Published in final edited form as:

Am Heart J. 2007 April ; 153(4): 666–673.

Epidemiology of angina pectoris: role of natural language processing of the medical record

Serguei Pakhomov, Ph.D.[†], Harry Hemingway, FRCP[‡], Susan A. Weston, MS[†], Steven J. Jacobsen, MD, Ph.D.[†], Richard Rodeheffer, MD^{*,†}, and Véronique L. Roger, MD, MPH^{*,†}

^{*} *Division of Cardiovascular Diseases, Department of Medicine, Mayo Clinic College of Medicine, Rochester Minnesota*

[†] *Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester Minnesota*

[‡] *Department of Epidemiology and Public Health, UCL Medical School, UK*

Abstract

Background—The diagnosis of angina is challenging as it relies on symptom descriptions. Natural language processing (NLP) of the electronic medical record (EMR) can provide access to such information contained in free text that may not be fully captured by conventional diagnostic coding.

Objective—To test the hypothesis that NLP of the EMR improves angina pectoris (AP) ascertainment over diagnostic codes.

Methods—Billing records of in- and out-patients were searched for ICD-9 codes for AP, chronic ischemic heart disease and chest pain. EMR clinical reports were searched electronically for 50 specific non-negated natural language synonyms to these ICD-9 codes. The two methods were compared to a standardized assessment of angina by Rose questionnaire for three diagnostic levels: unspecified chest pain, exertional chest pain, and Rose angina.

Results—Compared to the Rose questionnaire, the true positive rate of EMR-NLP for unspecified chest pain was 62% (95%CI:55–67) vs. 51% (95%CI:44–58) for diagnostic codes ($p<0.001$). For exertional chest pain, the EMR-NLP true positive rate was 71% (95%CI:61–80) vs. 62% (95%CI: 52–73) for diagnostic codes ($p=0.10$). Both approaches had 88% (95%CI:65–100) true positive rate for Rose angina. The EMR-NLP method consistently identified more patients with exertional chest pain over 28-month follow-up.

Conclusion—EMR-NLP method improves the detection of unspecified and exertional chest pain cases compared to diagnostic codes. These findings have implications for epidemiological and clinical studies of angina pectoris.

Introduction

While heart disease remains the number one cause of death in the Western World, angina pectoris, a central component of the burden of coronary disease, remains understudied and its burden may be under-appreciated. This is in part related to the fact that the diagnosis of angina is based solely on the characterization of the pain by the patient. Indeed, while characteristics of the pain may be recorded in the review of systems or other components of the clinical notes,

it may not surface among the list of diagnostic codes, such that studies that rely on diagnostic coding likely underestimate the burden of angina.

The verbal characterization of the symptoms conveyed by the patient, and recorded by the physician is central to the practice of clinical medicine, and increasing importance is attached to patient-centered clinical care(1). Electronic Medical Records (EMRs) contain physicians' free text description of symptoms, and diagnoses and these may be given one of several ICD-9 CM codes, or none at all(2). With the increasing adoption of the electronic medical record, free text (natural language) data on the clinical history can now be subjected to automated analysis in ways which are impossible or uneconomic with paper based records(3–5). A common, costly (6) example of a symptomatic condition in which history taking is central to management is chronic stable angina pectoris(7). Optimal methods of identifying stable angina patients remain unclear; many patients with typical symptoms are not diagnosed as angina(8) and age, sex, and ethnicity may influence physician's recommendations for diagnostic testing such as coronary angiography and the resulting ICD codes(9–12).

It is not known whether natural language processing of the EMR for identification of a chronic symptomatic condition adds information beyond code based searches. We have previously demonstrated that Natural Language Processing of the text of the EMR (EMR-NLP) can prospectively identify patients with ICD-9-CM codes for heart failure with 100% true positive rate(13). However, since ICD codes may not be the optimal reference standard(14–16), this study was limited by the lack of a reference group, that does not rely on coding. Additionally, chronic conditions may reveal themselves to medical care over a period of months and it is not known if the yield of EMR-NLP changes over time. More patients may be found by the EMR-NLP method over time due to multiple presentations of chest pain by the same patient. Alternatively, the code based methods may miss the patients found initially because of the lag in assigning diagnostic codes for billing which may exceed 30 days after the diagnosis was dictated and recorded in the EMR in free text. We hypothesized that processing the natural language in the EMR improved the ascertainment of angina over ICD codes. Our objectives were firstly to compare the EMR-NLP and the ICD methods to the Rose questionnaire, a standardized and validated(17) instrument for assessment of angina in the general population (18,19) and, secondly, to examine changes in performance over time of the two methods.

