Electronic Medical Records for Clinical Research: Application to the Identification of Heart Failure

Serguei Pakhomov, PhD; Susan A. Weston, MS; Steven J. Jacobsen, MD, PhD;
Christopher G. Chute, MD, DrPH; Ryan Meverden, BS; and Véronique L. Roger, MD, MPH

The electronic medical record (EMR) is increasingly used in healthcare.[1] Its clinical goals include streamlining clinical practice and improving patient safety. In addition to improving practice, the EMR offers promising methods for identification of potential study participants, which is essential for clinical research. Indeed, although the use of manually coded patient records in clinical research is a long-standing tradition,[2,3] these methods must allow for a delay between the diagnosis and the assignment of the code. In addition, coding systems have variable yields in identifying patients, depending on the disease under consideration, and are subject to shifts related to changing reimbursement incentives.[4] Use of coding systems to identify patients appears particularly problematic for heart failure (HF) because of its syndromic nature, which precludes its ascertainment from a single diagnostic test.[5,6] The EMR may enable efficient case identification by providing access to clinical reports as soon as they become transcribed; however, novel methods of identification that use the EMR require rigorous validation.[5]

Finding patient records that meet predefined clinical criteria lends itself well to statistical classification algorithms.[7-14] Natural language processing (NLP) systems such as the Medical Language Extraction and Encoding System have been used to identify cases of interest either directly by defining a terminologic profile of a case or indirectly by extracting covariates for predictive modeling.[13,15,16] To our knowledge, there have been no large-scale studies that examined the validity of both NLP and statistical methods for identification of patients with HF.

We report here on use of the EMR that currently is in place at the Mayo Clinic[17] for prospective recruitment of patients with HF.[18] The goal of our study was to validate 2 approaches to rapid prospective identification of patients with HF. One approach uses NLP of the EMR; the other uses predictive modeling.

Objective: To identify patients with heart failure (HF) by using language contained in the electronic medical record (EMR).

