.

Standard CHF marker information, including CAD, BNP values, ejection fraction, pulmonary edema, pleural effusion, were extracted and explored. Most of these clinical parameters are very sparse in either the codified EMR data or the clinical notes, therefore, are of limited utility for our case finding analytics. Inclusion of these parameters did not add any value to the current model in CHF case finding, and the *F*-score was reduced to 0.641 (PPV 0.893 and sensitivity of 0.500) comparing to the current model described in this work. Therefore, these CHF markers were not included by our current NLP workflow. Future effort to comprehensively collect these features might improve our CHF case findings.

Our case finding algorithm aims to be part of the live HIE system to construct CHF cohort. The engine has the capacity to analyze, within 12-h time frame, 881,347 notes of 2014 notes in the Maine HIE. Although the analytics utilized ICD codes which are often assigned post discharge, automated tracking and update of CHF patients' registry records can help the health care providers, under the accountable care organization (ACO) setting, to provide targeted care, therefore, allowing proactive CHF patient intervention.

Among the uncodified CHF cases, approximately 70% of the patients had CHF histories described in the clinical notes. Those patients, however, were not assigned with CHF diagnosis codes since CHF diagnoses might not be the clinical focus of relevant encounters in the database [36]. This observation indicated that under current coding guidelines, those patients missing CHF codings would not benefit from CHF targeted healthcare programs. In addition, the underlying issues, of both the apparent failure of the conventional coding methods and the flaws of the current coding system, need to be explored so that electronic solutions may be formulated.

Compared to patients currently identified as having CHF, uncodified patients with CHF histories may not have current acute symptoms or be taking any CHF related medications. This likely leads to a low level of awareness by both patients and care providers. However, these patients shall still possess risk factors that should be reviewed at a regular interval. Our CHF case finding algorithm presents an opportunity for individual providers to identify previously uncodified patients for proactive care interventions. Healthcare providers may be able to better manage all CHF patients through medication compliance, medication related adverse events, diet and exercise programs, along with the timely management of any emerging disease risks. As a result, these efforts may further reduce preventable CHF admission and readmission rates [5], and ultimately improve overall healthcare outcomes and patient quality of life.

5. Conclusion

A NLP based CHF case finding algorithm was developed and integrated in HIE live system across all demographic groups in the State of Maine. The NLP modeling results were validated with a cohort

Please cite this article in press as: Y. Wang, et al., NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records, Int. J. Med. Inform. (2015), http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007

from hospitals independent from the training sets, indicating the transfer of learning that is achievable between different hospitals and practitioners. The prospective cohort validation demonstrated the temporal effectiveness of the developed algorithm, supporting deployment in an existing and active HIE. Real time integration of CHF analytics into the Maine HIE workflow may empower population health exploration and specifically the monitoring of subjects at risk for CHF, providing an opportunity for proactive and preventive interventions. While HIE data represents an ideal source of clinical narrative notes for NLP analysis, operational HIEs are not present in all States. Our NLP based CHF finding algorithm can be applied to any clinical EMRs directly as well as private HIEs within hospital networks. Beyond the case finding for proactive CHF care, gaining a deeper understating of both the unique and common attributes of various sub-groups may further facilitate overall CHF management.

Conflict of interest

KGS, EW and XBL are co-founders and equity holders of HBI Solutions, Inc., which is currently developing predictive analytics solutions in health care. The other authors declare that they have no competing interests. From Stanford University School of Medicine, Stanford, California, YW, JL, SH, AYS, RL, LZ, Yingzhen Z, YH, YJ KGS, XBL conducted this research as part of a personal outside consulting arrangement with HBI Solutions, Inc.

Author contributions

BJ, XBL,YW, SH, DSC, STA, TR, KGS, and EW: Conceived and designed the experiments. YW, HX, JL, LW, Yifan Zhao, CZ, ZH,AYS, XBL, KGS, and EW: Analyzed the data. YW, JL, SH, KGS, AYS, and XBL: Wrote the paper. BJ, HX, ZH, CF, YH, Yingzhen Zhao, YJ, CZ, BJ, XBL, KGS, and EW: Acquired the data. XBL, SH, HX, RL, LZ, KGS, AYS, STA, FS, and EW: Critically revised the manuscript for important intellectual content. XBL: Provided statistical expertise. JL and LW: Provided chart review. XBL, SH, HX, RL, XD, LZ, KGS, AYS, STA, FS, and EW: Critically revised the manuscript for important intellectual content BJ, CZ, HZ, CF, DD, XBL, KGS, FS, and EW: Provided administrative, technical, or material support. XBL, KGS, FS, and EW: Supervised the study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007

