Evaluating E&M Coding Accuracy of GoCode as Compared to Internal Medicine Physicians and Auditors

by Rhonda R. Thomas, PhD, MBA; Constantin Mitocaru; and Judith Hanlon

Introduction

Change is continually upon us, and the challenge is to leverage technology to become more efficient professionals, better problem solvers, and true knowledge workers, using technology to enable us to work smarter, faster, and more accurately.

Reviewing free text has been a time-consuming and tedious manual task. Humans were needed to understand the meanings of synonyms, acronyms, and complex conceptual expressions. Statistics-based "trigger words" like *complications* or *history* often do not identify synonyms or closely related terms.[1] Just as technology has raced ahead of consumers in the past decade, it has also been being racing forward in development (albeit sometimes quietly and behind the scenes) for revolutionary new work in healthcare. New advances in natural language processing (NLP) allow the computer to "understand" what clinicians mean in their documentation, with little regard to how they say it.

This preliminary study examines both the qualitative and quantitative impact of new NLP technology, specifically found in GoCode, for assisting HIM professionals in achieving a more standardized interpretation of evaluation and management (E&M) coding guidelines in internal medicine encounters. Observations on how this technology impacts the coding behavior of human reviewers will also be tracked and discussed, both qualitatively, from their own words, and quantitatively, from observing changes to their subsequent coding behavior during and after their interaction with the system.

Background

Most healthcare payors require the clinician to provide a CPT (Current Procedural Terminology)[2] visit code in order to bill for patient services in the evaluation and management of their care. In fact, reimbursement from payors is largely based on the associated evaluation and management (E&M) level reported and supported by the exam documentation for reimbursement for a patient visit. However, CPT guidelines are in fact guidelines, subject to regional and individual interpretation. In fact, the complexity and variations in interpretations of CPT E&M levels is well documented in various studies.[3, 4]

Unstructured, free-text documentation of the patient encounter is certainly the norm throughout the largest single healthcare enterprise in the United States, the federal government, especially the Veterans Health Administration[5] and the Department of Defense.[6] Free-text documentation (from dictation, transcription, or voice recognition) is the preferred method of documentation currently and thus the "norm in most US practice settings."[7] However, extracting meaning from unstructured text is a complex task, previously only possible by human reviewers, of understanding syntax, negation, synonyms, and prose to unlock the meaning contained in narrative documentation.

Research Hypothesis: Null Hypothesis

The study sought to disprove the following null hypothesis: Semantic- and syntactic-based natural language processing technology, combined with a sophisticated medical coding engine known as GoCode, has no impact on improving the accuracy of human coder reviews for E&M coding of free-text patient encounter documents.

Methodology

The test production environment of a major, multispecialty medical ambulatory clinic was polled to extract all internal medicine patient encounter notes processed by the GoCode system from May 1, 2007, to July 26, 2007. This resulted in a total population of 2,515 internal medicine patient encounter notes. These notes were then processed with GoCode version 1.6 and compared with the E&M level of care indicated by the physician at the time of dictation (Table 1). Note that a difference of 0 means that the provider and GoCode were a perfect match in the E&M coding assignment, while negative differences (–1, –2, –3) indicate possible undercoding and positive differences (1, 2, 3, 4) indicate potential overcoding. Notes falling within +/–1 level of care difference were less concerning than the outliers (those greater than +/–1 level of difference), for which the coding issues need to be understood.

Notes that had a physician-entered E&M code within +/–1 level of agreement with the GoCode-assigned code were eliminated, resulting in a set of 131 discordant notes (disagreement by 2 levels or more) that would require human review and resolution prior to release for billing. In the typical workflow within GoCode, these 131 records (approximately 5.6 percent of the total) would have been routed back to both the physician and other human reviewers for evaluation as exceptions and would not be automatically released for billing.

Four certified coding professionals were selected by AHIMA staff to assist with this coding evaluation. AHIMA staff divided the sample set of 131 discordant notes among these four independent reviewers. Reviewers then evaluated the encounter note without the aid of GoCode findings, using the same audit form, and applied only the 1995 version of the American Medical Association (AMA)/Centers for Medicare and Medicaid Services (CMS) documentation guidelines for establishing the level of service.

Upon completion of the reviewers' evaluation, 64 notes (or 48.9 percent of the sample) demonstrated no agreement between either the physicians, GoCode, or the auditors. Human reviewers noted that three documents had incomplete or insufficient documentation for coding, and these documents were removed from the sample population and excluded from further analysis. In 12 cases the reviewers believed the providers selected the wrong type of encounter, resulting in an invalid CPT code, and these notes were also excluded from further evaluation. Thus, 116 notes remained for further analysis.