Methods: We validated 2 methods of identifying HF through the EMR, which offers transcription of clinical notes within 24 hours or less of the encounter. The first method was natural language processing (NLP) of the EMR text. The second method was predictive modeling based on machine learning, using the text of clinical reports. Natural language processing was compared with both manual record review and billing records. Predictive modeling was compared with manual record review.
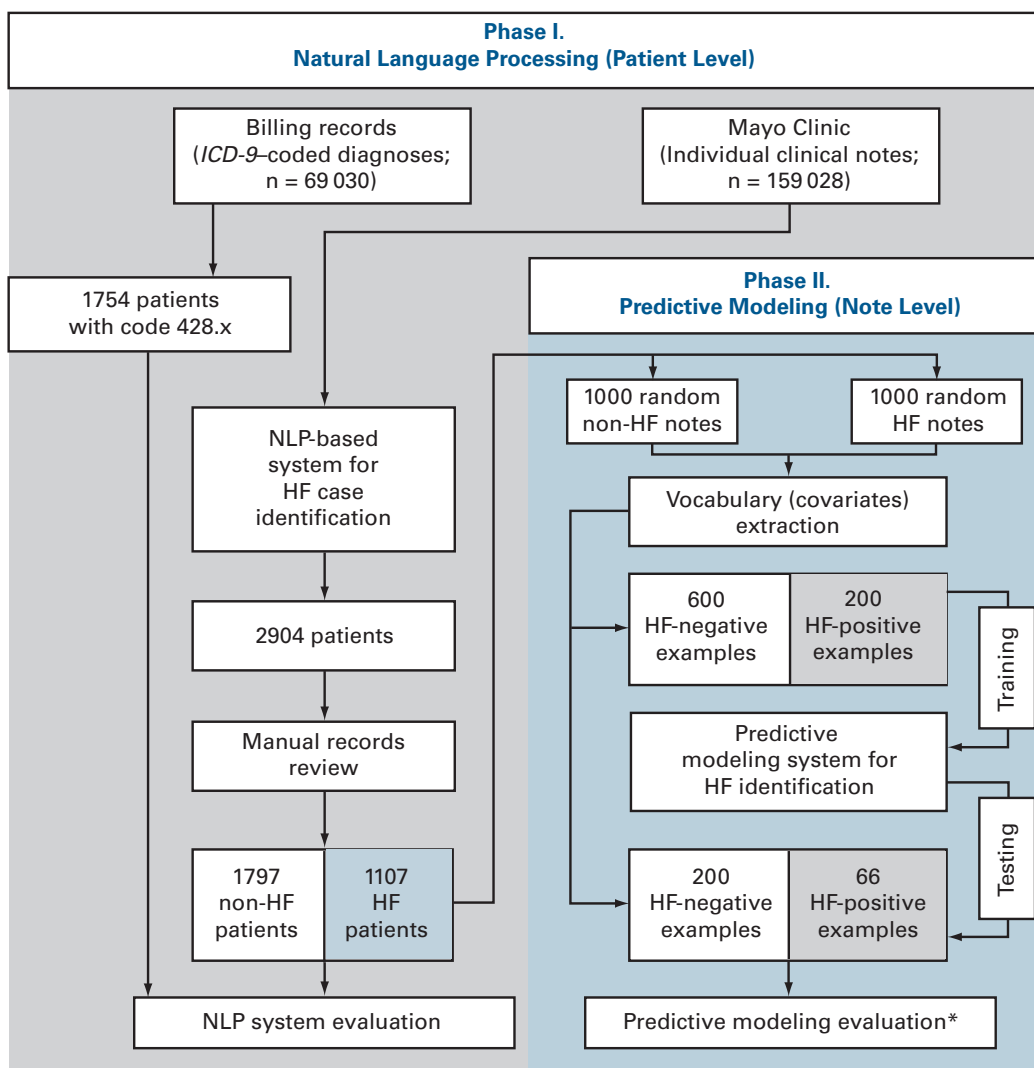
Results: Natural language processing identified 2904 HF cases; billing records independently identified 1684 HF cases, 252 (15%) of them not identified by NLP. Review of a random sample of these 252 cases did not identify HF, yielding 100% sensitivity (95% confidence interval [CI] = 86, 100) and 97.8% specificity (95% CI = 97.7, 97.9) for NLP. Manual review confirmed 1107 of the 2904 cases identified by NLP, yielding a positive predictive value (PPV) of 38% (95% CI = 36, 40). Predictive modeling yielded a PPV of 82% (95% CI = 73,93), 56% sensitivity (95% CI = 46, 67), and 96% specificity (95% CI = 94, 99).

Conclusions: The EMR can be used to identify HF via 2 complementary approaches. Natural language processing may be more suitable for studies requiring highest sensitivity, whereas predictive modeling may be more suitable for studies requiring higher PPV.

(Am J Manag Care. 2007;13(part 1):281-288)

METHODS

The study design—including the data sources, processing components, data flow, and evaluation—is shown in **Figure 1**.

Mayo Clinic Electronic Medical Record

For this study, we used 2 data sources available as part of the Mayo Clinic EMR: clinical notes and diagnostic codes.

For author information and disclosures, see end of text.

