Evaluating Healthcare Quality Using Natural Language Processing

Karen Brandt Baldwin

Abstract: Consistent monitoring for quality indicators as adverse events or missed screening opportunities remains a difficult proposition for most healthcare organizations. Much of the clinical data needed for quality reports is imbedded in narrative reports in the electronic health record. Narrative data most often require costly retrieval by manual data extraction. NUD*IST, a qualitative research computer program, was used as an automated natural language processing tool to extract and code data for analysis of screening and treatment for breast cancer. The study method demonstrated acceptable levels of precision and recall compared to large-scale natural language processing programs.

Key Words
electronic health record
natural language processing
quality of care

The need for regular monitoring of healthcare services has never been greater. In 2003, the Institute of Medicine (IOM) released a report describing the impact of medical errors on patient morbidity and mortality. Reporting of adverse events has traditionally been difficult to achieve (Cao, Stetson, & Hripcsak, 2003), even following the debate sparked by the IOM report. Missed opportunities for preventive screening may complicate the issue by creating situations in which patient conditions go unrecognized until adverse events occur. The most critical step in preventing errors and addressing missed screening opportunities is identification of gaps in service. Gaps in healthcare services are rarely identified in isolated quality assessments, however. Routine measurement of quality is more likely to recognize errors and gaps in service because data are regularly reported and analyzed over time.

Making routine measurements of healthcare quality is a difficult, time-intensive, and labor-intensive enterprise (Bates et al., 1999). Readily available databases such as administrative databases are limited in the amount and detail of clinically significant data. This limitation is a salient issue, given that clinical data are particularly useful in identifying gaps in healthcare services and adverse events.

The best clinical data are found in the electronic health record (EHR), considered the gold standard of data on healthcare quality (Cao et al., 2003; Iezzoni, 1997). Clinical data used in quality evaluations are present as flowcharts, integrated progress notes, and discharge summaries, as well as in radiology and pathology reports. These summaries and reports often contain explanations for missed screening opportunities, details of adverse events, and plans of care that deviate from recommended guidelines. The summaries and reports are most often recorded in narrative format in the EHR (Wilcox, Phil, & Hripcsak, 1999; Zeng et al., 2006). Narrative text provides insight into the quality of care that would otherwise be missed because it represents the thoughts of the provider, unrestricted by structured vocabularies found in many EHRs.

The difficulty with narrative text lies in data retrieval. Natural language processing allows retrieval of information locked in narrative reports and makes it available to healthcare professionals. Although some automated retrieval systems of narrative data exist, the majority of healthcare systems continue to rely on manual retrieval methods to extract narrative data because of the cost involved in the development of EHR automated systems (Hazlehurst, Mullooly, Naleway, & Crane, 2005). Healthcare organizations without automated natural language processing systems may not be able to provide routine quality reports on all areas of interest simply because of the cost and difficulties inherent in manual data retrieval of clinical data.

Background

Natural language processing (NLP) is a method of using computer programming for linguistic analysis of narrative text (Travers & Haas, 2003). In NLP, a computer software program breaks up the narrative text (termed *segmentation*) into sentences, phrases, or words, which are then mapped to a structured system of codes that can be analyzed by the computer.

For example, an NLP system may use a diagnosis coding system such as the *International Classification of Diseases (ICD)* codes as the structured framework for coding narrative

Journal for Healthcare Quality
Vol. 30, No. 4, pp. 24–29
© 2008 National Association
for Healthcare Quality

text. In this example, a discharge summary is imported into the software program, which searches the narrative discharge note for words or phrases that match the *ICD* codes for diagnosis noted in the EHR. When all matches are made, the program reports all the *ICD* codes that were present in the discharge summary. An NLP system such as this saves thousands of hours of manual assignment of *ICD* codes. *ICD* codes and Universal Medical Language Systems (UMLS) are the most commonly used structure systems in large-scale NLP systems (Hazlehurst, Frost, Sittig, & Stevens, 2005; Pratt & Yetisgen-Yildiz, 2003).

NLP systems face a number of challenges, primarily related to the difficulties in the consistent processing of unstructured text data (Pratt & Yetisgen-Yildiz, 2003). Difficulties arise when text contains phrases that have the same meaning but are entered into the EHR differently. For example, the triage note "head injury" has the same meaning as "injury to head," and yet an NLP system would code these entries as unique phrases.

Segmentation is another problem with natural language processors. Segmentation is the breaking apart of narrative text into identifiable components. An NLP system that is programmed to break text into sentences using the indicator of a period at the end of a sentence will have problems if a provider uses periods as abbreviations. Spelling errors, prefixes, and suffixes—all these naturally occurring parts of natural language pose potential problems for NLP systems.