Methods

Rose Angina Questionnaire as reference

The Rose angina questionnaire is the only validated instrument for assessing symptoms of typical angina pectoris in the general population, independent of medical care. It is highly (>90%) specific when compared against physician diagnosed angina,(20–22) is strongly associated with coronary artery calcification, (23) and subsequent risk of coronary events. (24,25) The Rose angina questionnaire was administered using a survey of a random sample of the Olmsted County, MN population (124,277 in 2000) aged at least 45 years. In addition to the Rose angina questionnaire, the survey included questions on symptoms other than chest pain including shortness-of-breath and cough. Of 4203 eligible residents, 2042 (49%) participated; analysis of a random sample of 488 showed no significant difference between responders and non-responders in terms of prior history of cardiovascular disease(26). A follow up questionnaire was sent between January 1, 2001 and June 1, 2003. Our current analysis is restricted to 892 (44%) participants who filled out the questionnaire between January 1, 2003 and June 1, 2003 because 2003 was the first year all services at the Mayo Clinic started to use the EMR. Patient characteristics are reported in Table 1.

ICD codes

The ICD-9-CM codes 413.x (angina pectoris), 414.0, 414.8, 414.9 (ischemic heart disease) and 786.5 (chest pain) were used to operationally define “ICD angina.” While we broaden the definition of ICD angina beyond the ICD-9 code 413.x, we also restrict the scope of the definition to conditions most likely indicative of chronic angina. For a sensitivity analysis, we also investigated ICD codes for atypical anginal symptoms 784.1 (throat pain), 723.1 (neck pain), 526.9 (jaw pain). We searched for these ICD codes in the billing system dated between January 1, 2003 and November 1, 2005 for all those patients who had at least one clinical note created between these dates. The search was not limited to the primary diagnosis – all diagnostic codes were evaluated.

EMR

The clinical notes used in this study are electronic documents with information dictated by Mayo Clinic physicians about each patient contact and transcribed by trained medical transcriptionists. Figure 1 shows a typical clinical note. Clinical notes are structured into conventional sections (including Chief Complaint, Current Medications, History of Present Illness, Social History, Family History, Impression/Report/Plan), each using language specific in terminology and structure. While restricting searches to specific sections (e.g. exclude Family History) might improve precision of patient identification, the reliability of segmenting medical dictation into sections is currently unknown. Therefore, for this study, we used the entire text of the clinical note to search for terms indicative of angina pectoris.

Processing the natural language – non-negated terms

We used a Text Analysis system(27) created at the Mayo Clinic for the purpose of structuring semi-structured clinical notes for subsequent indexing and retrieval. The Text Analysis system is part of a larger Mayo Clinic Life Sciences System whose purpose is to link disparate sources of clinical patient information including clinical reports, laboratory results, demographic and genomic data. As well as allowing automatic identification of terms for disorders and symptoms this system makes the crucial distinction between negated and non-negated terms as illustrated in the following examples:

1. Sixty-two-year-old male presents to the Emergency Department with a gradually increasing concern and problem with chest pain.
2. She has **not** experienced **any** syncope or near syncope and has indicated **no** chest pain or chest pressure.

The term “chest pain” in the Example 1 occurs in a non-negated context, while the terms “syncope”, “near syncope”, “chest pain” and “chest pressure” all occur in negated contexts identified through negation indicators such as “not”, “any” and “no.” In order to identify the negation, the Text Analysis system automatically determines boundaries of clinically relevant terms found in the text of clinical notes using a sliding window of 7 content words in all possible permutations. Each permutation of words within the sliding window is matched automatically to terms contained in a subset of the Unified Medical Language System (Metathesaurus®) (28). The subset consists of several widely used medical terminologies including SNOMED-CT®, ICD-9-CM, RxNorm and NCI Thesaurus. Information on the position of the term relative to the beginning of the document allows scanning the adjacent textual areas to the left and right of the found term for negation indicators. This algorithm for negation identification is similar to NegEx developed by Chapman et al. (29) and consists of scanning the adjacent text until a scope delimiter including a punctuation mark, a conjunction (“but”, “however”, “nevertheless”, “how”, “when”, etc.) or a personal pronoun in nominative form (“I”, “we”, “he”, “she”, “they”, “it”) is found. All of these indicators tend to signal a clause or a sentence

boundary. If a negation indicator is found within the adjacent areas bounded by scope delimiters, then the term is marked as negated.