■ **Figure 1.** Study Design and Methodology

*Predictive modeling was evaluated by repeating all steps in Phase II 1000 times to estimate 95% confidence interval for sensitivity, specificity, and positive predictive modeling.
ICD-9 indicates *Classification of Diseases, Ninth Revision*; EMR, electronic medical record; NLP, natural language processing; HF, heart failure.
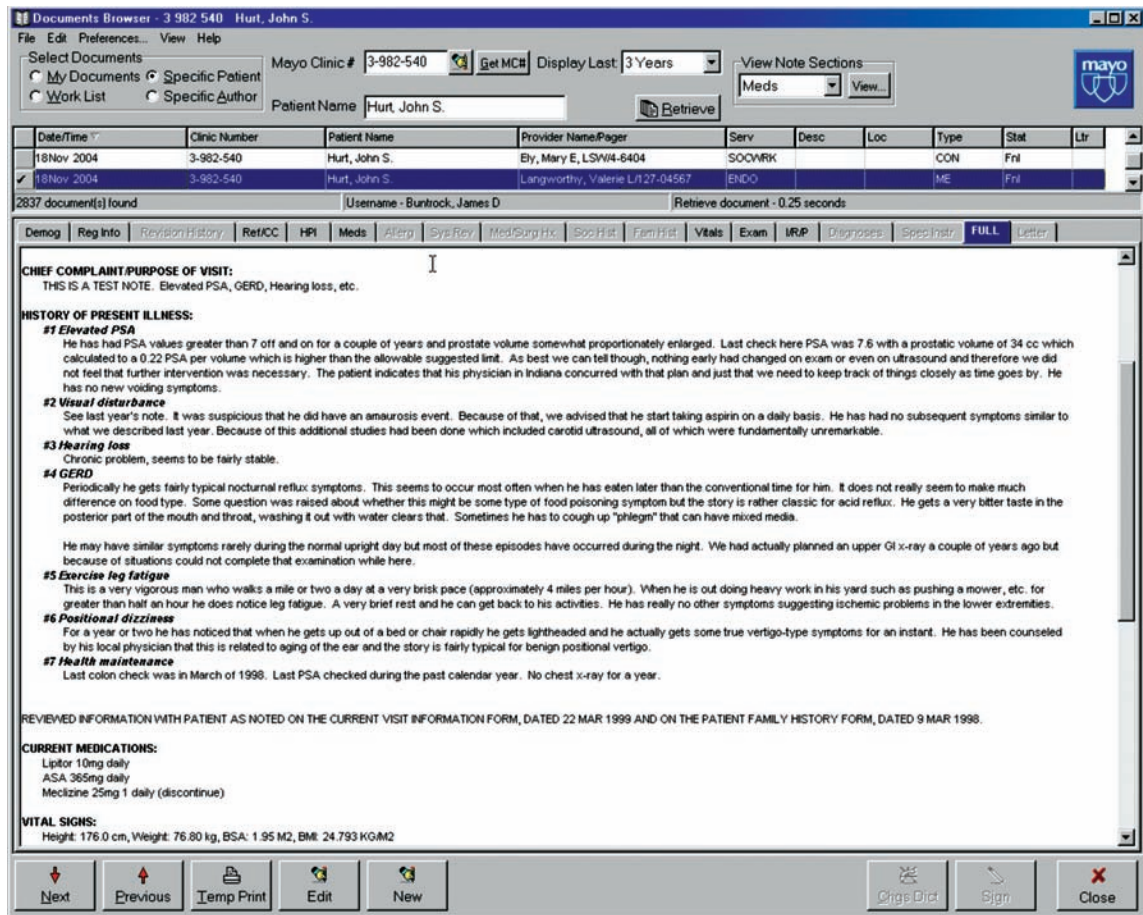
Clinical Notes. Clinical notes dictated by healthcare providers at the Mayo Clinic first became available electronically in 1994. These are electronic records that document each inpatient and outpatient encounter, and contain the text of the medical dictations transcribed by trained medical transcriptionists (for an example, see **Figure 2**). The Mayo Clinic EMR complies with the American National Standards Institute Clinical Document Architecture, which is a widely accepted standard for clinical documentation.[19] Most of the Mayo Clinic clinical notes are transcribed within 24 hours of the patient-physician encounter.

Diagnostic Codes. Patient-physician encounters are coded using *International Classification of Diseases, Ninth Revision (ICD-9)* diagnostic codes. The codes are assigned by trained medical coders as part of the routine billing process within 30 days of the encounter.

Use of Natural Language Processing

The NLP case-finding algorithm was piloted in October 2003.[20] The algorithm uses nonnegated terms indicative of HF: cardiomyopathy, heart failure, congestive heart failure, pulmonary edema, decompensated heart failure, volume over-

■ **Figure 2.** Example of a Clinical Note

This is an ambulatory care note for a nonexistent test patient. For this study, all available inpatient and outpatient notes were used.

