Use of the Delphi Method to Generate a Gold Standard for the Medical Coding Industry

by M. Elliott Familant, PhD; Stuart Covit; and Andrew B. Covit, MD

Abstract

A "gold standard" for medical coding is now an acknowledged need. However, the well-documented phenomenon of coder variability makes the generation of such a standard a challenge. A method is needed to efficiently generate a consensus view of how to code any given medical record. The Delphi Method, a consensus-generating methodology, is one such method.

In this study, we demonstrate how a modified version of the Delphi Method can be used to create a gold standard for medical record coding. We demonstrate that, compared to more traditional methods, the Delphi Method reduces undercoding and is less susceptible to problems due to limitations in human cognitive ability.

Finally, we describe an implementation of the Delphi Method as an online gaming environment. We discuss why a gaming environment is an ideal method for generating consensus cases and describe the benefits of this gaming environment to the greater coding community.

Introduction

Multiple studies have shown that medical coders will frequently differ in the number and type of codes they apply to medical charts. This is true for ICD-9-CM and ICD-10 coding[1–3] and CPT (Current Procedural Terminology) codes, including E&M coding.[4, 5] Coder variability is a significant challenge in developing any standard of "correctness" for medical coding. This is especially true for a gold standard that would be used to evaluate computer-assisted coding (CAC) systems. A gold standard, like any kind of measurement system, must be *valid*, meaning that the measurements derived from the system must be an actual reflection of the thing being measured (in this case the correctness of the coding system), and *reliable*, meaning that the measurements derived from the gold standard must be consistent, with the same measurements obtained each time the system is used under the same conditions.[6] Coder variability undermines both of these standards.

A typical method for creating a gold standard is to identify experts in the task under study (in this case medical coding experts), have them perform some reference task independently (code a set of medical records), and then use some resolution protocol to resolve differences in task output. For example, Resnik and colleagues (2006) had two experts code 720 radiology notes and then had a third expert code those notes when the first two experts disagreed. The gold standard for any radiology note in this study became the codes applied to notes that were agreed on by two of the three experts. Resnik and colleagues found that for 94 percent of the notes, two of the three coders agreed on all the CPT codes that should be applied.[7]

At first, the amount of agreement obtained in this study seems remarkable, especially in light of previous studies showing significant coder variability. However, a question can be raised about this kind of method for creating a gold standard. Radiology notes happen to be a particularly simple form of medical documentation.[8] Their simple structure and relatively limited vocabulary make them amenable to automation. This would also tend to reduce the amount of variance among medical coders working with these documents since both the number and variety of codes that could be applied to each note is small. Under these conditions, it is a reasonable possibility that the amount of agreement obtained was simply an artifact of the kinds of documents coded and the low threshold used to determine if coders agreed that a code should apply. (In fact, in the simple case, where only one code should be applied to a note and there are only two possible codes that could logically be applied, the chances that two out of three coders will agree on which of the two codes actually apply is 100 percent.) Would the codes generated as the gold standard in that study be reliably generated again if different experts were used? Can this technique be used for other, more complicated kinds of medical documentation?

Also, developing a gold standard by having expert coders independently generate codes is a method that is vulnerable to all the well-documented limitations of human information processing,[9, 10] limitations that contribute to the observed variability between medical coders. The alternative is to develop a method that incorporates techniques that specifically compensate for these limitations. One possibility, a modified version of the Delphi Method, is discussed in the next section.

The Delphi Method

Linstone and Turoff define the characteristics of the Delphi Method as follows:

- Receiving input from a variety of experts about a topic of interest, typically anonymously
- Obtaining this input in a structured way (e.g., a questionnaire, an opinion on a defined problem, a set of rating scales)
- Evaluating the input by using a set of criteria, and filtering and summarizing it if necessary
- Presenting this evaluation to the experts again and giving them an opportunity to comment on it and change their input based on the evaluation
- Evaluating this second round of input and presenting this second evaluation to the experts
- Iteratively repeating the process until the opinions of the experts are stable and, in some instances, have converged on a consensus opinion[11]

Since its development in the 1950s at the RAND Corporation, the Delphi Method has been used for a wide range of applications including project estimation, risk analysis, and technology projections.[12]

However, to date it has not been used for the more structured task of medical coding. Yet the fact that medical coding is a complicated task without a generally agreed-upon "solution" makes it a candidate for this approach.

