

Assessing the Accuracy of an Automated Coding System in Emergency Medicine

William C. Morris, PhD¹, Daniel T. Heinze, PhD¹, Homer R. Warner Jr., PhD²,
Aron Primack, MD FACP³, Amy E. W. Morsch, PhD¹, Ronald E. Sheffer, MA¹,
Mark A. Jennings, MS¹, Mark L. Morsch, MS¹, Michelle A. Jimmink, BA¹

¹A-Life Medical, Incorporated, ²3M Corporation, ³United States Uniformed Health Services

Accuracy and speed are imperative when it comes to coding medical records. Completely automated approaches to coding are faster than human coders, but are they as accurate? To measure accuracy, a “gold standard” is required; however, establishing a standard for medical records coding is problematic given the inherent ambiguity in some of the coding rules and guidelines. This paper presents statistics regarding the variability amongst experienced coders and compares this variability with an automated system, LifeCode®. The authors conclude that LifeCode is as accurate as the human coders used in this study and offers the potential for increased coding consistency and productivity.

Introduction

Automated coding systems hold the potential for increased coding speed and accuracy compared to unaided human coders. One automated system, LifeCode®, uses Natural Language Processing (NLP) techniques to provide both syntactic and semantic understanding to the complex job of autocoding [1]. LifeCode processes the transcribed text of emergency medicine records to assign procedure codes (Current Procedural Terminology or CPT© codes – including Evaluation and Management or E&M codes) and diagnosis codes (International Classification of Diseases, version 9, or ICD-9 codes). The issue of coding accuracy is of great concern for the entire industry, payers, physicians, and consumers alike. Coding accuracy, however, has been an elusive target. Not only is accuracy imperative, but so is speed. In an era of doing more with less, the productivity and efficiency of coding is crucial.

Healthcare payment by third party payers including Medicare is dependent upon having accurate coded data. Many researchers have studied coding accuracy. Over the past 25 years the literature shows human coding accuracy in medical records ranging from a low of 37.7 percent [2,3] to a high of 90.6 percent [4,5]. Dunn suggested that a performance standard of 97 percent accuracy is not achievable by most coders [6]. Yet, if accuracy depends on an elusive “gold

standard” then scoring high may be a function of who is doing the scoring.

Experts in the medical records field have recommended a variety of strategies to improve coding accuracy including: employment of credentialed coders, staff development through workshops and in-service training, development of coding policies and procedures, implementation of quality control programs and use of automated encoders [7]. With the exception of coding automation, most of these coding improvement strategies are widely used but on the whole have not produced both improved accuracy and speed. Because automated coding has the potential to be at least as accurate as experienced human coders while making a significant impact on productivity, a study was conducted to acquire a better understanding of the accuracy of an automated coding system, LifeCode, as compared to experienced human coders.

Methodology

A study was designed to test LifeCode against both production coders and expert consultants in emergency medicine coding. In order avoid bias from the study sponsors (A-Life Medical, Inc. which markets LifeCode) and also to keep the human coders from being biased by knowing the source and purpose of the study, the authors contracted with an outside market research firm to select, contact, and handle the transactions with the participating coders. The marketing firm was presented with a master list of coders and was asked to select a representative sample from among individual experts, premium billing companies and standard billing companies. Six human coders were selected to participate in the study, and one coding expert was selected to act as an auditor to blindly collate and review the results from the six coders and LifeCode.