One approach when dealing with these issues is to limit the vocabulary used in the narrative text. NLP programmers have had success creating systems that code radiology and pathology reports (i.e., RadTRAC [Radiology Text Report Analyzer and Classifier]) because the reports rely heavily on standardized terminology (Mendonca, Haas, Shagina, Larson, & Friedman, 2005). Programmers have had limited success in retrieving data from progress and discharge records, which typically have wide variations in free-text narratives. NLP systems such as MedLEE (Medical Language Extraction and Encoding System) that code the entire narrative text from all areas of the EHR are very costly to develop and take years to perfect, because the programs are unique to each EHR (Hripcsak et al., 1995).

For most acute and primary care organizations, the investment in NLP systems designed for their EHR is not financially feasible. A modest computer system adaptable to multiple quality initiatives and EHRs would be an advantage to organizations that need to collect data for quality, safety, or surveillance reports. NUD*IST (Nonnumerical, Unstructured Data Indexing, Searching, and Theorizing; the latest version is called N6) is a software program designed to assist researchers in categorizing narrative text into themes for analysis in qualitative research. Although NUD*IST has not been used as a natural language processor, the program contains several features that may make it useful as a tool for quality assessments.

The specific aims of this study were to (1) determine the efficiency of the study NLP method compared to manual coding or retrieval of breast cancer prevention data from the EHR and (2) compare values of recall and precision of the study NLP method to those of other large-scale NLP systems.

Study Method

Setting and Sampling

The setting for this study was a large Midwestern academic medical center's women's health center; approximately 3,200 patient visits are made to the center annually. A stratified convenience sample was taken of the EHRs of 60 women, ages 40 and older, who visited the Women's Health Center in 2001. Women of ages 40 and older were chosen to allow comparisons of present screening practices to National Cancer Institute (NCI) guidelines for women ages 40 and older. The only exclusion criterion was pregnancy, because it may affect provider decisions about screening, treatment, and follow-up for positive breast cancer diagnoses.

The study time frame was a 2.5-year period beginning in 2001 and ending in July 2003. NCI guidelines recommend routine screening every 2 years for women ages 40 and older. The extension of 6 months captured potential follow-up for suspicious or abnormal findings of tests performed in late 2002.

The sample size was designed to limit the amount of text NUD*IST would be importing and coding. It was anticipated that with a sample size of 60 and a time frame of 2.5 years, NUD*IST might process up to 7,500 words. This sample size was considered adequate for testing NUD*IST as a method for natural language processing.

Variable Selection

Variables in this study included demographic risk factors for breast cancer, provider interventions, and outcome variables related to breast cancer screening and treatment. Demographic variables included age, race, and gender. Although the EHRs only of women were used in this study, gender was included as a variable to test the accuracy of the experimental data-retrieval strategy. Gender should be retrieved as "female" for each record.

Risk factors included positive genetic screening and positive family and personal history of cancer (NCI, 2002). Current breast complaints were included because breast pain, masses, and nipple discharge may indicate a need for specialized testing and follow-up care.

Provider variables included physical exams, biopsies, treatments, and recommendations for diagnostic testing and follow-up visits. Because a woman may choose not to follow up with recommendations for a mammogram or other diagnostic tests, retrieving information that a mammogram has or has not been performed was insufficient to assess the quality of a provider's intervention. Therefore, this study included variables that directly reflect the provider's intentions: "recommend mammogram," "order given for mammogram," "referral" (to capture referrals to oncologists or other specialists), and "recommend follow-up visit."

Outcome variables predominantly reflected testing or physical exam results. A possible, though typically long-range, outcome of breast cancer is mortality; this variable was therefore included in the study.

Methodology

A hierarchical model similar to a medical decision tree was developed in NUD*IST 4.0 using the variables described. Following institutional review board approval, narrative data from the Demographic, Progress Notes, and Result sections of the EHR were transposed into sentence-length strings using the Edit function in Microsoft Word for Windows.

After data were segmented into sentences, they were imported into NUD*IST using an automated command programmed by the author. A second automated command coded the narrative data to the hierarchical model containing the variables of interest. To determine how accurately NUD*IST functioned as a natural language processor, the same data were manually coded, using the same hierarchical quality framework.

