Can Natural Language Processing Aid Outpatient Coders?

Save to myBoK

by Homer R. Warner, Jr., PhD

As HIM vendors strive to perfect automated coding, one supporting technology has reached an exciting stage of development. Natural language processing (NLP) is a major challenge for the computer industry because of the complexity and variability of human speech. NLP differs from simple word search programs because it considers sentence structure (syntax), meaning (semantics), and context to accurately interpret a physician's note.

For example, good NLP systems can distinguish between "no shortness of breath, chest pain aggravated by exercise" and "no chest pain, shortness of breath aggravated by exercise," which look the same to a word search engine. NLP systems also know that the toe is part of the foot. Further, some systems also claim to be able to discriminate between "the patient thinks they have angina" and "doctor thinks the patient has angina," which have different meanings from a coding perspective.

Promising new NLP products are beginning to emerge in certain medical arenas. Emergency medicine is an excellent proving ground for NLP because a significant percentage of emergency room (ER) charts are dictated and transcribed into ASCII text, which is a prerequisite for NLP systems.

Today's NLP autocoding promise is to provide improved data management productivity and consistency without sacrificing coding accuracy. To validate the claims made by an NLP vendor on these two points, 3M Health Information Systems designed and performed a study of ICD-9-CM and CPT Evaluation and Management (E&M) coding in ER charts.

Study Components

The NLP system used for the study requires that the text have some structure, including section headers like "History of Present Illness," "Physical Exam," and "Diagnosis," so it can separate subjective from objective data. An acute care hospital serving as a regional trauma center provided emergency medicine data in 996 sequentially selected charts. Patient and provider names were removed from all patient data used in the study to ensure privacy and confidentiality.

All charts used in the study had been previously dictated and existed in a database in ASCII format. These transcribed charts were taken to the NLP vendor's data center for batch processing. The NLP output consists of computer-generated ICD-9-CM and E&M codes, and each code is linked to the part of the medical record that justifies the code assignment, thus providing an audit trail. The batch processing of all 996 study cases took the NLP engine about four hours to complete.

The NLP engine processed the charts and divided them into three categories:

- **red cases** (13 percent—charts that did not meet the NLP's formatting requirements, requiring the client to remedy the formatting and resubmit the document or code it manually
- **yellow cases** (54 percent)—the NLP was not sufficiently "confident" in assigning a code. For example, no association was found between diagnosis and CPT procedure codes. These are returned to the client for manual coding along with the NLP analysis and audit trail
- **green cases** (33 percent)—the NLP was confident in its ability to assign the correct codes and no human intervention was required

Because by definition red and yellow cases require human review, 102 (31 percent) of the 328 green cases were examined to determine accuracy and productivity.

The study participants were three coders with six to 15 years of hospital coding experience but not extensive ER coding experience, a coder with 12 years of ER coding experience, and two NLP technologists who submitted the transcribed charts to their NLP system. At the beginning of the study, the three hospital-experienced coders worked together in coding a preliminary set of ER charts to become familiar with E&M coding methodology, using 1995 E&M HCFA/AMA documentation guidelines. A consensus code was established among the three coders when the codes assigned by at least two of the three coders matched.

Then, each coder worked on cases independently to code the 102 charts and the codes produced by each participant were scored. For E&M coding, a score of either 0 (no match) or 1 (match) was assigned for agreement between various participants. Kappa statistics were used for pair-wise comparisons because they provided a conservative way to compare how well one coder agreed with another, taking into account chance agreement.[1] If there was partial agreement (for example, if NLP agreed with a coder on each of the ICD codes but not in the order of the codes, then partial agreement score was given (.10 to .90). A Kappa score greater than .70 is considered excellent agreement, between .40 and .70 is considered moderate agreement, and below .40 is considered poor agreement.

Variance between the already-assigned hospital codes and those assigned by the other participants may be explained in part by the fact that the hospital coder had access to more information about each case than was in the dictated note.

ICD-9-CM Coding Agreement

Perfect Agreement on Primary ICD	Experienced Hospital Coders (Group A)	NLP vs. Group A Consensus	Group A vs. ER-experienced Coder (B)	Group A vs. Hospital	Coder B vs. NLP	NLP vs. Hospital	Coder B vs. Hospital
63%	91%	90%	86%	82%	82%	75%	75%

Assessing NLP Accuracy

ICD-9-CM Coding

"ICD-9-CM Coding Agreement," above, shows the extent of agreement for ICD-9-CM coding among the study participants. Perfect agreement among all participants occurred in only 63 percent of cases. The highest agreement score was found among the three hospital-experienced coders (Group A) at 91 percent, followed by the agreement between the NLP engine and Group A's consensus codes at 90 percent. Then, Group A and the experienced ER coder (Coder B) produced 86 percent agreement. Group A's consensus and the hospital's assigned codes matched 82 percent of the time, as did Coder B with the NLP system. The lowest levels of concordance were observed between NLP and the hospital and the hospital and Coder B, both at 75 percent.

