Defining the Standards for Automated E&M Coding through Coding Consistency Methodology

by James Flanagan, MD, PhD, FACP; and Mariana Casella dos Santos, MD

Abstract

A number of studies have shown that evaluation and management (E&M) coding even by experienced professional coders exhibits poor reliability as shown by a number of studies. Centers for Medicare and Medicaid Services (CMS) guidelines for E&M coding allow a large range of interpretation, requiring a number of ad hoc decisions (by carriers, provider institutions, auditors, and individual coders) in order to implement the guidelines.

This report investigates whether ad hoc standards are a major source of poor reliability. Our approach to this question develops a method to create a meaningful standard for an automated E&M coding tool. To investigate this question, we developed a process for revealing E&M coding variation, for analyzing the root cause of the variation, and for attempting to remove the variation. The process utilized two teams of coders: professional coders working for a large provider organization and physicians hired to work on an automated coding product. Clinical documents were coded for E&M by all individuals unaware of the others' results or automated results. Variances among individual coders were analyzed to determine issues that were consistent sources of variation.

Based on an iterative process, we found 60 separate issues (21 history, 17 exam, and 22 complexity) to which coders attributed variation among their decisions. Each issue was resolved to a Coding Consistency Standard through reference to documentation from CMS, carrier publications, or knowledge of auditor practices. At the outset of this process, both groups of coders had approximately a 66 percent agreement with the majority code of their respective groups. After several iterations and a stable set of Coding Consistency Standards, the average agreement with the majority code rose to 86–90 percent. Among coders unaware of the Coding Consistency Standards, the agreement remained at 66 percent.

We conclude that it is possible to achieve fairly high reliability for E&M coding if a consistent and detailed set of guidelines is applied. However, even after enormous effort there remains a significant amount of variation. One source of this variation falls under the heading of "inference." By this we refer to all cases in which a decision about a CMS guideline depends on making an inference based on what is actually stated in the document and using medical knowledge. Whether or not to make the inference depends both on the medical knowledge of the coder and on whether or not the individual coder believes an inference is justified. Examples of such inferences typically affected HPI duration and whether or not a problem addressed is considered new. Eliminating the need for such inferences is the current focus of our effort in developing Coding Consistency Standards for E&M coding.

Introduction

To create and deploy automated coding tools, it is essential to compare the automated tool to a standard for a number of reasons: to gain initial acceptance, to provide ongoing monitoring for regression, and to pass the ultimate test of a coding audit. For evaluation and management (E&M) coding tools, this is quite difficult because there is no standard. The Centers for Medicare and Medicaid Services (CMS) guidelines allow such a large range of interpretation that a number of ad hoc decisions are required to implement the guidelines. In some cases these decisions are made by the carriers, in others by the provider institutions, and in still others by auditors, but a large number of decisions are made by individual professional coders. We asked whether such ad hoc standards could be the main reason for the well-known poor reliability of E&M coding that has been shown by a number of studies.[1–4]

Discovering Issues

To investigate this question, we developed a process for revealing variation among individual professional E&M coders and for analyzing the root causes of the variation. We worked with two separate groups of four coders per group. One group was drawn from the pool of professional coders employed by a large healthcare provider. The individuals in this set were trained and accredited for this role and performed E&M coding tasks on a daily basis according to their understandings of national standards, regional directives, and local policy as well as their own personal conventions. In the discovery phase of the project, these four coders were freed from their normal duties for five consecutive days during each of several iterations in order to code sample documents. Initially, they were instructed to code documents according to their normal standards but to do so without conferring with each other. All reached their codes independently while being monitored. Their individual results, including the codes and subscores, were collected and recorded. Following this, each document was analyzed by the monitor and the group of coders to discuss every factor each coder identified as influencing the E&M code. Each case of discrepant decisions regarding the contributing factors was reviewed and discussed. After the discussion, the group was asked to reach a consensus on the code and subscores for each document.

The second set of coders were all physician-trained informatics specialists whose coding experience was informal, none of whom were accredited, and whose regular duties did not involve scoring clinical documents for E&M codes. Their regular duties did involve various aspects of the development of an automated E&M coder, and all were familiar with the general requirements of computer algorithms for precise definitions. In the discovery phase of the project, these four physicians were freed from their normal duties for five consecutive days during each of several iterations to code sample documents. They reviewed published CMS guidelines but had no supplemental training of any kind. They were instructed to code documents without conferring with each other. Their individual results, including the codes and subscores, were collected and recorded. Each case of discrepant decisions was reviewed and discussed for the contributing factors. After the discussion, the group was asked to reach a consensus on the code and subscores for each document.

This process was repeated with each group for three successive sets of 100 documents by each individual within two independent sets of three or four coders. There was no intent to compare the accuracy or reliability between the two sets of coders. The entire purpose of this exercise was to elicit issues on which knowledgeable coders disagreed. During the discovery all their results were analyzed identically. All variances among coders for individual documents were analyzed to determine what specific factors led to the discrepancy.

Consensus: Coding Consistency Standards

As noted above, in the process we asked coders to identify the factors contributing to the discrepancies and for consensus on the discrepant codes.