Statistical Analysis

Efficiency is an important consideration in routine quality assessments. If an NLP system cannot perform more efficiently than the traditional manual retrieval of quality of care data, it will not be cost-effective. In this study, efficiency was calculated on the basis of the time the NLP study method took to import and code data compared to manual coding of the same data.

NUD*IST was used as an NLP tool; it was therefore important to use evaluation methods appropriate for NLP systems to test its success. Two methods of evaluation consistently noted in the NLP literature are precision and recall (Chinchor, Hirschman, & Lewis, 1993; Sager, Lyman, Tick, Than Nhan, & Buckman, 1994). Precision and recall are components that demonstrate the ability of an NLP system to accurately and consistently retrieve narrative data. Accuracy is measured by evaluating recall and precision of the system. *Recall* may be defined as the ability of the system to locate and code all of the desired variables. *Precision* refers to the system's ability to locate and code only desired variables without coding undesired variables. The evaluations of recall and precision address the essential question of whether the information generated by a natural language processor is reliable.

According to Sager and colleagues (1994), *precision* may be determined by the amount of information correctly retrieved from a database. This measure may be presented as a rate:

$$\frac{\text{number of documents for which desired information is retrieved}}{\text{number of documents for which desired and undesired information is retrieved}}$$

The denominator in this calculation equals the total of the number of documents with correct processing plus the number of documents with errors due to incorrect processing. Incorrect processing may be noted when there are missing data, incorrectly coded data, or incompletely coded data. The numerator of this equation equals the number of entries that existed in the original data that were completely and correctly coded.

Recall refers to the degree to which an NLP system completely downloads and codes data (Sager et al., 1994). If *precision* reflects accuracy of the processing system, *recall* specifies the consistency of the system. Recall may also be expressed as a rate:

$$\frac{\text{number of documents with desired information that are retrieved}}{\text{total number of documents in which desired information is present in database}}$$

It can be noted that the closer this equation approximates 1, the more consistent the results.

Results

Efficiency of the Study NLP Method

To measure the efficiency of the study NLP method, time was recorded for two phases of data retrieval. Phase 1 included the time for identifying and collecting the desired variables. Phase 2 included the time for coding and formatting the collected variables for analysis. **Table 1** summarizes the time measurements for manual retrieval versus retrieval using the study NLP method.

Results of paired *t* tests indicated a significant difference ($p < .05$) between the study NLP method and the manual method for total efficiency. The study NLP method demonstrated greater efficiency, although the time spent writing automated commands was not included in calculations because it was a single event that would not be repeated with each EHR. Nonetheless, writing all automated commands for coding took 118 hours. If this time is included in the overall efficiency calculation, the study NLP system is much less efficient than the manual method.

Frequency of Reported Variables: Study NLP Method

This study compared the study NLP method to the manual retrieval method in three main categories: demographics, risk factors, and screening and test results or follow-up. The categories represent variables of interest in the quality evaluation and an area in which an NLP system should demonstrate consistent coding. Frequencies for retrieving gender and race variables were 100% for both methods, with no missing variables. Pearson's chi-square analysis indicated that, overall, the study NLP method demonstrated no significant difference ($p = .156$ with 3 degrees of freedom) from the manual retrieval method.

The presence of risk factor variables in the study population was low, although some variables were reported in the EHR. The EHR did not require the provider to enter data into the Risk Factor field. No risk factors were noted in this field; however, some risk factors were independently reported in the Progress Notes. The risk factor reporting rate was too low to merit comparative analysis.

Comparisons of breast cancer screening data showed differences between the study NLP method and manual data retrieval. Screening variables included "mammogram completed," "mammogram ordered," "mammogram recommended," "previous mammogram," "mammogram refused," "clinical breast exam," and "CBE" (clinical breast exam). NCI recommends that both mammography and clinical breast exams be performed. For this reason, results included findings for reporting of both screening methods **(Table 2).**

Chi-square tests determined a significance level of $p = .000$ ($p < .05$), indicating a difference between methods. Further analysis of the study NLP method revealed a false positive rate of 0 and a false negative rate of 0.35 (21/60). These findings suggest that although the study NLP method did not code data that were not present, it did have a high percentage of missing data that were not coded. Differences in methods were primarily due to missed automated coding of "recommended mammogram," "previous mammogram," and "mammogram refused."