Search for patients with angina

“NLP angina” was defined as a set of terms and their variants that are synonymous to the descriptions corresponding to the ICD-9-CM codes used in the ICD method. Table 2 shows a list of terms that were derived by querying the Unified Medical Language System (<http://www.nlm.nih.gov/research/umls/>) which represents a compendium of over 130 biomedical vocabularies containing over 1 million concepts and relationships among them including synonymy(30). In order to maximize the matching between these synonyms and the terms identified in the clinical notes with the Text Analysis system, we manually inserted wildcard characters (i.e. search operators that match on any character –Table 2) and reduced some of the words to their stem (coronary -> coron). While sophisticated stemming tools are available such as the Lexical Variant Generator (LVG) developed at the National Library of Medicine, we chose to do the stemming manually for this study in order to maximize control over a potential source of variability.

For sensitivity analysis, we also defined a set of search terms that parallel the additional ICD codes in the expanded definition of ICD angina. These terms refer to atypical locations of anginal pain and include “%jaw%pain%”, “%pain%jaw%”, “%throat%pain%”, “%pain%throat%”, “%neck%pain%”, “%pain%neck%”.

Patients in the Rose angina cohort, who filled out the questionnaire between January 1, 2003 and June 1, 2003 were identified as positive for NLP angina if at least one of the patient’s clinical notes generated between January 1, 2003 and November 1, 2005 contained a match on any of the non-negated terms in Table 2.

Other possible causes of chest pain

Patients presenting with chest pain may be diagnosed with non-cardiac conditions such as dyspepsia, heartburn of unknown etiology or other non-coronary cardiac conditions which may be accompanied by chest pain. In order to examine whether this may be a factor in identification of patients based on this particular symptom, we examined the billing records of the patients that indicated exertional chest pain on the Rose questionnaire for the following ICD-9 codes: 390-459 (CV disorders), 781.1 (heartburn) and 536.5 (dyspepsia). The CV disorders range was further split into 3 groupings (Coronary Heart Disease (CHD), Non-CHD Diseases of the Heart, Non-cardiac Circulatory Diseases) according to the AHA guidelines for classifying causes of cardiovascular death(31).

Follow up

We tracked the number of patients with any chest pain identified by the two methods at half-year intervals until the end of our study (November 1, 2005). Since “angina equivalents” may also present as breathlessness(32) we identified a cohort of people who reported shortness-of-breath on the general population survey and followed them up for ICD angina diagnoses. In order to gauge the specificity of effect we also included the symptom of cough, which was hypothesized not to be related to angina diagnosis.

All aspects of this study were approved by the Institutional Review Board.

Statistical Analysis

Baseline characteristics are presented as means $\pm$ standard deviations for continuous variables and as frequencies for categorical variables. True positive and true negative rates are presented as percentages with 95% confidence intervals. Differences among groups were tested using

the Kruskal-Wallis nonparametric test for continuous variables and the chi-square test for categorical variables. Differences in true positive and true negative rates were tested using the chi-square test for the difference between two proportions in a paired sample since both the ICD and EMR methods were tested on the same sample.

Results

Rose chest pain and subsequent clinical notes

Eight hundred seventy one (98%) of 892 participants had at least one clinical note dictated between January 1, 2003 and November 1, 2005. Of these 871, 202 (23%) reported chest pain. Eighty five (10%) of 871 reported exertional chest pain. Baseline characteristics are reported in Table 1. Restricting the cohort by date and presence of clinical notes in the EMR did not change the results.

Comparison between the EMR-NLP and the ICD codes

The true positive rate of the EMR-NLP system in identifying patients who reported any chest pain on the questionnaire was 62% (95% confidence interval (CI): 55, 67) compared to 51% (95%CI: 44, 58) using diagnostic codes ($p<0.001$) (Table 3). For exertional chest pain, the true positive rate of the EMR-NLP was 71% (95%CI: 61, 80) compared to 62% (95%CI: 52, 73) using diagnostic codes ($p=0.10$). The two approaches had identical true positive rates (88%) for the detection of definite Rose angina. The true negative rate of the EMR-NLP system was 10% lower than the diagnostic codes system for chest pain, exertional chest pain and Rose angina ($p<0.001$ for each diagnostic level).

Sensitivity analysis

We tested several modified definitions of EMR-NLP and ICD angina to account for the fact that chest pain may not be due exclusively to angina, and that anginal pain may occur in locations other than the chest.

Removing “chest pain” from search—We sought to measure whether removing the ICD code 786.50 for chest pain from the ICD angina definition and the term “chest pain” from the NLP angina definition would affect the true and false positive rates of the EMR-NLP and ICD patient identification methods. Excluding chest pain from the definitions results in lower true positive but higher true negative rates.

