Computer-assisted Coding at Its Limits: An Analysis of More Complex Coding Scenarios

by Mark Morsch, MS; Carol Stoyla, CLA; Michael Landis; Stacy Rogers; Ronald Sheffer, Jr., MA; Michelle Vernon, RCC; and Michelle A. Jimmink, MBA

Abstract

Healthcare organizations that have adopted computer-assisted coding (CAC) applications have benefited from enhanced coder productivity, elimination of manual processes, flexibility of remote coding, improved coding accuracy and consistency, and an overall more manageable and auditable coding process. However, coding and compliance managers, even early adopters, are often more comfortable using CAC technology based on natural language processing (NLP) for the simple cases. For radiology, this might be screening mammograms and two-view chest x-rays. In emergency medicine, it may be the cases with low medical risk, like otitis media and ankle sprains. The justification is that CAC technology works well for the simple cases, but it just isn't there yet for more complex coding. In this paper, we define three dimensions that characterize complexity beyond categorization by medical specialty or setting. Also, we present an analysis of CAC performance for two more complex coding areas: evaluation and management (E&M) coding for emergency medicine and interventional radiology.

Introduction

In this paper, we examine the factors that make these different coding scenarios complex. This examination will, in part, provide more questions for the skeptical coding manager to ask when considering computer-assisted coding (CAC) technology. However, our focus will be to present three defining characteristics of these complex scenarios. Since these scenarios are challenging for CAC applications, this analysis should help shape the dialogue for further discussions and give insights into future directions of CAC technology development.

In addition, we present test results for two of the complex coding scenarios, evaluation and management (E&M) for emergency medicine and interventional radiology. The results will be presented as a measure of statistical accuracy when comparing a CAC system to coding experts. The statistics will indicate the level of agreement for these two scenarios and highlight some of the challenges in evaluating results for complex coding scenarios.

Background

Usage of CAC applications based on natural language processing (NLP) is commonly identified with medical specialty (radiology, pathology, emergency medicine, etc.) and provider setting (inpatient or outpatient). In its 2004 report, the AHIMA e-HIM Work Group on Computer-Assisted Coding stated, "CAC works best within medical domains that have a limited vocabulary. NLP-based tools in particular work best where there is a limited number of source documents that must be analyzed for code selection and less extensive coding guidelines."[1] It's clear that fewer types of documents and smaller vocabularies

simplify performance requirements for CAC systems. However, what really defines "extensive [or complex] coding guidelines" from a CAC perspective?

For an HIM professional, proficiency with more complex coding guidelines typically involves higher levels of education and experience. However, even with a demonstrated high level of coder proficiency, the intuition is that complex coding scenarios are correlated with greater variability in code assignment. Two coding scenarios where we have observed higher levels of variability are E&M coding for emergency medicine and interventional radiology. Figure 1 reproduces the results from a study by Morris and colleagues (2000) that measured the level of agreement among seven coding experts and one CAC system.[2] The average rate of agreement with the consensus E&M level was 70.3 percent.

With regard to coding for interventional radiology, A-Life Medical conducted a small test with 33 radiology cases that were coded by four different expert coders. Three of the four were external resources, one overseas and two domestic. All were given the cases on paper and instructed to use their normal coding process to assign the ICD-9 and CPT (Current Procedural Terminology) codes for professional services. Table 1 summarizes the results of the test. Full agreement indicates that all four coders assigned the same ICD-9 and CPT codes, including modifiers. For complex interventional cases (which included selective catheterizations, angioplasties, and stent placements), zero out of nineteen cases had full agreement. The differences among the nineteen cases were quite variable, a mix of ICD-9 code, CPT code, and modifier differences with no single outlier identifiable. While this study was small, we believe that it, along with the previous work by Morris and colleagues,[3] substantiates the intuition that more complex coding scenarios lead to greater variability in code assignment.