A straight application of the Delphi Method could be used to create a gold standard, but there are potential problems with this. Traditionally, comments made by other experts have been used as part of the information received by participants. This feedback is used to reach a consensus by persuading some participants to change their positions on the topic in question. But presenting other participants' comments is a cumbersome process, and the potential variability of these comments, coupled with the fact that persuasion through written discourse is a complicated and difficult-to-control interaction, could have a deleterious effect on the reliability of the Delphi Method.

The Delphi Method of Medical Coding

To address the concerns raised above, we modified the Delphi Method to make it more streamlined and less susceptible to bias. The Delphi Method of Medical Coding (DMMC), a patent-pending process developed by Artificial Medical Intelligence, consists of three steps:

1. Multiple coders are presented with identical sets of medical records and are asked to code them as they would in a normal clinical setting. This first stage is used to identify the universe of possible codes that could reasonably be applied to a chart. For each record, the codes are consolidated across coders into a single set so that every code that was generated by at least one coder is represented once in the set. One consolidated set of codes is created for each medical record reviewed.

2. The consolidated sets of codes gathered in step 1 are then presented to a second set of coders along with the associated medical records that were used in step 1. The second group of coders is asked to make a simple yes/no decision about each code. Specifically, should this code be applied to the associated record or not?

3. For each code, the percentage of coders who decided that the code should apply to the associated record is calculated. Then comparing that percentage to some a priori criterion (either a fixed percentage or some derived statistic, for example, one derived from a binomial distribution), one of three decisions is made about the code:

    a) The code should be applied to the chart (and hence should be part of the gold standard) because the percentage of coders who voted "yes" (or the statistic derived from this percentage) exceeds some criterion.

    b) The code should not be applied to the chart because the percentage of coders who voted "no" (or the associated statistic) exceeds some criterion.

    c) The code is indeterminate, meaning that a consensus either for or against applying the code to a chart cannot be derived.

A Pilot Study That Used the DMMC

Method

To explore the effectiveness of the DMMC, we ran a pilot study. During phase 1, six coders, all with coding certification and hospital coding experience, were presented with five inpatient records. The records were e-mailed to the coders with the instruction to code the records as they normally would using whatever coding resources (manuals, encoders, etc.) they usually employ. Coders entered the codes into a spreadsheet and mailed them back to the experimenter. Consolidated sets were obtained for each record. During step 2, the consolidated sets and associated medical records were presented to seven additional coders. The order in which the medical records were presented was randomized with the constraint that no coder received the medical records in the same sequence. Coders were again instructed to use whatever resources they would normally use but were requested not to consult with other coders. Coders were provided with a simple form with the codes from each consolidated set pre-entered. Next to each code were two boxes, one for "yes" and one for "no." They were instructed to check the "yes" box if they felt the code should apply and the "no" box if they felt the code should not apply. Coders were also provided the opportunity to write additional codes that they felt should be applied to the record but were not listed in the consolidated set. Only one coder chose to do this, adding three additional codes, but this was done for only one record. Because of this, these additional codes were excluded from the final analysis.

Results

How does the pattern of agreement between coders change when making simple yes/no judgments compared with the more traditional methods where coders independently generate codes? Figure 1 shows the level of agreement between coders during step 1 of the DMMC. Coders at this step of the method perform a task very similar to the task used to create a gold standard in the study by Resnik and colleagues.[13]

We divided the codes into the following three categories: codes in which two-thirds or more of the coders generated the code for the record, codes in which one-third or fewer of the coders generated the code for the record, and codes in which between one-third and two-thirds of the coders generated the code for the record. If consensus is assumed to occur when two-thirds of the coders agree that a code should or should not be applied to a record, then the level of agreement in step 1 seems to be quite high: consensus is obtained for 91 percent of the codes, a percentage similar to the one obtained by Resnik and colleagues.[14]

Figure 2 shows the pattern of agreement among coders during step 2 of the DMMC.