Reasons for disagreement varied. When the coders in Group A disagreed with each other it was usually because one coder either added or omitted a code or one of the three coders choose a different ICD-9-CM code than the other two. Group A disagreed with the NLP 10 percent of the time for some of the same reasons but also for the following additional reasons:

1. **The NLP coded symptoms even after a diagnosis was established.** For example, Group A coded 300.00, anxiety state, unspecified, and NLP coded both 786.50, unspecified chest pain, and 786.05, shortness of breath, in addition to coding 300.00. The ER coder coded 786.50 and 300.00 for this case. Coding symptoms integral to the disease process reflected in a diagnosis code already assigned violates ICD-9-CM coding guidelines. When the symptom is unrelated or does not always occur with a specific diagnosis, then the symptom may be reported in addition to the primary code.

2. **Differences in the fourth and fifth digit for some codes** (e.g., NLP coded 250.00, diabetes type II, NIDDM, while Group A coded 250.01, diabetes type I, IDDM, as supported by the dictated note).

3. **NLP coded the physician's probable diagnosis.** For example, NLP coded 442.9, aneurysm of unspecified site, while Group A coded 784.0, headache (facial pain), 780.2, syncope and collapse, V42.0, kidney replaced by transplant, 753.12, polycystic kidney, unspecified type for the same case. According to official coding guidelines for outpatient reporting, it is not appropriate to code any condition that is still "probable," "suspected," or still to be "ruled out."

E&M Coding
On average, 43 percent of the study cases were coded level 4 (99284) and only 7 percent were coded level 1 or 2 by the participants. "Agreement Among Group A Coders," below, points out the variation in E&M coding among the three novice E&M coders. There was moderate-to-good agreement between coders 1 and 2 but poor agreement between coder 3 and the other two. Coder 3 had the least coding experience of the group, which could partly explain the variation. Perfect agreement among these three coders occurred in only 43 percent of cases, while coding agreement between at least two of the three coders occurred in 97 percent of the cases.

Agreement Among Group A Coders

.	**Coder 1 vs. Coder 2**	**Coder 2 vs. Coder 3**	**Coder 1 vs. Coder 3**
Agreement:	76%	57%	52%
Kappa:	0.64	0.37	0.32

Perfect agreement among all participants was only 18 percent. The best agreement was found between Group A and NLP with a kappa of .68 followed by Group A and the hospital-assigned codes (kappa: .48) (see "Pair-wise E&M Agreement," below). Group A consensus agreed perfectly with the experienced E&M coder (Coder B) 64 percent of the time (kappa: .37) while the hospital agreed perfectly with NLP 62 percent of the time (kappa: .40). Coder B compared to NLP and in particular the hospital compared to Coder B had very low kappa scores, suggesting wide disagreement in the E&M coding. Of the 49 percent of cases where the hospital codes differed from Coder B, eight differed by more than one E&M level (e.g., 99283 vs. 99285) in contrast to only one when comparing Group A with NLP.

Pair-wise E&M Agreement

.	**Group A vs. NLP**	**Group A vs. Hospital**	**Hospital vs. NLP**	**Group A vs. Coder B**	**NLP vs. Coder B**	**Hospital vs. Coder B**
Agreement:	80%	90%	86%	82%	82%	51%
Kappa:	0.68	0.48	0.40	0.37	0.31	0.16

Furthermore, the most experienced coders participating in the study were also the most aggressive in assigning E&M levels (e.g., assigning level 5 over level 4). This could mean that higher reimbursement results from using more experienced E&M coders provided that their codes are accurate.

Agreement levels among all study participants was higher for diagnosis code assignment than for E&M visit level coding. E&M coders, human or automated, agreed with each other somewhere between 43 to 78 percent of the time. Because E&M coding guidelines are general rules for measuring and categorizing the work of clinicians, the specific application of the guidelines is subject to human interpretation.

Does NLP Meet the Coding Gold Standard?
Without an established coding "gold standard" for the charts used in this study, it is impossible to measure the coding accuracy of any of the participants. However, one could argue that in the absence of a "gold standard," which is often the case when an HIM director is trying to asses the coding skills of an employee, the degree of concordance with other coding professionals could be considered to be a reasonable measure of accuracy. Using this metric, NLP compares favorably with the other participants in the study.