Analysis

Thus, if we accept the premise that correlates complexity with variability, then we can identify the characteristics of the coding guidelines that make certain scenarios more complex. Based on our experience developing a CAC system, we have identified three general characteristics that are associated with complex coding guidelines:

1. **Language generalization**—The language used in the coding guidelines cannot be easily generalized into actual usage in clinician documentation.
2. **Multi-stage reasoning**—One or more levels of reasoning, beyond one-to-one matching of a definition from the guidelines to a statement in the physician note, are required to arrive at the correct coding.
3. **Domain knowledge**—Significant knowledge of human anatomy, treatment protocols, or disease processes is necessary to fully capture meanings of certain codes.

We illustrate each of these with an example. Language generalization is the process of understanding how to map the language used in real-world medical documents to the definitions in the coding guidelines. This is similar to the process a newly trained HIM professional goes through in learning coding for billing. A good example of this is CPT surgical procedures. Typical CPT code definitions are one or two sentences in length, but the surgeon's notes may be a page or more with detailed descriptions of each step of a procedure, including the site preparation, anesthesia dosage and delivery, instruments used, surgical techniques, and condition of the patient throughout the procedure. Simple language processing techniques, such as synonym substitution and keyword matching, are inadequate to handle this level of complexity.

Multi-stage reasoning is common in the ICD-9 and CPT guidelines. This reasoning can involve quantifying aspects of a condition or procedure, or applying specific logic based upon a combination of individual facts. A simple example, which actually can be challenging for a CAC system to handle in general, is ICD-9 coding of rib fractures. The fifth digit of the ICD-9 code for rib fracture specifies the number of ribs fractured. The CAC system must sum up the ribs listed, in whatever form they may be expressed. Other examples of this type of logic in ICD-9 include coding the loss of consciousness associated with skull fractures and expressing the body area percentage affected by burns. E&M coding is

also a good example of multi-stage reasoning with its multiple components and various data elements making up each component.

Extensive domain knowledge is required for certain coding scenarios. This exceeds the language generalization problem by requiring a coder to have clinical knowledge beyond standard coding references. A good example is CPT coding of catheterization procedures. These procedures are classified based upon the route of the catheter through the blood vessels and the vessel branches crossed during the route. Knowledge of the vascular anatomy is essential, with specific understanding of vessel interconnections and potential anatomy variants. Other scenarios that require extensive domain knowledge include obstetrics and complications from surgery.

Results

To assess the performance of a CAC system in complex coding scenarios, we analyzed data produced by the LifeCode engine,[4] an NLP-based CAC technology. The first scenario is E&M coding for emergency medicine. Our analysis measured the rate of agreement for E&M codes between the CAC system and users for a three-week period during the month of July 2007, a total of 31,399 cases. The users are 33 production coders who performed coding in their normal workflow, using the CAC system with visibility of the CAC system results. Figure 2 shows the agreement rates with the CAC system for each of the 33 coders. The overall average rate of agreement was 79.8 percent. Note that there is significant variability in the number of cases handled by each coder, from 1 case for coders 17, 20, 31, and 33 to 4,841 cases for coder 25.

The second scenario is coding for interventional radiology. The data analyzed is from 12 facilities over a five-month period in early 2007, a total of 5,960 interventional radiology cases with an average of 2.66 procedure codes per case. The coding process included two levels of review. All cases were processed through a CAC system and reviewed by a professional coder. A second level of review was performed on all cases by an auditor, who reviewed the results of the professional coder and identified any discrepancies. The statistics presented here compare the results of the CAC system to the auditor. Rates of recall and precision were calculated for the CPT codes. Recall is the percentage of CPT codes assigned by the auditor that matched the CAC system. Precision is the percentage of CPT codes assigned by the CAC system that matched the auditor. Higher rates of recall mean lower false negatives. Higher rates of precision mean lower false positives.