load, and fluid overload. To maximize sensitivity, all available synonyms (n = 426) for these terms were used as well. The synonyms were found by automatically searching a database of 16 million problem-list entries comprised of diagnostic phrases expressed in natural language. These phrases are manually coded by trained staff as part of the Rochester Epidemiology Project,[3] using a hospital adaptation of the *International Classification of Diseases*.[21] Diagnostic phrases were considered synonymous if they were assigned the same code (eg, phrases such as heart failure, CHF [congestive heart failure], biventricular failure, and cardiopulmonary arrest were treated as synonymous).[22] In addition to synonyms, the NLP algorithm relies on finding nonnegated terms by excluding those terms that have negation indicators (eg, "no," "denies," "unlikely") in their immediate context (±7 words). In order to identify potential cases of HF, the algorithm searched for the terms indicative of HF and their synonyms in the text of clinical notes as soon as the notes were dictated, transcribed, and

became available electronically. Once a term was found, a determination with respect to its negation status was made. If this particular instance of the term was negated, it was ignored for the purposes of identification of evidence of HF in the clinical note. However, the note was identified as containing evidence of HF if another instance of the same term was found in a nonnegated context. The algorithm was implemented in Perl programming language as an application that runs inside a JBoss Application Server. The application continually "listened" to the live stream of clinical notes that are generated within the Mayo Integrated Clinical Systems production environment.

After the pilot, we conducted periodic verifications of the method to ensure that no patients with HF were being omitted by comparing the results of the algorithm with the billing codes. We extracted all unique patient identifiers using *ICD-9* code 428.x (heart failure) for the period between October 10, 2003, and May 31, 2005, and compared them

■ **Figure 3.** Example of a Clinical Note Represented as a "Bag of Words"

Note ID	HF status	Covariate #1 "heart"	Covariate #2 "pulmonary"	. . .	Covariate #10179 "laceration"
45893484-02	positive	3	2		0
34695234-01	negative	1	0		0

The values in the cells represent the frequency of the covariate in the clinical note and are used as weights during predictive modeling.
HF indicates heart failure.

with the patients identified by the NLP system. All cases found by the NLP system since October 2003 were reviewed by nurse abstractors for HF criteria as part of the ongoing Heart Failure Surveillance Study.[23] The results of this manual review were used for validation of the NLP method (Figure 1, Phase I).

Use of Predictive Modeling

Predictive Modeling Algorithm. Prior studies have reported on using predictive modeling techniques including logistic regression, classification trees, and neural networks for clinical decision support and outcomes prediction.[24,25] A comparative validation between these 3 approaches shows that logistic regression outperforms the other methods.[26] Although traditional logistic regression relies on small sets of well-defined clinical covariates, predictive modeling based on the text of clinical notes involves an unlimited number of predictive covariates based on the vocabulary of the clinical notes and may include more than 10 000 items whose relative contribution to the categorization decisions is unknown. Thus, large-scale predictive modeling based on clinical notes requires algorithms specifically designed to process large numbers of covariates. Naïve Bayes[27] is one such approach that is robust, highly efficient, and widely used in text classification. It has been shown to be functionally equivalent to logistic regression.[28] This algorithm chooses the most likely outcome given a set of predictive covariates. In the present study, the outcome is dichotomous (HF positive vs HF negative) and the covariates are words found in the clinical notes. The likelihood of an outcome is computed based on its co-occurrence frequency with each of the predictive covariates. One of the advantages of naïve Bayes compared with other more sophisticated techniques is that it is robust and fast to train, and does not require large amounts of computing resources.

Covariate Extraction. To extract covariates from text, we split the text of the clinical notes into single words listed in no particular order ("bag-of-words" representation[29]). We

collected 2048 random clinical notes manually verified to contain evidence of HF (HF-positive examples) and 2048 random notes with no HF (HF-negative examples). Each note was then represented in terms of the vocabulary contained in all notes (see **Figure 3**), with the exception of 124 stop words (eg, "the," "a," "on"). We used the entire vocabulary of 10 179 covariates without any restrictions.

Training and Testing Data. We sampled 1000 HF-positive and 1000 HF-negative examples at random from the entire collection of 4096. We set aside 200 (20%) of each half for testing and used the remaining 800 (80%) for training (Figure 1, Phase II). The *test set* was created by combining one third of the HF-positive testing examples with two thirds of the HF-negative testing examples to reflect the proportion of HF-positive examples in the data. (The proportion was determined during periodic verifications of the NLP method. A little more than one third of all patients identified by the NLP method were manually confirmed to have HF.) The *training set* was created by combining 200 HF-positive examples with 600 HF-negative examples to force the predictive modeling algorithm to favor HF-negative cases and thus maximize the positive predictive value (PPV). A predictive model was then trained using the training set and tested on the test set.