Atypical pain locations—We also sought to measure whether the true and false positive rates of the EMR-NLP and ICD patient identification methods would be affected by adding ICD codes 784.1, 723.1, and 526.9 for atypical pain locations to the ICD angina definitions and their natural language equivalents (throat, neck and jaw pain respectively) to the NLP angina definitions. Adding atypical symptom locations raised the true positive rates of both approaches while lowering their true negative rates. The results of this sensitivity analysis are shown in Table 3.

Longitudinal analysis

We found that initially on June 1, 2003, 30% of the patients who reported any chest pain were identified by the EMR-NLP method as positive for angina while 17% were identified by the ICD method. Both ratios steadily increased over time to reach 62% for EMR-NLP and 51% by the ICD method (Fig. 2). No convergence between the EMR-NLP and the ICD methods was detected.

We identified 164 patients who reported shortness-of-breath; of these ICD angina increased from 14% on June 1, 2003 to 54% on November 1, 2005, similar to that observed for patients reporting chest pain. We identified 238 patients who reported cough, of whom 13% had ICD angina on June 1, 2003 and 36% November 1, 2005. Reporting of chest pain was more strongly associated with reporting shortness-of-breath (OR = 3.44 (95% CI 2.47–4.78), $p < 0.001$), than with cough (OR = 1.86 (1.38, 2.49), $p < 0.001$).

Possible non-cardiovascular causes

The examination of the billing records of the 85 patients with exertional chest pain revealed only 1 (1.2%) patient with code 781.1 (heartburn) and 8 (9.4%) patients with 536.5 (dyspepsia). The results for the 3 groups of CV disorders (Coronary Heart Disease (CHD), Non-CHD Diseases of the Heart, Non-cardiac Circulatory Diseases) show 60 (70.6%) patients with ICD-9 codes indicating any of the three groups.

Discussion

Our findings indicate that, compared to diagnostic coding, natural language processing of the EMR results in higher true positive rate to identify patients in the general population with self-reported exertional chest pain and any chest pain. To the best of our knowledge, this is the first study that relies on an advanced EMR system and expertise in population-based research and medical informatics computer science and linguistics to validate the use of natural language processing of the EMR to identify a chronic symptomatic condition within a geographically-defined population.

Stable angina or chest pain is commonly the initial presentation of coronary disease, particularly in women(33) and may precede myocardial infarction by years. Diagnosis is notoriously difficult, and increasingly it is observed that there is an appreciable risk of subsequent coronary events(34) among patients with “non-cardiac” chest pain(35), non-obstructed angiograms(36) and admissions with “chest pain” without further diagnosis. Structured patient-reportable information may be of assistance for earlier diagnosis(37). Physical symptoms account for half of all outpatient visits(38) and are commonly not diagnosed (39). NLP methods that take advantage of symptoms reported in electronic physician’s notes may provide insight into causes and consequences of these symptoms and help in preventive approaches, particular for lifestyle and risk factor modification(40).

This methodology can be easily implemented for other conditions including gastrointestinal, musculoskeletal, respiratory and other chronic disease, which rely heavily on patient’s presentation of symptoms that may not be systematically captured by coding systems. The simplicity of our approach (reliance on term lists) minimizes the amount of adjustments particularly for many European languages, for which medical language processing systems have been developed(41). The Unified Medical Language System Metathesaurus® also contains terminologies in multiple languages that can be used for synonym finding.

Only half of the patients with Rose chest pain are identified by ICD codes while the EMR-NLP method identifies 62% of them. This may be due to patient with Rose chest pain not presenting to care or presenting to care but receiving ICD codes unrelated to cardiac conditions. As chronic symptomatic conditions may reveal themselves to medical care gradually over time, we examined longitudinal trends and found that EMR-NLP remained consistently superior to coding with no convergence over time between the 2 methods. This persistent under-ascertainment of angina is consistent with Hemingway et al.’s(10) report that more than half of the subjects with typical symptoms did not have a diagnosis of angina recorded by their physician, while 60% of the patients who repeatedly presented with angina remained undiagnosed over 5 years of follow up. The greater yield of EMR-NLP compared to coding

remained constant overtime indicates that EMR-NLP has the potential of improving the ascertainment of clinical angina.

Limitations and strengths

Limitations should be acknowledged to aid in the interpretation of the data. The number of Rose angina cases was relatively small, despite the large number of participants in the survey. However, cases of exertional chest pain have been shown to also have prognostic importance for future coronary events(42). Although the Rose angina questionnaire has been extensively validated against physician diagnosed angina,(20–22) coronary artery calcification,(23) and risk of coronary events,(24,25) some cases may have no underlying myocardial ischemia. It is also possible that in some cases chest pain resolves spontaneously between the response to Rose questionnaire and the next patient visit to the clinic. These cases would lower the true positive rate of both approaches. Our NLP-EMR system used a negation filter in order to minimize false positives. Excluding negation identification results in higher true positive rate but much lower true negative rate (data not shown). However, manual records review for the additional patient identified by the EMR method after removing negation indicated that this was a false positive.