Figure 3 and figure 4 show the recall and precision statistics for the eight larger facilities. The four smaller facilities are not displayed because, in total, they represented less than 0.2 percent of the total cases. The overall recall was 56.1 percent with a high of 76.5 percent and a low of 20.0 percent, and the overall precision was 80.8 percent with a high of 93.1 percent and a low of 53.4 percent.

Discussion

We have presented evidence to support the intuition that more complex coding scenarios lead to greater variability in coding results. To better define complex coding related to CAC systems, we have defined three characteristics of coding guidelines associated with complexity: language generalization, multi-stage reasoning, and domain knowledge. Lastly, we have analyzed the performance of a CAC system for two types of complex coding. The feedback from users is consistent with other studies[5] that have identified improved productivity and consistency. However, we believe a key point for all parties involved in CAC standards to appreciate are the niches of complexity throughout the coding guidelines. These are, in some sense, guidelines within guidelines, each with its own set of rules and unique features. Standards must have the flexibility to be applied both across coding systems and within these niches.

Mark Morsch, MS, is the vice president of NLP and software engineering at A-Life Medical, Inc., in San Diego, CA.

Carol Stoyla, CLA, is the director of compliance, coding, and software QA at A-Life Medical, Inc., in San Diego, CA.

Michael Landis is an NLP software engineer at A-Life Medical, Inc., in San Diego, CA.

Stacy Rogers is a lead research linguist at A-Life Medical, Inc., in San Diego, CA.

Ronald Sheffer, Jr., MA, is the manager of NLP development at A-Life Medical, Inc., in San Diego, CA.

Michelle Vernon, RCC, is a lead radiology analyst at A-Life Medical, Inc., in San Diego, CA.

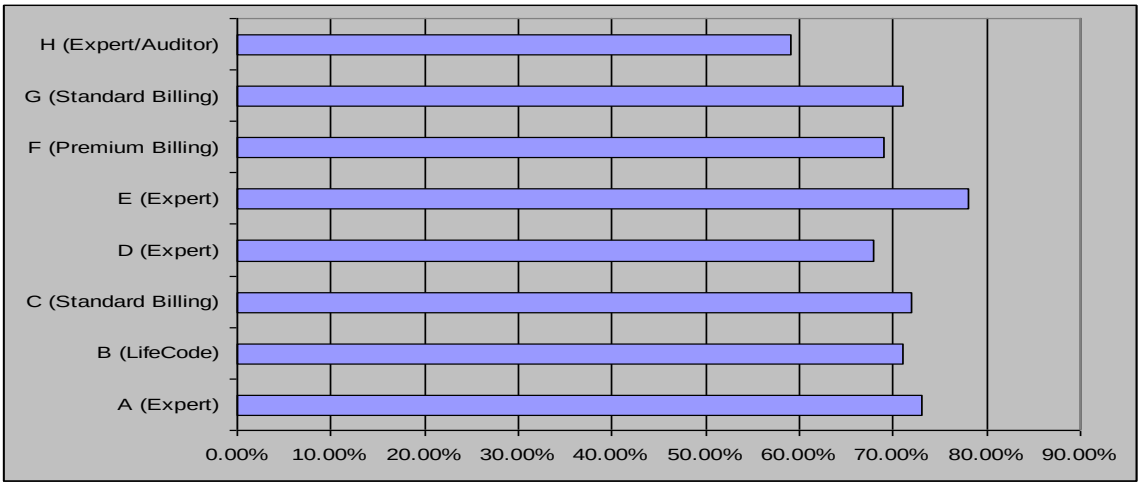
Michelle A. Jimmink is an advanced research linguist for A-Life Hospital in San Diego, CA.