References

[1] Hospitalization for congestive heart failure: United States, 2000–2010. Centers for Disease Control and Prevention Website. <http://www.cdc.gov/nchs/data/databriefs/db108.pdf/> (accessed 16.01.15).
[2] A.S. Go, D. Mozaffarian, V.L. Roger, et al., Heart disease and stroke statistics – 2013 update: a report from the American Heart Association, Circulation 127 (1) (2013) e6–e245.
[3] P.A. Heidenreich, N.M. Albert, L.A. Allen, et al., Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association, Circ. Heart Fail. 6 (3) (2013) 606–619.
[4] Readmission reduction program. Centers for Medicare and Medicaid Services Website. <http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html/> (accessed 16.01.15).
[5] A.B. Bindman, K. Grumbach, D. Osmond, M. Komaromy, K. Vranizan, N. Lurie, J. Billings, A. Stewart, Preventable hospitalizations and access to health care, JAMA 274 (4) (1995) 305–311.
[6] G.C. Fonarow, Heart failure disease management programs: not a class effect, Circulation 110 (23) (2004) 3506–3508.
[7] D.E. Feldman, C. Thivierge, L. Guérard, V. Déry, C. Kapetanakis, G. Lavioie, E. Beck, Changing trends in mortality and admissions to hospital for elderly patients with congestive heart failure in Montreal, CMAJ 165 (8) (2001) 1033–1036.
[8] M. Onofrei, J. Hunt, J. Siemienczuk, D.R. Touchette, B. Middleton, A first step towards translating evidence into practice: heart failure in a community practice-based research network, Inform. Prim. Care 12 (3) (2004) 139–145.
[9] S.M. Meystre, G.K. Savova, K.C. Kipper-schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, Yearb. Med. Inform. 12 (2008) 8–44.
[10] R.J. Byrd, S.R. Steinhubl, J. Sun, S. Ebadollahi, W.F. Stewart, Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records, Int. J. Med. Inform. 83 (12) (2014) 983–992.
[11] M. Fiszman, J.H. Peter, Using medical language processing to support real-time evaluation of pneumonia guidelines, in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2000.
[12] E.A. Mendonça, J. Haas, L. Shagina, E. Larson, C. Friedman, Extracting information on pneumonia in infants using natural language processing of radiology reports, J. Biomed. Inform. 38 (4) (2005) 314–321.
[13] S. Dublin, E. Baldwin, R.L. Walker, L.M. Christensen, P.J. Haug, M.L. Jackson, J.C. Nelson, J. Ferraro, D. Carrell, W.W. Chapman, Natural language processing to identify pneumonia from radiology reports, Pharmacoepidemiol. Drug Saf. 22 (8) (2013) 834–841.
[14] T.A. Holt, C.L. Gunnarsson, P.A. Cload, S.D. Ross, Identification of undiagnosed diabetes and quality of diabetes care in the United States: cross-sectional study of 11. 5 million primary care electronic records, CMAJ Open 2 (4) (2014) E248–E255.
[15] S. Pakhomov, S.A. Weston, S.J. Jacobsen, C.G. Chute, R. Meverden, V.L. Roger, Electronic medical records for clinical research: application to the identification of heart failure, Am. J. Manag. Care 13 (6 Part 1) (2007) 281–288.
[16] S.V. Pakhomov, J. Buntrock, C.G. Chute, Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier, J. Biomed. Inform. 38 (2) (2005) 145–153.
[17] Clinical classifications software for ICD-9-CM. Healthcare Cost and Utilization Project Website. <http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp#overview/> (accessed 16.01.15).
[18] T.Y. Kim, N. Hardiker, A. Coenen, Inter-terminology mapping of nursing problems, J. Biomed. Inform. 49 (2014) 213–220.
[19] P.M. Nadkarni, J.A. Darer, Migrating existing clinical content from ICD-9 to SNOMED, J. Am. Med. Inform. Assoc. 17 (5) (2010) 602–607.
[20] Medical subject headings. National Center for Biotechnology Information Website. <http://www.ncbi.nlm.nih.gov/mesh/> (accessed 16.01.15).
[21] Tm R. package. <http://cran.r-project.org/web/packages/tm/tm.pdf/> (accessed 16.01.15).
[22] BMI classification. World Health Organization Website. <http://apps.who.int/bmi/index.jsp?introPage=intro_3.html/> (accessed 16.01.15).
[23] <http://www.heart.org/HEARTORG/Conditions/HighBloodPressure/AboutHighBloodPressure/Understanding-Blood-Pressure-Readings_UCM_301764_Article.jsp/> (accessed 07.04.15).
[24] <http://www.diabetes.org/diabetes-basics/diagnosis/> (accessed 07.04.15).
[25] XML R. package. <http://cran.r-project.org/web/packages/XML/XML.pdf/> (accessed 16.01.15).
[26] J.A. Strauss, C.R. Chao, M.L. Kwan, S.A. Ahmed, J.E. Schottinger, V.P. Quinn, Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm, J. Am. Med. Inform. Assoc. 20 (2) (2013) 349–355.
[27] OpenNLP package. <http://cran.r-project.org/web/packages/openNLP/openNLP.pdf/> (accessed 16.01.15).
[28] <https://code.google.com/p/negex/downloads/detail?name=medinfo_2013_multilingual_negex_lexicon_v1_April30th2013.zip&can=2&q=/> (accessed 07.04.15).
[29] <http://www.vocabulary.cl/english/family-members.html/> (accessed 0.04.15).
[30] L. Breiman, Random forest, Mach. Learn. 45 (1) (2001) 5–32.
[31] Random forest R package. <http://cran.r-project.org/web/packages/randomForest/randomForest.pdf/> (accessed 16.01.15).
[32] AUC R. package. <http://cran.r-project.org/web/packages/AUC/AUC.pdf/> (accessed 16.01.15).
[33] J. Wiens, J. Guttag, E. Horvitz, A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions, J. Am. Med. Inform. Assoc. 21 (4) (2014) 699–706.
[34] A. Maisel, B-Type natriuretic peptide levels: diagnostic and prognostic in congestive heart failure: what's next, Circulation 105 (2002) 2328–2331.
[35] M. Maeder, D. Kaye, Heart failure with normal left ventricular ejection, J. Am. Coll. Cardiol. 53 (11) (2009) 905–918.
[36] ICD-9-CM coding. <http://nursing.advanceweb.com/article/icd-9-cm-coding.aspx?CP=2/> (accessed 29.01.15).

Please cite this article in press as: Y. Wang, et al., NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records, Int. J. Med. Inform. (2015), http://dx.doi.org/10.1016/j.ijmedinf.2015.06.007