Statistical Analysis

Both the 159 028 clinical notes and 69 030 billing records (Figure 1) represent the same set of Mayo Clinic patients; however, not every clinical note generated a billing record. Both billing and NLP were compared on a per patient basis during the evaluation, even though the initial unit of analysis was different for the 2 methods due to the technical aspects of the systems used to store billing records and the text of clinical notes. Although it is possible to construct a query to extract patient-level statistics from the billing database, the NLP must operate on the text of individual notes. In the latter case, the results of processing individual notes were aggregated at the patient level. The average number of notes per patient was 5.6, with a median of 3 notes; 92% of all notes originated from outpatient visits. The NLP and predictive modeling methods were evaluated using sensitivity, specificity, and PPV. For NLP, 95% confidence intervals

■ **Table.** Comparison Between the NLP-based Method of Case Finding and the Method Based on the Dismissal-code Billing System

Method	No. Correctly Identified	Sensitivity, % (95% CI)	No. Correctly Identified	Specificity, % (95% CI)	PPV (95% CI)
	Billing HF-positive (n = 1754)		**Billing HF-negative (n = 67 276)**		
NLP	1432	81.6 (79.8, 83.5)*	65 804	97.8 (97.7, 97.9)	49.3% (47.5, 51.1)†
	Manual Review HF-positive (n = 66)		**Manual Review HF-negative (n = 200)**		
Predictive modeling	37	56.0 (44.1, 68.0)	192	96.0 (93.3, 98.7)	82.2% (73.1, 93.4)

*A total of 322 patients were identified with the billing system but not with the NLP system. See the Results section, Natural Language Processing, for details.
†The PPV obtained after manual review of the cases identified with NLP was 38% (95% CI = 36, 40).
NLP indicates natural language processing; CI, confidence interval; PPV, positive predictive modeling.

(CIs) for a single proportion were reported. For predictive modeling, the random sampling, splitting, and training and test set creation were repeated 1000 times on the entire set of 4096 positive and negative examples. Upon each repetition, the training and testing sets did not overlap. Results are reported as the mean of the 1000 samples with the 95% CIs created from the 25th and 975th observations, after numerical sorting.

RESULTS

Natural Language Processing

Diagnostic Codes as Reference. We used the billing records that contained patient diagnoses manually coded with the *ICD-9* classification to validate the NLP method. The Table summarizes the results. There were 1754 Olmsted County residents age 20 years and older who had *ICD-9* code 428.x and 69 030 patients who had any code for the time period between October 10, 2003, and May 31, 2005. For the same time period, the NLP system identified 2904 HF patients age ≥20 years who were Olmsted County residents. A total of 3226 patients were identified by either system. Of these, 1432 (44%) patients were identified by both systems, 1472 (46%) patients were identified by the NLP system but not the billing system, and 322 were identified by the billing system but not the NLP system. A total of 65 804 patients were not identified by either system. Of the 322 false negatives, 70 were non–Olmsted County residents, had the diagnosis date after May 31, 2005, or had insufficient information necessary to determine active HF. From the remaining 252, an abstractor reviewed a random sample of 25 patients (10%) to determine whether they satisfied inclusion criteria. All these patients were false positives. Thus, the NLP-based method provides

81.6% sensitivity (95% CI = 86, 100) and 97.8% specificity (95% CI = 97.7, 97.9), with a traditional diagnostic code-based approach as a reference. The PPV was 49% (95% CI = 47.5, 51.1).

Manual Review as Reference. For additional validation, the PPV of the NLP system was determined by manual record review conducted as part of the ongoing Heart Failure Surveillance Study. Two nurse abstractors with more than 15 years of experience performed the review as part of ongoing cohort recruitment. Typically, a review of a single patient record takes between 2 and 8 hours depending on the complexity of the case.

Of the 2904 patients identified by the NLP system as having active HF, record review showed that 1107 of these patients met HF criteria,[18] leaving 1797 false positives and resulting in a PPV of 38% (95% CI = 36, 40). Of the 1472 cases identified by the NLP method but not by the diagnostic code method, 210 (14%) were manually confirmed to have active HF.