Errors in collecting clinical breast exam data by the study NLP method were primarily re-lated to differences in the providers' description of the clinical breast exam, primarily in the use of the acronym "CBE" or "breast exam." NUD*IST had more difficulty collecting data when the breast exam was included in the general physical exam, although phrases used in the physical exam (e.g., "breast exam normal," "no

Table 1. Comparison of Methods: Data Retrieval Times

Phase	Study NLP Method (Minutes per EHR)	Manual Retrieval Method (Minutes per EHR)
Phase 1: Identification and gathering of data	3.1	5.0
Phase 2: Coding and formatting data for analysis	2.3	3.8
Total	**5.4**	**8.8**

Note. EHR = electronic health record; NLP = natural language processing.

Table 2. Comparison of Methods: Screening per EHR

Mammogram and Clinical Breast Exam	Study NLP Method	Manual Retrieval Method
Present	15	36
Not present	45	24
Total	**60**	**60**

Note. EHR = electronic health record; NLP = natural language processing.

nipple discharge") were included as automated commands for data capture.

Regardless of the method of data collection, the screening rates for mammography and clinical breast exams were lower than those recommended by the NCI guidelines. Other variables such as breast biopsy and treatment options (e.g., tamoxifen, chemotherapy, or radiation therapy) were not present in any EHR.

The researcher also observed follow-up (repeat visit in less than 1 year) and referral rates for women with abnormal or suspicious findings on the mammogram or clinical breast exam. It was expected that in all cases of abnormal or suspicious findings, some follow-up would be noted in the EHR **(Table 3).**

The results of the chi-square analysis suggest no significant differences between the study NLP method and the manual method of data retrieval. The additional counts for the study NLP method were the result of false–positive reporting of variables.

Variables related to test results (mammograms, clinical breast exam, breast ultrasound results) were reported 100% of the time with both methods. No breast biopsy results or mortalities were reported in this time frame.

Recall and Precision of the Study NLP Method

Evaluation of the study NLP method using the manual retrieval method to verify accuracy revealed a recall rate of .293 (144 correctly coded variables from 491 desired variables). This rate is low compared to that achieved by seasoned large-scale NLP systems, which typically have recall and precision rates of 80%–87% (Romacker, Hahn, Schulz, & Klar, 2000).

The precision rate for the study NLP method was .709 (144 correctly coded variables divided by 203 correct and incorrect variables). Precision rates for this trial of the study NLP method were acceptable in comparison to precision rates of large-scale systems.

Table 3. Comparisons of Methods: Follow-Up for Abnormal or Suspicious Findings

	Study NLP Method	Manual Retrieval Method*
Number of patients with follow-up for abnormal or suspicious findings	4	2

Note. NLP = natural language processing.
*Pearson's chi-square value = .702; degree of freedom = 1; p = .402.

Discussion

The study method demonstrated moderate success as an NLP system. As with larger systems, the study method fared better in retrieving variables that were narrowly and consistently defined. Variables such as gender, race, and test results remained consistent over time and demonstrated the highest rates of recall and precision.

Other variables proved more challenging to retrieve. Variables that had multiple definitions were difficult to capture because of the variations in the terminology that providers used to express the same concept. The variations in reporting normal findings in the clinical exam created problems for the study method. As was expected, precision rates were acceptable, although recall rates were poor. This finding suggests that when automated codes matched the providers' descriptions of the variables, data retrieval was successful. Recall rates were poor because of the high rate of nonretrieved narrative phrases.

Risk factor variables were also difficult to retrieve for two reasons. First, provider descriptions varied in much the same way that screening variables varied. Second, risk factor retrieval was hindered by the lack of reporting in the EHR. Until providers routinely report risk factors, variables will be unavailable for quality initiatives. Part of the challenge in retrieving data from the EHR was related to using NUD*IST to create automated retrieval and coding of data. NUD*IST is designed for manual coding of qualitative data, and the programming function of NUD*IST 4.0 is not well developed. Other software programs designed for qualitative research (e.g., Atlas) are available but were not used as NLP tools in this study.

These improvements notwithstanding, the study method demonstrated respectable efficiency with variables that were well defined. NUD*IST performed moderately well when used as a tool for NLP. Once the automated commands were written, data retrieval and coding were significantly faster than manual retrieval and coding of the data. The efficiency of autocoding was directly related to the precise identification of the variables. As is typical with large-scale NLP programs, missed words, misspellings, and variations in sentence structure contributed to decreased recall. Understanding of medical terminology was critical to identifying variables that captured the health concept in question.

This method could be adapted for a number of uses in healthcare applications, including routine assessments of quality initiatives, such as unit-based indicators that are analyzed on a quarterly basis or surveillance monitoring of the EHR for potential adverse events. It may also be useful for identifying adherence to standardized treatment protocol after a new protocol has been introduced. It is not known whether NUD*IST is compatible with software used in healthcare settings. Future research may explore this possibility with other programs associated with the EHR. This study demonstrated modest results; the sample size was limited, and the programmer was a novice. Additional research with this method using a larger EHR database and improved programming of automated commands will provide a better understanding of the usefulness of this method for quality-of-care initiatives.