Notes

1. AHIMA e-HIM Work Group on Computer-Assisted Coding. "Delving into Computer-assisted Coding" (AHIMA Practice Brief). *Journal of AHIMA* 75, no. 10 (2004): 48A–H.
2. Morris, William, et al. "Assessing the Accuracy of an Automated Coding System in Emergency Medicine." *Proceedings of the AMIA Annual Symposium* (2000): 595–599. Available at http://www.alifemedical.com/documents/LifeCodeEMPerformanceAMIA2000.pdf.
3. Ibid.
4. Heinze, Daniel, et al. "LifeCode: A Deployed Application for Automated Medical Coding." *AI Magazine* 22, no. 2 (2001): 76–88. Available at http://www.alifemedical.com/documents/LifeCodeAIMagazine.pdf.
5. Resnik, Phillip, et al. "Using Intrinsic and Extrinsic Metrics to Evaluate Accuracy and Facilitation in Computer Assisted Coding." AHIMA/FORE Computer-Assisted Coding Software Standards Workshop, Arlington, VA, September 6–8, 2006.

Figure 1

E&M Agreement Rates with Consensus



Source: Morris, William, et al. "Assessing the Accuracy of an Automated Coding System in Emergency Medicine." *Proceedings of the AMIA Annual Symposium* (2000): 595–599. Available at http://www.alifemedical.com/documents/LifeCodeEMPerformanceAMIA2000.pdf.