One hundred charts were selected for the study and sanitized by removing the provider and patient names. Half of these charts were selected to be representative of cases that are most common in the emergency department and half were selected at

random. All charts were dictated, transcribed, and presented to the study participants in ASCII text format. To help eliminate potentially confounding factors in the experiment, all coders were required to agree in writing to the following guidelines:

1. All 100 records must be coded in a single session of no more than 7 hours, inclusive of breaks.
2. All charts are to be coded. If for some reason a chart cannot be completely coded, coders are to move on to the next chart and not return to the uncompleted chart.
3. Charts must be coded in the order presented, and once a chart has been coded it cannot be reexamined for review or changes.
4. The coding session shall start at the coder's normal business day which can be on a weekend.
5. Coding shall be conducted in an environment that is free from interruptions and distractions.
6. All coding shall be done according to the HCFA 1995 Documentation Guidelines for Evaluation and Management (E&M) Services. A summary of these guidelines is given to the coders for use as a reference.
7. ICD-9 codes are to include both E and V codes, and CPT codes are to include E&M codes and modifiers.
8. Coding is to include the correct sequencing of both ICD-9 and CPT codes.
9. Paper-based 1999 ICD-9 and CPT manuals, the attached summary of the 1995 HCFA guidelines, and a medical dictionary are the only reference materials to be used. The summary of the HCFA guidelines should be reviewed before starting the study.
10. Consultation with other individuals should neither be sought nor accepted.

All participants, including the auditor, coded each and every chart. Each chart was also processed in LifeCode to produce both an E&M code and the associated diagnoses and procedures codes. LifeCode has a "self-awareness" capability in terms of its own limits as compared to the complexity of a particular medical chart. It will, on difficult cases, request human assistance with regard to complex diagnoses and procedures or, in production settings, with respect to complex provider and payer specific billing requirements. Even though ICD and CPT codes were recorded and compared, the authors have chosen to focus only on the E&M CPT codes in this report because they represent about 80 percent of the reimbursement for emergency medical services and because they are the primary target for payer audits.

The CPT E&M level of service codes (99281 - 99285) for emergency medicine were recorded by each participant for all 100 charts. A chart coded as 99281 for professional services represents a patient case that is far less complicated and requires less physician work (e.g., uncomplicated insect bite) than one that is coded as 99285 (e.g., myocardial infarction). This type of patient severity level coding

was established to reimburse healthcare providers in proportion to the amount of time and effort required by the provider to work-up and treat the patient (i.e., a simple case gets a lower reimbursement than a complicated case). The rules governing the coding in this area were created by the AMA, and the guidelines for using them developed by HCFA and the AMA together.

Analysis of inter-coder agreement was done using the Kappa statistic which provides a conservative way of comparing how well one coder agrees with another taking into account chance agreement [8]. Kappa is calculated by dividing the difference between observed and random agreement by the maximum possible difference between observed and random agreement:

$$K = \frac{P_o - P_r}{1 - P_r}$$

where P_o is the observed agreement, P_r is the random agreement, and 1 is the maximum observed agreement. A Kappa value less than 0.40 indicates poor agreement. Kappa between 0.40 and 0.70 indicates fair to good agreement. Kappa above 0.70 indicates strong agreement. Since raw agreement rates (i.e., how often coder A agreed exactly with coder B in the aggregate) are widely used, results are also summarized using both exact agreement and Kappa. Each coder was also compared to the study group consensus.

Results

Inter-coder agreement results are presented in Table 1. In columns A-H, Table 1 reports the pair-wise comparisons between each coder including LifeCode and the auditor. Coders A, D and E are nationally recognized experts in emergency medicine coding. Coder F is a coder assigned from a premium billing company that is nationally recognized for quality. Coders C and G are coders assigned from standard, competent billing companies that are considered typical of the industry. Coder B is LifeCode. Coder H is the auditor who has an expert status comparable to coders A, D and E.

The left-most matrix in Table 1 is laid out as two half-tables: on the lower left are the agreement rates between the coders, on the upper right are the corresponding Kappas.

The Consensus columns in Table 1 shows the score of each coder against the consensus opinion on each chart. Because there were eight participants coding each chart, if there was a tie then the auditor's code

Table 1: A-H Inter-rater agreement (*italics*) and Inter-coder Kappa (roman); Agreement with the consensus; Kappa with the consensus; Number of charts deviating -2/-1/+1/+2 from consensus; Average RVU for EM levels.