The innovative nature of our study lies in the use of the free text of the EMR which offers several advantages. It provides direct access to the physician's assessment of the patient's condition and is thus more responsive to evolving clinical practice compared to relatively static coding systems. Further, it considers the totality of the clinical note, not just the final diagnosis section, typically used for coding. Important information may be recorded in the history of present illness sections for patients with chronic diseases such as angina, heart failure and rheumatoid arthritis. The text can be accessed immediately after transcription of the clinical report in contrast with the considerable delay (over 30 days in most cases) for manually assigned diagnostic codes. The present findings demonstrate the value of EMR-NLP methods for the comprehensive identification of patients with cardiovascular disease compared to administrative coding. Indeed, code-based medical service claims data have high true negative rate but low true positive rate to ascertain heart failure(14,15). To this end, Onofrei et al. examined the physician provided problem list coded with ICD-9-CM to identify patients with HF and compared that to two 'gold standards' defined by a left ventricular ejection fraction $\leq 55\%$ and $\leq 40\%$. The true positive rate of case finding was 44% and 54% respectively(16). Another important strength is that the present study population is geographically-defined, includes outpatient and inpatient cases, which optimizes the generalizability of our findings.

Conclusions

Processing the natural language of the physician's rendering of the patient's history in the electronic medical record is a critical component of identifying patients with angina. This will help in the ascertainment of the burden of angina pectoris, a central manifestation of coronary disease.

Acknowledgements

This work was in part supported by grants for the Public Health Service (RO1 HL 59205, HL 72435), the Rochester Epidemiology Project (GM14321 and AR30582), and the NIH Roadmap Multidisciplinary Clinical Research Career Development Award Grant (K12/NICHD)-HD49078. Dr. Hemingway is supported by a Department of Health Public Health Career Scientist Award. We acknowledge the assistance of Ryan Meverden and Jill Killian with data collection and analysis.

References

1. Crossing the Quality Chasm. A New Health System for the 21st Century. Washington, DC: Institute of Medicine; 2001.

2. Subramanian U, Fihn SD, Weinberger M, Plue L, Smith FE, Udris EM, et al. A controlled trial of including symptom data in computer-based care suggestions for managing patients with chronic heart failure. *Am J Med* 2004;116(6):375–84. [PubMed: 15006586]
3. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. *J Am Med Inform Assoc* 1994;1(2):142–60. [PubMed: 7719796]
4. Friedman C. A broad-coverage natural language processing system. *Proc AMIA Symp* 2000:270–4. [PubMed: 11079887]
5. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc* 2005;12(4):448–57. [PubMed: 15802475]
6. Javitz HS, Ward MM, Watson JB, Jaana M. Cost of illness of chronic angina. *Am J Manag Care* 2004;10(11 Suppl):S358–69. [PubMed: 15603245]
7. Hemingway H, McCallum A, Shipley M, Manderbacka K, Martikainen P, Keskimaki I. Incidence and prognostic implications of stable angina pectoris among women and men. *JAMA* 2006;295(12):1404–11. [PubMed: 16551712]
8. Philpott S, Boynton PM, Feder G, Hemingway H. Gender differences in descriptions of angina symptoms and health problems immediately prior to angiography: the ACRE study. Appropriateness of Coronary Revascularisation study. *Soc Sci Med* 2001;52(10):1565–75. [PubMed: 11314852]
9. Schulman KA, Berlin JA, Harless W, Kerner JF, Sistrunk S, Gersh BJ, et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. *N Engl J Med* 1999;340(8):618–26. [PubMed: 10029647]
10. Hemingway H, Shipley M, Britton A, Page M, Macfarlane P, Marmot M. Prognosis of angina with and without a diagnosis: 11 year follow up in the Whitehall II prospective cohort study. *BMJ* 2003;327(7420):895. [PubMed: 14563744]
11. Research finds low EHR adoption rates for physician groups. Medical Group Management Association Center for Research and University of Minnesota School of Public Health. In; 2005.
12. Ford EW, Menachemi N, Phillips MT. Predicting the adoption of electronic health records by physicians: when will health care be paperless? *J Am Med Inform Assoc* 2006;13(1):106–12. [PubMed: 16221936]
13. 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(2):145–53. [PubMed: 15797003]
14. Ahmed F, Janes GR, Baron R, Latts LM. Preferred provider organization claims showed high predictive value but missed substantial proportion of adults with high-risk conditions. *J Clin Epidemiol* 2005;58(6):624–8. [PubMed: 16028341]
15. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. *J Clin Epidemiol* 2004;57(2):131–41. [PubMed: 15125622]
16. Onofrei M, Hunt J, Siemieniczuk 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(3):139–45. [PubMed: 15606986]
17. Sorlie PD, Cooper L, Schreiner PJ, Rosamond W, Szklo M. Repeatability and validity of the Rose questionnaire for angina pectoris in the Atherosclerosis Risk in Communities Study. *J Clin Epidemiol* 1996;49(7):719–725. [PubMed: 8691220]
18. Rose GA. The Diagnosis of Ischemic Heart Pain and Intermittent Claudication in Filed Surveys. *Bulletin of World Health Organization* 1962;27:645–658.
19. Rose, GA.; Blackburn, H.; Gillum, RF.; Prineas, RJ. *Cardiovascular Survey Methods*. 2. Geneva: World Health Organization; 1986.
20. Lawlor DA, Adamson J, Ebrahim S. Performance of the WHO Rose angina questionnaire in post-menopausal women: are all of the questions necessary? *J Epidemiol Community Health* 2003;57(7): 538–41. [PubMed: 12821705]
21. Fischbacher CM, Bhopal R, Unwin N, White M, Alberti KG. The performance of the Rose angina questionnaire in South Asian and European origin populations: a comparative study in Newcastle, UK. *Int J Epidemiol* 2001;30(5):1009–16. [PubMed: 11689512]
22. Blackwelder WC, Kagan A, Gordon T, Rhoads GG. Comparison of methods for diagnosing angina pectoris: the Honolulu heart study. *Int J Epidemiol* 1981;10(3):211–5. [PubMed: 7287281]