Based on an iterative process, we found that 60 separate issues (21 history, 17 exam, and 22 complexity) accounted for a large part of the variation among coders' decisions. These 60 issues reflect

variation in frequent practices among professional coders. Many of the standards were resolved by reference to documentation from carriers and knowledge of auditor practices. We referred to the consensus resolutions as Coding Consistency Standards. We are aware from discussions with coders from other organizations that there would be significant variation among the groups as to what their consensus decisions would have been.

We do not purport to have created a standard that will satisfy all coders, provider organizations, or auditors. Each of the standards raised an issue concerning which one of several possible decisions was reached. Other decisions could very well be considered correct. As such, in presenting the standard, we present only the issue addressed and not the standard chosen.

Issues Considered in the Coding Consistency Standards

See Table 1 for general factors affecting the coding issues.

Results

After agreeing on the Coding Consistency Standards, we asked if the standards could reduce the variation among coder practices for subsequent sets of documents. At the outset, both groups of coders had approximately the same degree of reliability. There was an average agreement of approximately 66 percent with the majority code within each group. After developing a stable set of Coding Consistency Standards, we reevaluated. In both groups familiar with the Coding Consistency Standards, the individual agreement with the majority code was 86 and 90 percent for the two groups, respectively. Simultaneously, the result among a new set of professional coders unfamiliar with the Coding Consistency Standards was again a 66 percent individual agreement with the majority code. This third group of coders was then able to achieve 86 percent agreement with the majority code after becoming familiar with the Coding Consistency Standards.

Discussion

The results suggest it is possible to improve the reliability of human coding through rigorous definition of and adherence to Coding Consistency Standards. However, even after enormous effort, there remains a significant amount of variation.

One source of variation is the decision about whether a statement in the document is sufficient for meeting the stated guideline. For instance, at one extreme, the merest mention of a problem can be taken as evidence of a problem addressed. At the other extreme, one could require that each problem addressed be summarized in a special section and be associated with some plan of action. As another example, one document may simply report an entire system as negative where another document reports details about the system. This source of variation is reduced by defining a very specific standard of the kind we created in our Coding Consistency Standards. Creating such a standard is an effective method for removing this source of variation.

Another source of E&M coding variation falls under the heading of “inference.” By this we refer to all cases in which a decision depends on making an inference based on what is actually stated in the document and on medical knowledge. Whether or not such an inference is made depends both on the medical knowledge of the coder and on whether or not the individual coder believes there is more than one possible interpretation. Some coders have more medical knowledge than others. Some coders are willing to accept a likely inference, while others make an inference only if it is certain. Examples of such inferences include deciding whether there is sufficient information to infer the HPI duration, determining whether a problem addressed is new, or determining whether a problem is worsening.

An inference often depends on knowing that there is a connection among various statements based on knowing that specific symptoms or laboratory results could be connected to diagnoses named in other sentences. There is no question that such inferences are valid and consistent with CMS guidelines if they can be defended. The point is that there is often a leap that can be taken from what is actually stated to the information that bears on some extremely vague CMS guidelines. Not everyone makes these leaps with

the same ability or with the same certainty. Inferences as a source of variation could perhaps be controlled using Coding Consistency Standards that define very precise rules for allowable inferences. Unfortunately, such rules can be complicated and numerous, making it difficult for humans to remember them. As a result, inferences remain a significant source of human variation in coding behavior.

Summary

Defining Coding Consistency Standards provides several advantages to a project using an automated E&M coding tool. The same standard can be used to configure the E&M coding tool and to train those responsible for evaluating the tool. Using Coding Consistency Standards, human coders can create a gold-standard set of coded documents for evaluating the automated tool. Our experience is that the automated tool influences the human coders to more reliably adhere to the Coding Consistency Standards.

James Flanagan, MD, PhD, FACP, is the chief medical officer for Language and Computing in Bethesda, MD, as well as an internal medicine hospitalist.

Mariana Casella dos Santos, MD, is the chief ontologist for Language and Computing in Bethesda, MD.

Notes

1. M. S. King, M. S. Lipsky, and L. Sharp. *Archives of Internal Medicine* 162 (2002): 316–320.

2. W. C. Morris et al. "Assessing the Accuracy of an Automated Coding System in Emergency Medicine." *Proceedings of the AMIA Annual Symposium* (2000): 595–599.

3. Morsch, Mark L., David S. Byrd, and Daniel T. Heinze. "Factors in Deploying Automated Tools for Clinical Abstraction and Coding." Unpublished work. 2007.

4. Stoner, Jean, Michael Nossal, Philip Resnik, Andrew Kapit, and Richard Toren. "Assessing Coder Change Rates as an Evaluation Metric." Unpublished work. 2007.

Table 1

General Factors Affecting the Coding Issues

Chief Complaint, Family History, Past Medical History, Social History, Problems Addressed	Conceptual content required Section location, explicit definitions (New, Established, Worsening, Workup)
HPI Elements	Explicitly linked to a presenting problem Section location, explicit definitions
Double Dipping	Define permission for each type
Physical Exam	Define the bullets credited for common abbreviations