	A	B	C	D	E	F	G	H	Consensus Agreement	Consensus Kappa	-2	-1	+1	+2	RVU
A (Expert)	*	0.35	0.24	0.31	0.34	0.43	0.51	0.34	0.73	0.57	3	19	4	0	2.00
B (LifeCode)	0.58	*	0.24	0.30	0.36	0.33	0.36	0.28	0.71	0.51	0	13	12	0	2.13
C (Standard Billing)	0.51	0.55	*	0.28	0.57	0.37	0.27	0.30	0.72	0.57	3	18	4	0	2.17
D (Expert)	0.51	0.52	0.49	*	0.30	0.45	0.40	0.37	0.68	0.46	3	12	11	0	2.14
E (Expert)	0.55	0.63	0.72	0.50	*	0.44	0.38	0.30	0.78	0.63	0	7	11	0	2.18
F (Premium Billing)	0.58	0.53	0.55	0.57	0.65	*	0.33	0.34	0.69	0.54	0	2	18	1	2.36
G (Standard Billing)	0.67	0.58	0.54	0.58	0.59	0.53	*	0.38	0.71	0.54	3	18	2	0	1.95
H (Expert/Auditor)	0.51	0.49	0.52	0.55	0.48	0.53	0.54	*	0.59	0.42	3	21	8	0	1.99

was removed from consideration in order to break the tie. The columns -2 to +2 report on the number of charts per coder that deviated by the indicated magnitude and direction from the consensus. In other words, if coder A scored a chart with 99283 but the consensus scored the same chart with 99281 then A would receive a +2 rating for that chart because A coded two levels higher than the consensus.

The RVU column represents the average relative value unit per the E&M codes of each coder. RVUs are used for comparison instead of the average E&M level because the value scale is not linear. Here the E&M level to RVU correspondence is per the December 1999 Federal Register [9]. Coders B, C, D and E all lie less than one-half standard deviations from mean RVU of 2.12, whereas A, G and H are one or more standard deviations below the mean and F is more than two standard deviations above the mean.

The differences in inter-rater Kappas as compared to individual rater versus the consensus is significant. The mean inter-rater Kappa is 0.38 (poor) with a standard deviation of only 0.08, whereas the mean rater versus consensus Kappa is 0.53 (fair to good) with a standard deviation of 0.06. The mean inter-rater Kappa plus one standard deviation (0.46) is still less than the mean rater versus the consensus minus one standard deviation (0.47). Because no individual had a strong Kappa (> 0.70) as compared to the consensus, we would hesitate to raise the consensus to a “gold standard”. This hesitancy seems further justified by the observation that it is possible to be very close to the consensus Kappa but significantly above (coder F) or below (coder G) the mean RVU level.

Expert study participants appear to show the least agreement between themselves. In terms of consensus agreement, LifeCode is about average amongst participants. With respect to the consensus,

LifeCode produced the most balanced distribution of codes, whereas expert coder A had the most prominent downward skew and Premium Billing company F had the most prominent upward skew.

If the consensus level for each chart is used as a gold standard, then the codes for most charts should cluster around the consensus and be evenly balanced one level above and below the consensus with a few outliers [10]. However, this is not the case. The typical pattern, as shown in Table 1, is significant deviation from the consensus, with clusters one level above or one level below rather than being balanced on both sides of the consensus.

Discussion

Among production coders in hospitals and billing companies, the most common check on coding accuracy is the level of agreement between a coder and some other coder who is accepted as an expert or standard. If the level of agreement is calculated between coders within the same organization who have been trained in the same practices and are held accountable to the same internal coding standards, the level of agreement can be quite high. Chao [11] noted a very high E&M agreement level for nurses trained in a similar methodology in a primary care environment. In order to achieve a high agreement level, methodology training was required and direct observations were obtained by the nurses themselves (coding was not done from the physician’s notes).