23. Oei HH, Vliegenthart R, Deckers JW, Hofman A, Oudkerk M, Witteman JC. The association of Rose questionnaire angina pectoris and coronary calcification in a general population: the Rotterdam Coronary Calcification Study. *Ann Epidemiol* 2004;14(6):431–6. [PubMed: 15246332]
24. Rose G. Variability of angina. Some implications for epidemiology. *Br J Prev Soc Med* 1968;22(1): 12–5. [PubMed: 5636970]
25. Lampe FC, Whincup PH, Shaper AG, Wannamethee SG, Walker M, Ebrahim S. Variability of angina symptoms and the risk of major ischemic heart disease events. *Am J Epidemiol* 2001;153(12):1173–82. [PubMed: 11415952]
26. Jacobsen SJ, Mahoney DW, Redfield MM, Bailey KR, Burnett JC Jr, Rodeheffer RJ. Participation bias in a population-based echocardiography study. *Ann Epidemiol* 2004;14(8):579–84. [PubMed: 15350958]
27. Pakhomov, S.; Buntrock, J.; Duffy, P. High Throughput Modularized NLP System for Clinical Text. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics 2005; Demonstrations and Interactive Posters.
28. UMLS: Unified Medical Language System. In: National Library of Medicine; 2004.
29. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. Evaluation of negation phrases in narrative clinical reports. *Proc AMIA Symp* 2001;105–9. [PubMed: 11825163]
30. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. *J Am Med Inform Assoc* 2005;12(4):486–94. [PubMed: 15802483]
31. American Heart Association. Heart Disease and Stroke Statistics-2005 Update. Dallas, Texas: 2005.
32. Pepine CJ, Wiener L. Relationship of anginal symptoms to lung mechanics during myocardial ischemia. *Circulation* 1972;46(5):863–9. [PubMed: 4628188]
33. Murabito JM, Evans JC, Larson MG, Levy D. Prognosis after the onset of coronary heart disease. An investigation of differences in outcome between the sexes according to initial coronary disease presentation. *Circulation* 1993;88(6):2548–55. [PubMed: 8252666]
34. Hsia J, Aragaki A, Bloch M, LaCroix AZ, Wallace R. Predictors of angina pectoris versus myocardial infarction from the Women's Health Initiative Observational Study. *Am J Cardiol* 2004;93(6):673–8. [PubMed: 15019867]
35. Sekhri N, Feder G, Junghans C, Hemingway HJ, Timmis AD. How effective are rapid access chest pain clinics? Prognosis of incident angina and non-cardiac chest pain in 8762 consecutive patients. *Heart*. 2006In press
36. Bugiardini R, Bairey Merz CN. Angina With “Normal” Coronary Arteries: A Changing Philosophy. *JAMA* 2005;293(4):477–484. [PubMed: 15671433]
37. Wang SJ, Ohno-Machado L, Fraser H, Kennedy L. Using patient-reportable clinical history factors to predict myocardial infarction. *Computers in Biology and Medicine* 2001;31:1–13. [PubMed: 11058690]
38. Hing E, Cherry DK, Woodwell DA. National Ambulatory Medical Care Survey: 2004 summary. *Adv Data* 2006;(374):1–33.
39. Lamberg L. New mind/body tactics target medically unexplained physical symptoms and fears. *JAMA* 2005;294(17):2152–4. [PubMed: 16264150]
40. Kroenke K. Studying symptoms: sampling and measurement issues. *Ann Intern Med* 2001;134(9 Pt 2):844–53. [PubMed: 11346320]
41. Rassinoux A, Michel P, Wagner J, Baud R. Current trends with natural language processing. *Medinfo*. 19958 Pt 2
42. LaCroix AZ, Guralnik JM, Curb JD, Wallace RB, Ostfeld AM, Hennekens CH. Chest pain and coronary heart disease mortality among older men and women in three communities. *Circulation* 1990;81(2):437–46. [PubMed: 2297854]