Predictive Modeling

The test set for predictive-modeling validation consisted of 66 HF-positive and 200 HF-negative examples. The naïve Bayes predictive model identified 37 of the 66 HF-positive examples and 192 of the 200 HF-negative examples correctly, yielding 56% sensitivity (95% CI = 44.1, 68.0), 96% specificity (95% CI = 93.3, 98.7), and a PPV of 82% (95% CI = 73.1, 93.4).

DISCUSSION

Two complementary approaches, NLP and predictive modeling, can be used to identify subjects with HF from the EMR.

Natural language processing, which captured all cases identified with the billing system, yielded additional cases not captured by the billing system, thereby providing more comprehensive case identification. Predictive modeling, on the other hand, achieved a high PPV. A major advantage of either of these approaches over traditional methods that rely on diagnostic codes is that they enable case identification as soon as the text of a clinical note becomes available electronically, avoiding the delays and biases associated with manual coding. We also observed that the PPV determined by manual review is 10% lower than the PPV determined using diagnostic codes ($P <$ .001), which indicates that the billing system may overestimate the prevalence of HF in the population. The identification of the condition when it is not the major reason for visit is a possible cause of this discrepancy. Another possibility is that NLP searches for a set of symptoms including pulmonary edema, ankle edema, and volume overload in addition to the diagnosis of HF; therefore, NLP may find borderline cases that were never coded as HF but do qualify according to the study criteria.

Comparisons with previously published studies are limited by differences in design, clinical data, and evaluation methods. Fiszman et al compared 2 keyword-search algorithms, an NLP system, physicians, and lay persons with respect to their ability to identify patients with bacterial pneumonia from 292 chest X-ray reports.[30] They reported 95% sensitivity and 85% specificity with the NLP system. An earlier study by Hripcsak et al reported 81% sensitivity and 98% specificity for identification of 6 conditions using NLP on radiology reports.[31] The language of chest X-ray reports is different from that of discharge summaries[32] and, therefore, is likely to be different from that of clinical notes, which are functionally closer to discharge summaries in terms of grammatical patterns and vocabulary. Melton and Hripcsak evaluated the ability of the Medical Language Extraction and Encoding System to automatically detect adverse events in 1000 discharge summaries.[33] They reported 28% sensitivity, 99% specificity, and a PPV of 45%. Although a direct comparison with this study is not possible because of different goals (identification of adverse events vs identification of patients with HF), the study by Melton and Hripcsak is comparable to our study with respect to the type of clinical documents used and the large number of participants. Our study extends this previous work by showing the utility of NLP for highly sensitive identification of patients with HF.

Our results are also important in light of previous studies that examined the validity of diagnostic information contained within medical service claims and billing data. Wilchesky et al showed that using the medical service claims data results in highly specific (96%) but highly nonsensitive (42%) case ascertainment for CHF.[34] Similar results were found by Ahmed et al: the specificity of using a claims-based *ICD-9* algorithm for high-risk conditions including cardiovascular disease was 99% but the sensitivity—12%—was very low.[35] Onofrei et al examined the physician-provided problem list coded with *ICD-9* as a source of diagnostic information to identify patients with HF and compared that with 2 gold standards, defined by documented left ventricular ejection fractions of ≤55% and ≤40%.[5] The sensitivity for case finding was 44% for the ejection fraction ≤ 55% and 54% for the ejection fraction ≤ 40%. All of these studies question the validity of using sources of coded diagnostic information for case finding of HF. Our study shows that using the text of the EMR enables case identification that is at least as sensitive as using coded diagnoses, with the added benefit of timeliness.

The results of our project are particularly relevant in the context of identification of syndromic conditions in a practice setting, where a number of disorders may not lead to hospitalizations or billing codes. For example, subsequent to the project reported in this article, we extended our methodology to identification of patients with angina pectoris,[36] whose initial diagnosis relies on patient-reported symptoms. By the same token, our methodology may be extended to other conditions that are diagnosed based on the initial presentation of symptoms by the patient, including rheumatoid arthritis, gastrointestinal disorders, psychiatric conditions, obesity, and drug abuse.