Codes were again divided into three categories. However, this time agreement was determined by the percentage of coders voting "yes." It might seem that the level of agreement actually declines when voting on codes as opposed to generating codes. We assume that codes for which the level of agreement is between one-third and two-thirds of the coders have an ambiguous status in that it is not clear whether they should or should not be applied to the associated chart. The percentage of codes that fall into this middle category increases by almost 67 percent, from 9 to 15 percent, between step 1 and step 2. But to conclude that voting on codes amplifies disagreements between coders is to miss the broader point about what is really taking place. This is shown in Figure 3.

The figure shows that the number of coders agreeing that any one code should be applied to a medical record increases between step 1 and step 2. Also, a greater number of codes were applied by a larger number of coders in step 2. The average percentage of coders that agree that any given code should be applied to a record increased between step 1 and step 2 from 49 percent to 61 percent ($p < .05$). At the same time, the number of outliers drops dramatically once a list of "potential" codes is provided. It would appear that one of the reasons coders seem to agree more often in step 1 (and in traditional methods for producing gold standards) is that they simply did not think of certain codes that could have been applied. An implication of this is that traditional methods for producing gold standards are susceptible to undercoding problems. Presenting coders with a list of codes that they need to vote on as opposed to requiring coders to generate codes reduces this source of error.

Discussion

Although the current study is only a pilot, it suggests that the method of creating a gold standard by having coders independently generate codes and then using a resolution protocol to resolve differences has significant limitations. At a minimum, such a method is likely to result in gold standards with significant amounts of undercoding.

The DMMC represents an alternative. Agreement is dependent on independent judgments on the applicability of codes rather than independent generation of the codes themselves. Because the former task is easier, it eliminates a significant source of error and, as the current study suggests, results in more completely coded gold standards.

Traditionally, the number of experts used to create gold standards has been small. However, using small sample sizes increases the likelihood of variability, reducing the reliability of the standard. This suggests that significantly increasing the numbers of coders who participate in the creation of a coding standard is desirable. One way to do this is to implement the DMMC as an online game like the one developed by Artificial Medical Intelligence, called The Coding Game, a registered trademark of Artificial Medical Intelligence.[15]

In this game, coders perform both the task of entering codes and defining the universe of possible applicable codes (step 1) and the task of voting on the applicability of codes (step 2), just as is done in the DMCC, but as a competition. The coders receive points based on the correspondence between their responses and consensus of coders participating in the game.

Online games that produce useful output as a byproduct are part of the emerging field of human-based computation.[16] Online games have been shown to be effective means of accomplishing other tasks including labeling images[17] and collecting common-sense knowledge.[18] As with other online games, coders are motivated to participate for several reasons:

- The game is engaging and fun to play.
- It allows coders to demonstrate their skills and receive a rating.
- Coders can win prizes and cash based on their performance.

The Coding Game will produce a national gold standard for medical coding based on the responses of potentially thousands of coders. It will simultaneously be able to produce a quantitative score that will measure a coder's ability to code, based on the gold standard. It can be adapted to audit coding at individual hospitals and could even be used to do new coding of medical records. The game is also adaptable to any coding system, including ICD-9, ICD-10, and CPT. Because it implements the DMMC, it inherits all the advantages of this methodology that have been demonstrated in this study. We believe that an online game like The Coding Game is the most effective, robust, and economical way to create a coding gold standard that can be used to evaluate CAC solutions, coders, and hospital coding.

M. Elliott Familant, PhD, is chief technical officer of Artificial Medical Intelligence in Eatontown, NJ.

Stuart Covit is executive vice president of marketing and administration of Artificial Medical Intelligence in Eatontown, NJ.