In discussions with the authors, various billing companies have claimed agreement rates of 95 percent to 98 percent. Similar agreement rates are reported by Lloyd and Layman [7]. However, these are situations where the supervisor is reviewing the coder’s work – not double-blind studies. In the declining number of cases where billing companies send a sample of their charts to another company

(one selected for similarity of philosophy and practice), much lower agreement rates are achieved. Rates on the order of 80 percent to 85 percent have been reported to the authors. Finally, double blind outside audits that A-Life has observed, show billing company codes versus the auditor in the 50 percent agreement range [12].

When audited, a coding organization must somehow justify its methodology and defend its discrepancies with the auditor. It becomes important that the type of errors (or discrepancies) generated by the coder or system, be identified, explained and justified.

The subjective nature of coding: Subjectivity in medical records coding results from the wide variety of medical conditions reported, the language used by the care giver, the various interpretations of coding guidelines, and the complexity of the documentation guidelines. Human coders are required to extract a large number of details and distill these details into the few codes that accurately describe the patient's visit.

Differences of opinion may not necessarily mean that one person is right and the other is wrong. A coder's perspective or work environment often influences the way charts are interpreted. Because coding guidelines are general rules for measuring and categorizing the work of clinicians, the specific application of the guidelines is subject to human interpretation. Published coding guidelines contain one or two prototypical examples of each E&M level of service code. From these guidelines alone there is no way to address every possible combination of presenting problems, medical history, medical decision making, and final diagnosis. Human coders apply their best judgement to assign level of service codes. This is the seed of defensible disagreement. However, in an audit, the auditor is considered to be always right, whether the coder's position is defensible or not. However, in the eyes of the auditor some types of errors are more forgivable (less expensive) than others.

Coding errors: There are broadly two categories of errors that coders make in decision-making tasks, performance errors and systematic errors [13,14,15,16,17]. Performance errors consist of mistakes and slips. In the realm of medical coding, mistakes involve misreading words in the source document or missing details altogether, resulting in a misunderstanding of document content. An example of a mistake would be failing to see negation or failing to pull details together from across the document. Slips involve failing to carry out the

intended coding assignment, for instance the transposition of two digits in a code.

Systematic errors, on the other hand, consist of knowledge- and rule-based errors. Knowledge-based errors occur because the coder lacks the medical knowledge to grasp the situation at hand – often due to inadequate training and experience in the field. Rule-based errors involve a misapplication of rules. The coder may well understand the document and correctly assess the medical case it describes, but incorrectly assign a code. This type of error might result in systematic up-coding – a serious problem from an auditor's perspective. For example, all procedures from similar classes might get lumped together and assigned the code appropriate only for the highest-class procedure. In a legal or regulatory dispute, local policies that result in systematic up-coding can be disastrous for the billing company or the physician involved. Systematic up-coding that is felt to be intentional to charge more money is considered fraudulent and resulting penalties can be severe.

Errors in automated coding systems: Performance errors, like "mistakes" or "slips," are rare in computerized coding systems, but systematic errors can occur. Automated systems can be overly sensitive to typographical errors, formatting changes and stylistic variants. All processing and understanding is based on previously seen texts. Ability to create generalization and handle novel expressions is less than for humans.

Idiosyncratic or obscure terminology, use of idioms, and occasionally tortuous syntax, can cause the computer to make a "performance error." Phrases like "a change of heart" or references to a "court hearing" or to "Palm Beach" can cause a system to postulate problems with the heart, the ears, or the hand. NLP systems must acquire significant real-world knowledge to avoid such mistakes.

The systematic errors of automated decision-making systems are very similar to those of human decision-makers. Both knowledge-based errors and rule-based errors can occur. Knowledge-based errors are straightforward; an inadequate database causes the system to "fail to sufficiently understand" the document, leading to an inadequate assessment of the medical case being coded. As with a human coder, a rule-based error occurs when the system correctly assesses the medical situation, but applies the wrong rule and assigns the wrong code to the case.

For human coders both performance and systematic errors can be addressed with better tools and more training. For automated coding systems the corresponding errors can be addressed with spell-checking, with greater linguistic sophistication in the information extraction algorithms, and by increasing the size and sophistication of the knowledge bases.