Figure 1.
A Clinical note example (this is a note for a non-existent test patient)

Figure 2. Increase over time in recognition of patients reporting chest pain on the Rose questionnaire according to EMR-NLP and ICD codes
 (06/01/2003 (the date of the last Rose angina questionnaire)-11/10/2005).
 N =222 of whom 202 had at least one clinical note.

Table 1

Baseline characteristics of general population survey participants reporting no chest pain, any chest pain or exertional chest pain

	No Chest Pain	Chest Pain	Exertional Chest Pain	p-value
	N=669	N=117 ^a	N=85 ^a	
Physical characteristics				
Age	62.4 ± 9.6	61.3 ± 9.8	62.8 ± 9.5	0.39
Male	309 (50.6)	49 (44.1)	39 (53.4)	0.38
Height	168.4 ± 9.7	167.4 ± 8.9	168.8 ± 9.3	0.48
Weight	80.5 ± 18.0	77.8 ± 15.0	83.3 ± 17.2	0.12
CV Risk Factors				
Smoked at least 100 Cigarettes	297 (48.8)	54 (48.7)	43 (58.9)	0.26
Smokes Presently	39 (13.2)	7 (13.0)	9 (20.9)	0.39
Diabetes	42 (6.9)	7 (6.3)	8 (11.0)	0.41
Verified Hypertension	179 (29.3)	33 (29.7)	30 (41.1)	0.12
Total Cholesterol	204.0 ± 34.3	206.3 ± 32.5	196.0 ± 32.7	0.10
HDL Cholesterol	46.8 ± 14.1	49.0 ± 13.8	41.0 ± 10.8	<0.001
Prior CV medical history				
CABG	15 (2.5)	2 (1.8)	7 (9.7)	0.003
CAD	50 (8.2)	17 (15.3)	28 (38.4)	<0.001
Angina	35 (5.7)	9 (8.1)	18 (24.7)	<0.001
Unstable Angina	13 (2.1)	9 (8.1)	18 (24.7)	<0.001
COPD	23 (3.8)	1 (0.9)	8 (11.0)	0.003
Restrictive Lung Disease	4 (0.7)	0 (0.0)	2 (2.7)	0.092
CV Accident	10 (1.6)	1 (0.9)	1 (1.4)	0.84
TIA	11 (1.8)	2 (1.8)	1 (1.4)	0.97
Aortic Aneurysm Abdominal/Thoracic	5 (0.8)	1 (0.9)	0 (0.0)	0.73
PVD	9 (1.5)	2 (1.8)	2 (2.7)	0.71
CHF	6 (1.0)	2 (1.8)	5 (6.9)	<0.001
MI	16 (2.6)	7 (6.3)	12 (16.4)	<0.001
Medications				
Beta Blocker	84 (14.8)	17 (15.6)	23 (31.5)	0.001
Calcium Channel Bl	31 (5.5)	8 (7.3)	10 (13.7)	0.026
Ace Inhibitor/ARB	57 (10.1)	12 (11.0)	13 (17.8)	0.14
Diuretic	86 (15.2)	16 (14.7)	24 (32.9)	<0.001
Anticoagulant	11 (1.9)	3 (2.8)	4 (5.5)	0.17
Digoxin	12 (2.1)	2 (1.8)	4 (5.5)	0.19
Nitrate	8 (1.4)	5 (4.6)	8 (11.0)	<0.001
Vasodilator	21 (3.7)	3 (2.8)	2 (2.7)	0.83
Antiarrhythmic	5 (0.9)	1 (0.9)	0 (0.0)	0.72
Antilipemic	101 (17.8)	18 (16.5)	25 (34.3)	0.003

^aThe Rose chest pain and Rose exertional chest pain patients are mutually exclusive. The total number of patients in these two categories adds up to 202. The patients in the Rose chest pain category (n=117) marked "yes" for any chest pain but "no" for pain brought on by exertion.