Andrew B. Covit, MD, is chief executive officer of Artificial Medical Intelligence in Eatontown, NJ.

Notes

1. Office of Inspector General, Office of Healthcare Inspections. *Department of Veterans Affairs*. March 19, 1993. Available from the Department of Veterans Affairs Web site at http://www.va.gov/oig/54/reports/no37.htm (retrieved August 13, 2007).

2. Morris, W. C., D. T. Heinze, H. R. Warner, Jr., A. Primack, A. E. Morsch, and R. E. Sheffer. "Assessing the Accuracy of an Automated Coding System in Emergency Medicine." *Proceedings of the 2000 AMIA Annual Symposium* (pp. 595–599). Philadelphia: Hanley & Belfus, 2000.

3. Nilsson, G., H. Petersson, H. Ahlfeldt, and L. E. Strender. "Evaluation of Three Swedish ICD-10 Primary Care Versions: Reliability and Ease of Use in Diagnostic Coding." *Methods of Information in Medicine* 39 (2000): 325–331.

4. Morris, W. C., D. T. Heinze, H. R. Warner, Jr., A. Primack, A. E. Morsch, and R. E. Sheffer. "Assessing the Accuracy of an Automated Coding System in Emergency Medicine."

5. King, M. S., M. S. Lipsky, and L. Sharp. "Expert Agreement in Current Procedural Terminology Evaluation and Management Coding." *Archives of Internal Medicine* 162, no. 3 (2002): 316–320.

6. Cook, T. D., and D. T. Campbell. *Quasi-Experimentation: Design and Analysis for Field Settings.* Chicago, IL: Rand McNally, 1979.

7. Resnik, P., M. Niv, M. Nossal, G. Schnitzer, J. Stoner, A. Kapit, et al. "Using Intrinsic and Extrinsic Metrics to Evaluate Accuracy and Facilitation in Computer Assisted Coding." *Computer-Assisted Coding Software Standards Workshop.* Arlington, VA: AHIMA, 2006.

8. Lai, J., and J. Vergo. "MedSpeak: Report Creation with Continuous speech Recognition." *Proceedings of CHI '97* (pp. 431–438). Atlanta, GA: ACM Press, 1997.

9. Norman, D. A. *The Psychology of Everyday Things.* New York: Harper & Row, 1988.

10. Reason, J. T. *Human Error.* New York: Cambridge University Press, 1990.

11. Linstone, H. A., and M. Turoff. *The Delphi Method: Techniques and Applications*. 2002. Available at the *Delphi Method: Techniques and Applications* Web site at http://is.njit.edu/pubs/delphibook/index.html (retrieved August 13, 2007).

12. Ibid.

13. Resnik, P., M. Niv, M. Nossal, G. Schnitzer, J. Stoner, A. Kapit, et al. "Using Intrinsic and Extrinsic Metrics to Evaluate Accuracy and Facilitation in Computer Assisted Coding."

14. Ibid.

15. Familant, M. E. United States Patent No. 44617. 2007.

16. Kosorukoff, A. "Human-based Genetic Algorithm." *IEEE Transactions on Systems, Man, and Cybernetics* (2001): 3464–3469.

17. von Ahn, L., and L. Dabbish. "Labeling Images with a Computer Game." *ACM CHI* (2004): 319–326.

18. von Ahn, L., M. Kedia, and M. Blum. (2006). "Verbosity: A Game for Collecting Common-Sense Knowledge." *ACM Conference on Human Factors in Computing Systems, CHI Notes* (2006): 75-78.

References

Kikano, G. E., M. A. Goodwin, and K. C. Stange. "Evaluation and Management Services: A Comparison of Medical Record Documentation with Actual Billing in a Community Family Practice." *Archives of Family Medicine* 9, no. 1 (2000): 68–71.