Next, a random-number generator was applied to each reviewer's note set, and five notes from each of the four reviewers (n = 20) were randomly selected for a second pass involving a group evaluation and interaction with GoCode processing detail. The group leader was given Web access to the processing detail provided by GoCode for this second pass. Through a virtual meeting environment, the group was able to review findings from GoCode, discuss interpretations of the documentation, and record comments and/or disagreements in interpretation of the record's content. A final E&M coding level was reviewed and recorded.

Analysis

Discussion with HIM professionals regarding interpretive and often subjective interpretations of coding guidelines has generated some consensus that coding results within +/–1 level of agreement should be acceptable.

A statistical test of proportions was conducted to evaluate the change in the proportion of records falling within +/–1 level of agreement prior to interaction with GoCode and with AHIMA staff coding (agreement/disagreements) based on interaction with the GoCode tool.

Based on statistical test results from the reviewers' first pass to the changes in the second pass, the null hypothesis was rejected at the $p < .03$ or 97 percent confidence level on a two-tailed test of proportions, using a *t*-test statistic. Human review initially demonstrated that 52.6 percent fell within the +/–1 level of agreement, as compared to GoCode findings (see Table 2). Upon interacting with GoCode, reviewers achieved 80 percent agreement in the second pass in the +/–1 range of concordance with GoCode (see Table 3). It appears that in this preliminary study, interaction with GoCode technology did improve consensus in the free-text narrative documentation interpretation and coding outcome.

Using the auditing software on the LingoLogix Web portal, a second pass was made comparing the evaluation of the human coders with the software program. This resulted in a change in E&M level in 55 percent of the cases once the components of the documentation were evaluated using the tool. Initially, in only 1 of the 20 cases (5 percent) did GoCode and human reviewers agree (see Figure 1). At the close of this preliminary study, there were only four cases with a variance of more than one level between the human coder and the software results, and the causes for the variance are still under review.

The light blue bars in Figure 1 show the deviation of reviewers' E&M levels at the beginning of the second pass ($n = 20$). Note that the distribution is skewed toward the negative level, indicating a propensity to undercode the documents. After interaction with GoCode, the distribution shows results similar to those in Table 3, with much higher agreement and concurrence of human E&M coding with GoCode. Prior to interaction with GoCode, only 11 out of 20 documents (55 percent) were concurrent (within the +/–1 range) with GoCode versus 80 percent after interaction with GoCode.

Discussion

New technologies involving natural language processing based on semantic and syntactic rules (rather than string word matching or keyword searching) offer new and exciting opportunities to make health information management professionals more efficient, effective, and productive, while simultaneously improving accuracy.

Comments from the AHIMA staff participants include the following. (Note: These comments are those of individuals rather than official remarks by AHIMA.)

- The project was interesting and exciting.
- It was very valuable to look at the software to compare the E&M assignment since the software application referenced the note and displayed it in components required by the guidelines for review.
- Since E&M coding is so subjective, there were areas that the human coder varied with the interpretation of the software application—some were discovered to be facility specific guidelines (medication risk).
- One area that was evident to the coders examining the overall results was that GoCode frequently selected a higher level code for the case compared to the physician or the human coder. In many cases the documentation supported the higher level code.
- We were very impressed with the software and the capabilities and particularly with the fact that SNOMED-CT was utilized.
- The audit tool was very easy to use and helpful in organizing the component parts of the documentation available to evaluate for level of service.
- The educational value as a feedback loop to physicians is impressive.
- The ability to add organization specific criteria to support care levels is useful and in many cases not available to the human coder at the time of code selection.

Other specific observations from the reviewers were as follows:

1. In the History section, GoCode seemed to use isolated words, and out of context. For example, under history of present illnesses, this must be concerning the elements of the present condition, not some past condition, and the computer could not differentiate that.

Occasionally, some false positive findings may occur—especially between history of present illnesses and past medical history, if there is no clear definition between these sections, as the computer cannot tell the recent past from significant past history.

2. In the physical examination section, it is noteworthy that the body areas and the organ systems were randomly mixed, and it appeared that many times the system or area was counted more than once; however, this is not actually the case. For example, if you count an examination of the ears, nose, mouth, throat, then you cannot count head; cardiovascular, then you don't count chest; GI, then you don't count abdomen, and so forth. It also appeared that for these reasons GoCode gave a higher assignment to Examination. Even though GoCode is displaying all the body and organ systems, they are not being double counted. There was one particular record where the total exam was only four lines long, with two lines dedicated to constitutional (vital signs) and GoCode assigned a "detailed" level exam.
3. In the Medical Decision Making section, it also appears to code at a level higher than humans. There may be issues with the software recognizing what was current data and what was a test that appeared in the past medical history section. For example, statements of the patient having a colonoscopy in the past might be utilized in the amount of data to review. It also seemed that cases were inappropriately assigned to high medical decision making due to medication management at the high risk level, as knowledge of these high-risk medications was not taken into account by reviewers.