Table 2

50 search strings used to identify patients with angina pectoris based on the text of clinical notes. The “%” is a wild card search operator that matches any character. Thus, the string “%CP%exertion%” will match “CP on exertion” as well as “CP, exertional.”

Preferred term corresponding to an ICD-9-CM concept			
	Chest Pain (786.5)	Angina (413.x)	Chronic IHD (414.0, 414.8, 414.9)
1	“%chest%pain%”	“%angina%”	“%coron%heart%disease%”
2	“%thoracic%pain%”	“%VAP%”	“%coron%artery%disease%”
3	“%thorax%pain%”		“%coron%art%calcification%”
4	“%precordial%pain%”		“%calcific%coronary%artery%”
5	“%pain%chest%”		“%atherosclerosis%”
6	“%chest%discomfort%”		“%coron%artery%atherosclerosis%”
7	“%atypical CP%”		“%atherosclerotic%coronary%”
8	“%exertional CP%”		“%ASCAD%”
9	“%CP%exertion%”		“%increased%coronary%artery%calcium%”
10	“%anginal CP%”		“%high%coronary%artery%calcium%”
11	“%anginal pain%”		“%CHD%”
12	“%chest distress%”		“%CAD%”
13	“%acute CP%”		“%ASHD%”
14	“%typical CP%”		“%abnormal%CAC%score%”
15	“%chest aching%”		“%coron%calcification%”
16	“%aching chest%”		“%coron%arterioscle%”
17	“%typical chest pain%”		“%coron%atheroscle%”
18	“%chest%pain%”		“%coron%sclerosis%”
19			“%coron%artery%disorder%”
20			“%coron%atheroma%”
21			“%cardiac%atheroma%”
22			“%atherosclerotic%heart%disease%”
23			“%myocardial%ischemia%”
24			“%stenocard%”
25			“%IHD%”
26			“%ISHD%”
27			“%ischem%heart%”
28			“%ischem%heart%”
29			“%cardiac%ischem%”
30			“%coronary%ischem%”

Table 3

True positive rate and true negative rate results of the comparison between using the EMR-NLP and the ICD-codes to identify patients with Rose angina.

	N correctly identified	True positive rate (%) (95% CI)	p-value*	N correctly identified	True negative rate (%) (95% CI)	p-value*
		Positive Chest Pain (n=202)			Negative Chest Pain (n=669)	
EMR	125	62 (55, 67)	<0.001	421	63 (59, 67)	<0.001
ICD	103	51 (44, 58)		488	73 (70, 76)	
EMR (w/o chest pain)	92	46 (39, 52)	<0.001	490	73 (70, 77)	<0.001
ICD (w/o786.5)	68	34 (27, 40)		551	82 (79, 85)	
EMR (with chest, jaw, neck and throat pain)	136	67 (61, 74)	<0.001	393	59 (55, 62)	<0.001
ICD (with 786.5, 723.1, 784.1, 526.9)	109	54 (47, 60)		476	71 (68, 75)	
		Positive Exertional Chest Pain (n=85)			Negative Exertional Chest Pain (n=787)	
EMR	60	71 (61, 80)	0.1	473	60 (57, 64)	<0.001
ICD	53	62 (52, 73)		555	70 (67, 74)	
EMR (w/o chest pain)	49	58 (47, 68)	0.008	564	72 (69, 75)	<0.001
ICD (w/o786.5)	40	47 (36, 58)		640	81 (79, 84)	
EMR (with chest, jaw, neck and throat pain)	64	77 (68, 86)	0.03	438	56 (52, 59)	<0.001
ICD (with 786.5, 723.1, 784.1, 526.9)	56	66 (56, 76)		540	69 (65, 72)	
		Positive Rose Angina (n=8)			Negative Rose Angina (n= 864)	
EMR	7	88 (65, 1)	**	497	58 (54, 60)	<0.001
ICD	7	88 (65, 1)	—	586	68 (65, 71)	
EMR (w/o chest pain)	6	75 (45, 1)	>0.5	598	69 (66, 73)	<0.001
ICD (w/o786.5)	7	88 (65, 1)		684	79 (77, 82)	
EMR (with chest, jaw, neck and throat pain)	7	88 (65, 1)	—	458	53 (50, 56)	<0.001
ICD (with 786.5, 723.1, 784.1, 526.9)	7	88 (65, 1)		568	66 (63, 69)	

* p-value for comparing the EMR to the ICD

** p-value could not be established due to perfect agreement between EMR and ICD methods