Limitations and Strengths

Some limitations need to be acknowledged to facilitate the interpretation of the data in this study. The *ICD-9*–based billing system can be used only as an approximation of a criterion standard. The cases that were missed by both the billing system and the NLP system lie outside the range of the current evaluation. Also, the results of this study may not be readily generalizable to other diagnoses; therefore, the use of EMR for patient identification has to be validated on a case-by-case basis.

This study also has unique strengths. It used a large dataset (more than 3000 patients) that was developed over a period of 3 years and involved complete manual records abstraction for validation. Another strength is that this study addressed identification of HF patients, whose diagnosis is complex, and relies in part on the language found in the unrestricted text of the EMR. The 2 methods described here were tested on the same population, which made it possible to determine their respective yields in the same dataset and to define their potential application separately or in combination with predictable results. Another advantage of the NLP approach is

that it is not restricted to a specific data element or a specific location in the EMR. The Mayo Clinic EMR maintains a diagnostic problem list that is used to summarize the main findings entered into the clinical note. The notes contain problem-list entries as numbered items inside the Impression/Report/Plan and the Final Diagnosis sections. The NLP algorithm described in this article does not take advantage of the problem-list items. Instead, the search is performed across the entire text of the note in the attempt to capture symptom information. Therefore, our NLP strategy may be used in EMR systems that do not routinely use problem-list entries. The feasibility of using the NLP strategy in other EMR systems and for other conditions will be assessed in subsequent work.

Implications for Clinical Research

Although the highly sensitive NLP method may be more appropriate as a screening mechanism for observational studies, the predictive modeling method may be more suitable for clinical trials, when stricter inclusion criteria may be required. Indeed, the NLP method often involves subsequent manual abstraction of medical records, and a highly sensitive screening tool will direct manual data collection. The predictive modeling method is based on selected populations and is more concerned with the efficient enrollment of patients who fit study inclusion and exclusion criteria. For clinical trials, the predictive-modeling approach with a higher PPV is a better screening mechanism.

Take-away Points

Two methods, natural language processing and predictive modeling, were used to identify patients with heart failure from electronic medical records.

■ Both approaches enable accurate and timely case identification as soon as the text of a clinical note becomes available electronically, avoiding the delays and biases associated with manual coding.

■ Natural language processing may be more suitable for studies requiring the highest sensitivity such as observational studies.

■ Because of its higher positive predictive value, the predictive-modeling approach is a better screening mechanism for clinical trials.

REFERENCES

1. **Committee on Quality of Health Care in America, Institute of Medicine.** *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington, DC: Institute of Medicine; 2001.

2. **Kurland LT, Molgaard CA.** The patient record in epidemiology. *Sci Am.* 1981;245:54-63.

3. **Melton LJ.** History of the Rochester Epidemiology Project. *Mayo Clin Proc.* 1996;7:266-274.

4. **Psaty BM, Boineau R, Kuller LH, Luepker RV.** The potential costs of upcoding for heart failure in the United States. *Am J Cardiol.* 1999;84:108-109.

5. **Onofrei M, Hunt J, Siemienczuk J, Touchette DR, Middleton B.** A first step towards translating evidence into practice: heart failure in a community practice-based research network. *Inform Prim Care.* 2004;12:139-145.

6. **Hunt S.** ACC/AHA 2005 guideline update for the diagnosis and management of chronic heart failure in the adult: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Update the 2001 Guidelines for the Evaluation and Management of Heart Failure). *J Am Coll Cardiol.* 2005;46:e1-e82.

7. **Yang Y, Chute CG.** A linear least squares fit mapping method for information retrieval from natural language texts. In: *Proceedings of 14th International Conference on Computational Linguistics (COLING 92).* Vol II. Nantes, France: August 1992:447-453.

8. **Lewis D.** Naive (Bayes) at forty: the independence assumption in information retrieval. In: *Proceedings of the 10th European Conference on Machine Learning (ECML 98).* Berlin, Germany: Springer Verlag; 1998:4-15.

9. **Aronsky D, Haug PJ.** Automatic identification of patients eligible for a pneumonia guideline. In: Overhage JM, ed. *Proceedings of the 2000 AMIA Annual Symposium.* Bethesda, Md: American Medical Informatics Association; 2000:12-16.