Conclusions

This is a preliminary study that could be expanded to other specialty areas. Focusing on internal medicine is challenging because of the complexity and variability of the exams and documentation encountered that must be interpreted by the NLP technology against coding guidelines. More standardization of coding guidelines will help ensure that codes are interpreted and applied more uniformly across all areas of healthcare and across institutions.

Rhonda R. Thomas, PhD, MBA, is the president of LingoLogix in Dallas, TX.

Constantin Mitocaru is the lead software developer for LingoLogix in Dallas, TX.

Judith Hanlon is the chief operating officer and vice president of professional services for LingoLogix in Dallas, TX.

Notes

1. Brown, Steven, T. Speroff, E. Fielstein, et al. "eQuality: Electronic Quality Assessment from Narrative Clinical Reports." *Mayo Clinic Proceedings* 81, no. 11 (2006, November): 1472–1481.
2. *Current Procedural Terminology: CPT*. Chicago, IL: American Medical Association, 1999.
3. Lasker, Roz, and Susan Marquis. "The Intensity of Physicians' Work in Patient Visits." *New England Journal of Medicine* 341, no. 5 (1999): 337–341.
4. King, M. S., L. Sharp, and M. S. Lipsky. "Accuracy of CPT Evaluation and Management Coding by Family Physicians," *Journal of the American Board of Family Practice* 14, no. 3 (2001, May–June): 184–192.
5. Brown, Steven, T. Speroff, E. Fielstein, et al. "eQuality: Electronic Quality Assessment from Narrative Clinical Reports."
6. The authors have had meetings and conferences with Dr. Pak and others at Fort Detrick, 2007.
7. Brown, Steven, T. Speroff, E. Fielstein, et al. "eQuality: Electronic Quality Assessment from Narrative Clinical Reports."

Table 1

E&M Coding Level Deviation of Internal Medicine Notes

Difference between Provider and GoCode	Number of Notes
–3	5
–2	110
–1	866
0	1,425
1	93
2	13
3	1
4	2

Table 2

GoCode and Human Review without GoCode Assistance

Deviation	Frequency	Percent	
–3	0	0.0%	
–2	2	1.7%	
–1	4	3.4%	
0	6	5.2%	**52.6%**
1	51	44.0%	
2	52	44.8%	
3	1	0.9%	

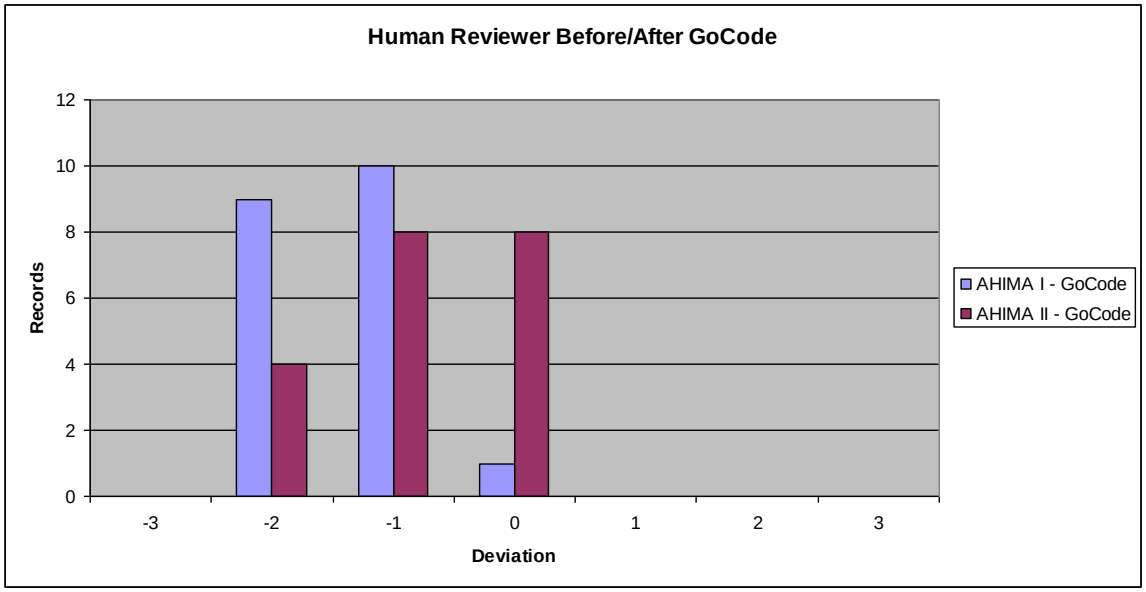
Table 3

GoCode and Human Review with GoCode Assistance (Second Pass)

Deviation	Frequency	Percent
–3	0	0.0%
–2	0	0.0%
–1	0	0.0%
0	8	40.0%
1	8	40.0%
2	4	20.0%
3	0	0.0%

80%

Figure 1


Human Reviewer Before/After GoCode
Records
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Deviation
AHIMA I - GoCode
AHIMA II - GoCode