Productivity Assessment for Emergency Medicine Charts

# Charts	Average # Minutes/Chart Spent by Group A Coders w/o NLP	Average # Minutes/Chart Spent by Group A Coders Checking NLP Notes	Time Difference in Minutes	Percent Improvement using NLP
102	6.32	3.29	3.03	48%

Assessing NLP Productivity

The next critical question is how much more efficient can the coding process become using NLP tools? In this study, productivity for emergency medicine coding was considered for two scenarios:

- An ER department treats the green cases like yellow cases and reviews them, making NLP a coder's decision-support tool.
- An ER department allows the NLP to process the charts in batch and reviews none of the green charts.

To measure coder productivity, each coder in Group A coded a set of charts (from the 328 "green" ER cases) assisted by off-the-shelf computer coding programs and coding manuals. Then each coder coded a different set of charts (from the same batch of cases) assisted by the NLP coder review workstation. This workstation allowed the coders to see the codes assigned by the NLP along side the text that correlated with each code. For cases in which the coder did not agree with the NLP code, the coder referred to other coding tools to determine the correct code. Each coder received one hour of training on the NLP workstation before the timed trial. The elapsed time for both steps was noted and is shown at above, in "Productivity Assessment for Emergency Medicine Charts." When NLP was used as a reference, coder productivity improved by 48 percent.

The second measure of coding productivity was based on the assumption that NLP can autocode the "green" charts with sufficient accuracy that no human intervention is required. Under this assumption, NLP was able to code all 328 "green" ER charts in approximately 82 minutes or 15 seconds per chart. When considering that the 328 charts represented 33 percent of the total emergency cases submitted to NLP, then an NLP client might expect a 33 percent reduction in coding workload when autocoding and not reviewing the cases, combined with a 48 percent productivity improvement when using NLP assistance for cases requiring review.

Buy or Wait?

The decision to begin using a new technology involves more than whether the technology produces accurate results and is faster as demonstrated by experimental results in a laboratory setting. A potential buyer must consider how the new technology will influence overall workflow and what the ripple effect of change will be on staff and budgets. Optimizing workflow must include providing adequate documentation and producing accurate coding, reducing physician time in the documentation process, reducing cost, and optimizing revenue. The cost of transcription services, turnaround time, and software, hardware, and interface costs all have to be considered.

Scale must also be a consideration in this decision. A large-volume ER department with 70,000 visits a year or an outsource physician billing service company with more than 1 million ER charts a year might see NLP as a way to reduce FTEs or take on more business without increasing labor costs. On the other hand, a small-volume healthcare provider with less than 25,000 ER visits a year would need to find other tasks for their outpatient coder to effectively capitalize on the increase in coding speed.

NLP in the Future

NLP programs have the advantage in being predictable, programmable, and fast. While automated coding systems still need improvement before they will meet hospital standards on coding accuracy, they may be able to provide an immediate benefit as decision-support tools for emergency medicine coding. The performance results from this study provide support for using NLP technology in coding because coding speed, like coding accuracy, affects expeditious reimbursement.

Emergency medicine is not the only medical domain where NLP is adding value. Radiology is emerging rapidly and there has been some significant progress in natural language understanding of discharge summaries, which will prove valuable for inpatient coding as well as data mining in the future. For now, with the shortage of qualified outpatient coders to meet the APC-related increased demand for professional services coding, HIM directors may want to consider NLP technology to improve coding productivity.

Note

1. Fleiss, JL. *Statistical Methods for Rates and Proportions.* 2d ed. New York: John Wiley & Sons, Inc., 1981.

References

Chao J. et. al. "Billing for physician services: A comparison of actual billing with CPT codes assigned by direct observation." *Journal of Family Practice* 47, no. 1 (1998).

Cohen, R.H. "Using an Expert System to Ensure Accurate Third-Party Reimbursement." *Patient Accounts* 17, no. 2 (1994).

Morgan, John D. "Computer-assisted encoding." *Topics in Health Record Management* 2, no. 3 (1982).

Morgan, John D. "Automating the Code Book in Not Enough." *Journal of the American Medical Record Association* 57, no. 2 (1986).

Morris, W.C., et al. "Assessing the Accuracy of an Automated Coding System in Emergency Medicine." *Journal of the American Medical Informatics Association*, publication pending, November 2000.

Homer Warner, Jr. *is a manager of business development at 3M Health Information Systems. He can be reached at hrwarner@mmm.com.*

Article citation:
Warner, Homer, Jr. "Can Natural Language Processing Aid Outpatient Coders?" *Journal of AHIMA* 71, no.8 (2000): 78-81.