Figure 2

Observed E&M Agreement Rates

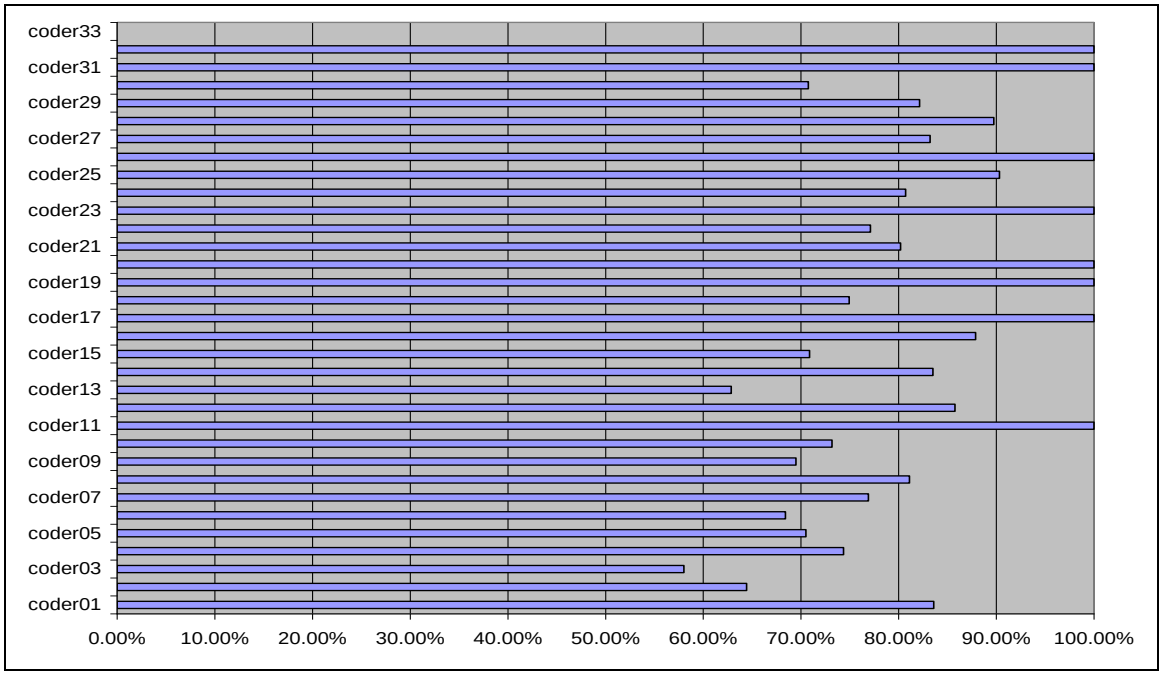


Figure 3

Interventional Coding Recall

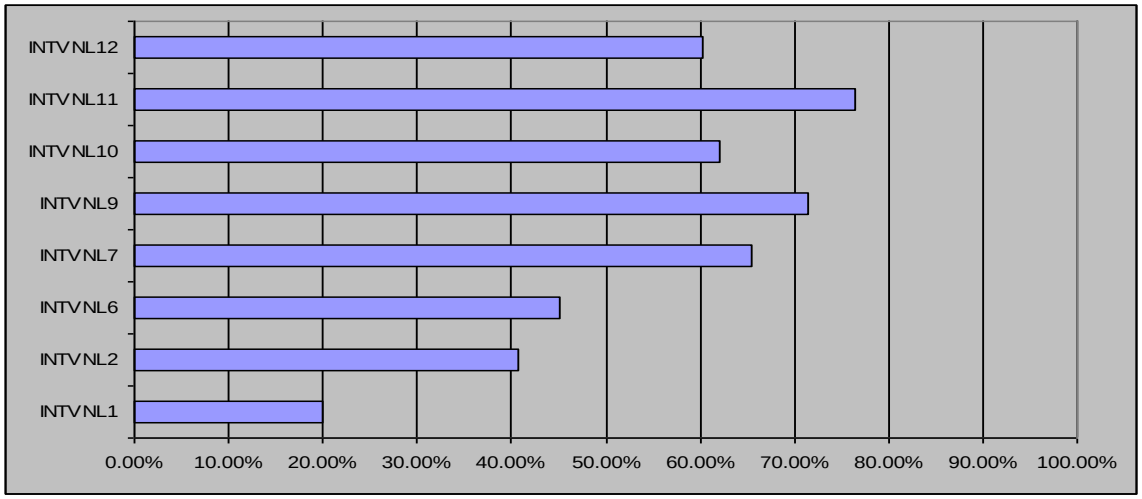


Figure 4

Interventional Coding Precision

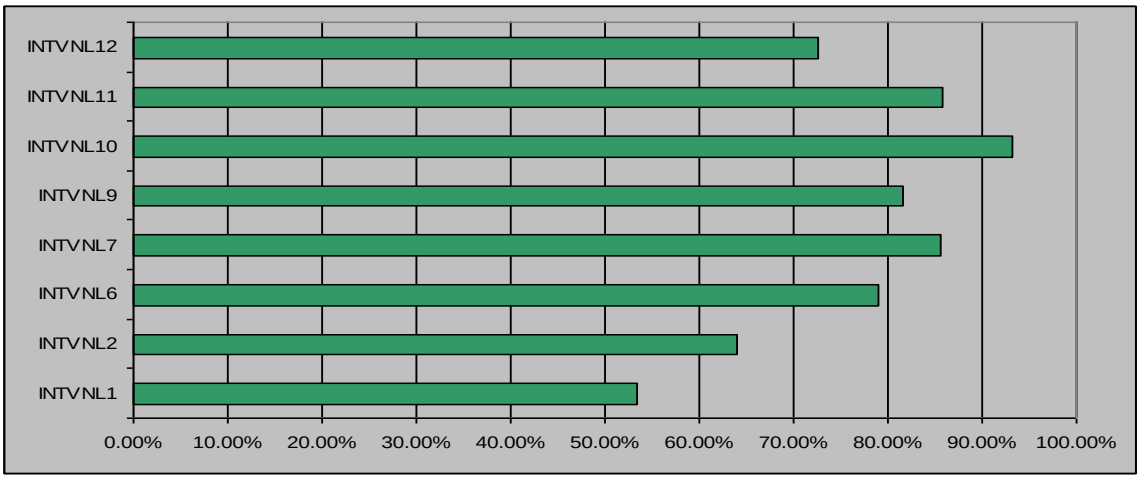


Table 1

Results of Four-Way Coding Test

Case Type	Number of Cases	Full Agreement
Diagnostic radiology	10	9/10
Simple interventional	4	3/4
Complex interventional	19	0/19