Morsch, M., D. Heinze, and D. Byrd. "Factors in Deploying Automated Tools for Clincal Abstraction and Coding." *IT in Health Care: Sociotechnical Approaches. Second International Conference.* Portland, OR: IT in Health Care, 2004.

Norman, D. A. "Categorization of Action Slips." *Psychological Review* 88 (1981): 1–15.

Zuber, T. J., C. E. Rhody, T. A. Muday, et al. "Variability in Code Selection Using the 1995 and 1998 HCFA Documentation Guidelines for Office Services." *Journal of Family Practice* 49, no. 7 (2000): 642–645.

Figure 1

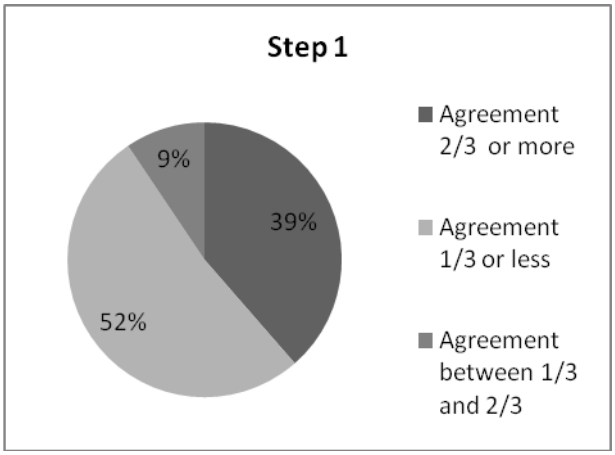

Step 1
9%
39%
52%
Agreement 2/3 or more
Agreement 1/3 or less
Agreement between 1/3 and 2/3

Figure 2

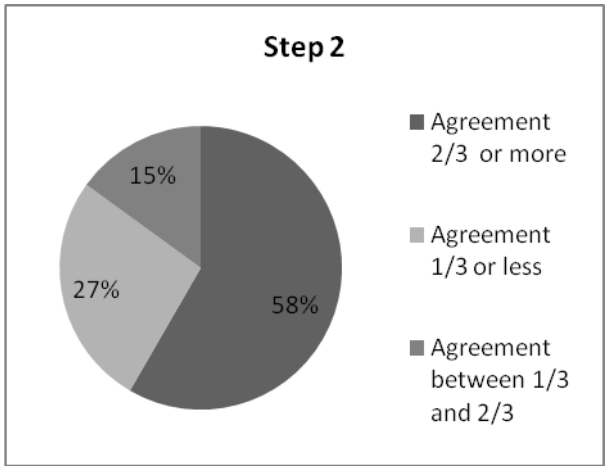
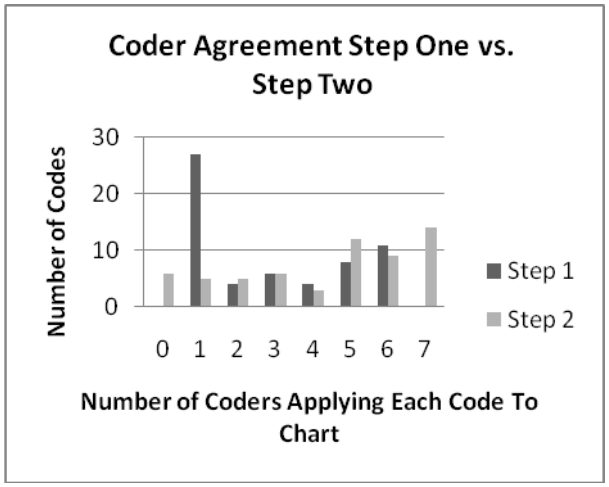

Step 2
15%
27%
58%
Agreement 2/3 or more
Agreement 1/3 or less
Agreement between 1/3 and 2/3

Figure 3


Coder Agreement Step One vs. Step Two
Number of Codes
30
20
10
0
0 1 2 3 4 5 6 7
Step 1
Step 2
Number of Coders Applying Each Code To Chart