10. **Johnson D, Oles F, Zhang T, Goetz T.** A decision-tree-based symbolic rule induction system for text categorization. *IBM Systems Journal.* 2002;41:428-437.

11. **Nigam K, Lafferty J, McCullum A.** Using maximum entropy for text classification. In: Joachims T, ed. *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering.* Stockholm, Sweden: August 1999:61-67.

12. **Yang Y.** Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In: Croft WB, van Rijsbergen CJ, eds. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: Springer-Verlag; 1994:13-22.

13. **Wilcox A.** *Automated Classification of Text Reports* [dissertation]. New York: Columbia University; 2000.

14. **Aronow DB, Fangfang F, Croft WB.** Ad hoc classification of radiology reports. *J Am Med Inform Assoc.* 1999;6:393-411.

15. **Jain NL, Friedman C.** Identification of finding suspicious for breast cancer based on natural language processing of mammogram reports. In: *Proceedings of the 1997 AMIA Annual Symposium.* Bethesda, Md: American Medical Informatics Association; 1997:829-833.

16. Hripcsak G, Austin JHM, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology.* 2002;224:157-163.

17. Carpenter P. The electronic medical record: perspective from Mayo Clinic. *Int J Biomed Comput.* 1993;34:159-171.

18. Roger VL, Weston SA, Redfield MM, et al. Trends in heart failure incidence and survival in a community-based population. *JAMA.* 2004;292:344-350.

19. Dolin RH, Alschuler L, Boyer S, et al. HL7 clinical document architecture, release 2. *J Am Med Inform Assoc.* 2006;13:30-39.

20. Pakhomov SV, Buntrock J, Chute CG. Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. *J Biomed Inform.* 2005;38:145-153.

21. Commission on Professional and Hospital Activities. *Hospital Adaptation of ICDA (H-ICDA).* 2nd ed. Ann Arbor, Mich: CPHA; 1973.

22. Pakhomov SV, Buntrock J, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *J Am Med Inform Assoc.* 2006;13: 516-525.

23. Roger VL, Killian J, Henkel M, et al. Coronary disease surveillance in Olmsted County: objectives and methodology. *J Clin Epidemiol.* 2002;55:593-601.

24. Steyerberg EW, Eijkemans MJC, Boersma E, Habbem JDF. Equally valid models gave divergent predictions for mortality in acute myocardial infarction patients in a comparison of logical regression models. *J Clin Epidemiol.* 2005;58:383-390.

25. Wolfe R, McKenzie DP, Black J, Simpson P, Gabbe BJ, Cameron PA. Models developed by three techniques did not achieve acceptable prediction of binary trauma outcomes. *J Clin Epidemiol.* 2006;59:82-89.

26. Terrin N, Schmid CH, Griffith JL, D'Agostino RBS, Selker HP. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *J Clin Epidemiol.* 2003; 56:721-729.

27. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques.* 2nd ed. San Francisco, Calif: Elsevier; 2005.

28. Roos T, Wettig H, Grunwald P, Myllymaki P, Tirri H. On discriminative bayesian network classifiers and logistic regression. *Machine Learning.* 2005;59:267-296.

29. Manning C, Shutze H. *Foundations of Statistical Natural Language Processing.* Cambridge, Mass: MIT Press; 1999.

30. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc.* 2000;7:593-604.

31. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med.* 1995;122:681-688.

32. Friedman C. A broad-coverage natural language processing system. In: Overhage JM, ed. *Proceedings of the 2000 AMIA Annual Symposium.* Bethesda, Md: American Medical Informatics Association; 2000: 270-274.

33. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc.* 2005;12:448-457.

34. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. *J Clin Epidemiol.* 2004;57:131-141.

35. Ahmed F, Janes GR, Baron R, Latts LM. Preferred provider organization claims showed high predictive value but missed substantial portion of adults with high-risk conditions. *J Clin Epidemiol.* 2005;58:624-628.

36. Pakhomov S, Hemingway H, Weston S, Jacobsen S, Rodeheffer R, Roger V. Epidemiology of angina pectoris: role of natural language processing of the medical record. *Am Heart J.* 2007;153:666-673. ■