Conclusion

Accuracy in E&M coding is relative. To the auditor, coding accuracy is precise agreement with their way of coding. To the physician, coding accuracy may be maximizing their reimbursement for work performed and defensibly documented. To the human coder, coding accuracy may be agreement with their supervisor's rules for coding. To the vendor of an coding automated system, accuracy may be avoidance of systematic up-coding. Perspective and motivation changes coding outcomes, especially when left with loose guidelines to govern behavior.

This study provides evidence of how significant disagreement arises when general guidelines are applied to specific situations. If an E&M coding consensus can be used to represent a "gold standard," then only moderate agreement was observed between any of the study participants and this gold standard. Under these circumstances, automated programs like LifeCode have a significant advantage over human coders in being predictable, repeatable, and fast.

In a regulatory and reimbursement environment where a "gold standard" does not yet exist, using a coding methodology that is as accurate and defensible as the norm is acceptable. LifeCode is as accurate as any auditor, it is more consistent than any auditor, and, because it is algorithmic, the basis of any of its decisions can be explicitly shown and reviewed for consistency with published regulations.

The issue then becomes speed and productivity. Given the shortage of experienced human coders, the tedious nature of medical records coding, and the inherent variability in human coding one can argue that the future of automated coding systems looks bright.

References

1. Heinze, DT, et.al., A Natural Language Processing System for Medical Coding and Data Mining, *AAAI - Twelfth Innovative Applications of Artificial Intelligence Conference*. Forthcoming.

2. Currie, MS. Clinical Data Quality: Impact on Revenue. *J Am Med Rec Assoc*. 57:15-17, 1985.
3. Schraffenberger, LA. Coding Errors Encountered in DRG Study. *J Am Med Rec Assoc*. 57:15-17, 1986.
4. Hsia, DC., et al. Accuracy of Diagnostic Coding for Medicare Patients user the Prospective-Payment System. *NEJM* 318:352-355, 1988.
5. Hsia, DC. et al., Medicare Reimbursement Accuracy under the Prospective Payment System, 1985 to 1988. *JAMA* 268:896-873, 1992.
6. Dunn, R. Performance Standards for Coding Professionals. *For the Record*, 15 Nov. 1993: 4-6.
7. Lloyd SS, Layman E. The Effects of Automated Encoders on Coding Accuracy and Coding Speed. *Topics in Health Information Management*. Vol. 17. No. 3, February, 1997.
8. Fleiss, JL, *Statistical Methods for Rates and Proportions*, 2nd Ed. 1981, John Wiley & Sons, Inc.
9. http://www1.access.gpo.gov/GPOAccess/sitesearch/su_docs_aces/aces140.html
10. Hripsak G, Kuperman GJ, Friedman C, Heitjan DF. A Reliability Study for Evaluating Information Extraction from Radiology Reports. *Journal of the American Medical Informatics Association*. Vol. 6, No. 2, Mar/Apr, 1999.
11. Chao J, Gillanders WG, Flocke SA, Goodwin MA, Kikano GE, Strange KC. Billing for Physician Services: A Comparison of Actual Billing with CPT Codes Assigned by Direct Observation. *The Journal of Family Practice*. Vol. 47, No. 1, July, 1998.
12. Private audits performed for customers of A-Life Medical, Inc.
13. Norman DA. Categorization of action slips. *Psychological Review*, 88, 1-15. 1981.
14. Norman DA. *The psychology of everyday things*. New York: Harper & Row; 1988.
15. Reason JT. Lapses of attention. In Parasuraman R & Davies R editors, *Stress and fatigue in human performance*. Chichester UK: Wiley; 1984.
16. Reason JT. *Human error*. New York: Cambridge University Press; 1990.
17. Wickens CD. *Engineering psychology and human performance*. New York: HarperCollins; 1992.

This paper was published in *Proceedings of the AMIA 2000 Annual Symposium*. American Medical Informatics Association: November 2000.