Conclusion

Retrieval and coding of healthcare data are essential components in improving healthcare quality and maintaining surveillance systems to monitor adverse events. This study tested a method of automated data retrieval and coding of breast cancer screening to determine whether NLP could improve efficiency without sacrificing reliability in reporting of data. NUD*IST, a software program used primarily in qualitative research, performed well when retrieving well-defined data but less well when retrieving all desired data variables. Improvements in the EHR, such as consistent reporting of risk factor data in a prescribed section of the health record, could increase the efficiency of data retrieval. The study method of NLP shows promise in easing the data burden created by the need for manual extraction of data for quality initiatives.

References

Bates, D. W., Pappius, E., Kuperman, G. J., Sittig, D., Burstin, H., Fairchild, D., et al. (1999). Using information systems to measure and improve quality. *International Journal of Medical Informatics, 53,* 115–124.

Cao, H., Stetson, P., & Hripcsak, G. (2003). Assessing explicit error reporting in the narrative electronic medical record using keyword searching. *Journal of Biomedical Informatics, 36*(1–2), 99–105.

Chinchor, N., Hirschman, L., & Lewis, D. (1993). Evaluating message understanding systems: An analysis of the Third Message Understanding Conference of Computer Linguistics. *Journal of AHIMA/American Health Information Management Association, 19*(3), 409–450.

Hazlehurst, B., Frost, R., Sittig, D. F., & Stevens, V. J. (2005). MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical records. *Journal of the American Medical Informatics Association, 12,* 517–529.

Hazlehurst, B., Mullooly, J., Naleway, A., & Crane, B. (2005). Detecting possible vaccination reactions in clinical notes. *Proceedings of AMIA* [American Medical Informatics Association] *Annual Symposium,* 306–310.

Hripcsak, G., Friedman, C., Alderson, P., DuMouchel, W., Johnson, S. B., & Clayton, P. D. (1995). Unlocking clinical data from narrative reports: A study of natural language processing. *Annals of Internal Medicine, 122,* 681–688.

Iezzoni, L. (1997). *Risk adjustment for measuring health care outcomes.* Chicago: Health Administrative Press.

Institute of Medicine. (2003). *Patient safety: Achieving a new standard for care* (P. Aspden, J. M. Corrigan, J. Wolcott, & S. M. Erickson, Eds.). Washington, DC: National Academies Press.

Mendonca, E. A., Haas, J., Shagina, L., Larson, E., & Friedman, C. (2005). Extracting information on pneumonia in infants using natural language processing of radiology reports. *Journal of Biomedical Informatics, 38,* 314–321.

National Cancer Institute. (2002). *Mammography guidelines.* Retrieved November 12, 2003, from http://newscenter.cancer.gov/pressreleases/mamstatement.31jan01html.

Pratt, W., & Yetisgen-Yildiz, M. (2003). Litlinker: Capturing connections across the biomedical literature. *Proceedings of International Conference of Knowledge Capture,* 105–112.

Romacker, M., Hahn, U., Schulz, S., & Klar, R. (2000). Semantic analysis of free text. *Proceedings of MIE 2000,* 438–443.

Sager, N., Lyman, M., Tick, L., Than Nhan, N., & Buckman, C. (1994). Natural language processing of asthma discharge summaries for the monitoring of patient care. *Proceedings of AMIA* [American Medical Informatics Association] *Annual Symposium,* 265–268.

Travers, D. A., & Haas, S. W. (2003). Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. *Journal of Biomedical Informatics, 36*(4–5), 260–270.

Wilcox, A., Phil, A., & Hripcsak, G. (1999). Classification algorithms applied to narrative reports. *Proceedings of AMIA* [American Medical Informatics Association] *Annual Symposium,* 455–459.

Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., & Lazarus, R. (2006, July 26). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making, 6,* 30. Retrieved March 30, 2007, from www.biomedcentral.com/1472-6947/6/30.

Author's Biography

Karen Brandt Baldwin, PhD RN, is an assistant professor at Northern Illinois University School of Nursing, DeKalb, IL. She has spent 15 years in acute healthcare, including 6 years as a clinical nurse specialist responsible for hospital-wide quality evaluations.

For more information on this article, contact Karen Brandt Baldwin at kbaldwin@niu.edu.