Assessing Coder Change Rates as an Evaluation Metric

by Jean Stoner, CPC, RC; Michael Nossal, MA; Philip Resnik,PhD; Andrew Kapit; and Richard Toren

Abstract

In a typical model for computer-assisted coding, the system automatically identifies codes, which are then improved by a human review coder in order to more closely approach the codes that would be assigned to the note by an idealized "gold standard" coder. Assuming such a model, a natural method for evaluating system accuracy might be the coder change rate—specifically, the percentage of notes approved by a coder without modification in a day-to-day business setting. In this paper we look more carefully at change rates as a measure of coding accuracy.

Introduction

Computer-assisted coding (CAC) is one in a family of applications in which an automatic system analyzes text data in order to add information that is then reviewed, and possibly corrected, by a human expert. Such tasks have been familiar in natural language processing (NLP) for a long time.[1-3] A useful illustration is the Penn Treebank project. Undertaken at the University of Pennsylvania in the early 1990s, the project used an approach one might call computer-assisted annotation to augment large bodies of text with information that is useful to linguists and NLP developers.[4] Such information included, for example, word-level grammatical categories and phrase-level constituency information. More generally, there is a considerable literature on processes and tools for creating and manipulating annotated corpora; see, for example, *Treebanks: Building and Using Parsed Corpora*, "Speech Annotation and Corpus Tools" in the special issue of *Speech Communication*, and *Proceedings of the ACL 2004 Workshop on Discourse Annotation*.[5-7] A computer program automatically analyzed text and generated the linguistic information, and then specially trained linguistic annotators corrected the output, operating a user interface carefully tailored for that purpose.

A persistent problem for those developing automatic or computer-assisted annotation, however, is how to evaluate the automatic systems. In order to achieve rigorous comparisons, the usual practice in NLP is to create a "gold standard" against which to compare uncorrected system output. Unfortunately, creating gold standards can be complicated and labor intensive. Furthermore, once created, gold standards are a static resource, even though in practice the real-world material to be annotated may be changing, and experience and insight may have led to changes in the annotation guidelines.

When looking at CAC as an annotation task, a potentially attractive solution to this problem presents itself. CAC generally takes place within an environment similar to the Penn Treebank annotation workflow, in that human experts review and/or correct (some subset of) the system's automatically generated output. In contrast to linguistic annotation projects, however, CAC takes place on a vastly larger scale, and it is embedded in an ongoing business process, day in and day out. Every time a human coder changes the codes for an automatically coded note, or approves the codes without changing them,

one might say an evaluation event has taken place as a part of the business itself. Why not, then, use those evaluation events to define what it means for an automatic system to be coding well?

In this paper we explore that possibility. Section 2 briefly describes the coding engine and workflow for a successful, widely deployed CAC solution. Section 3, the main body of the paper, drills deeper into a key issue that arises when attempting to use coder change rates as an evaluation measure—the problem of human reviewers making incorrect changes—including discussion of both systematically obtained and informally observed data. Section 4 concludes with a summary and discussion of what was learned.

CAC Workflow

The context for this study is a successful, widely deployed commercial CAC solution that utilizes a sophisticated NLP engine and a secure, Web-based coder workflow. Although the system is in use in a variety of medical specialties, we restrict our attention here to radiology. Medical dictations enter the system in electronic form (or are converted to electronic form), and undergo a process of analysis, structural normalization, and other forms of preprocessing. An NLP engine identifies relevant units of evidence within the note body and metadata, extracts relevant local and contextual features, and applies a combination of domain knowledge and statistical classification techniques to associate Current Procedural Terminology (CPT) and International Classification of Diseases (ICD) codes with evidence and feature combinations. A sophisticated logical component governs the selection and ordering of codes to report given the underlying evidence, extracted features, and predicted codes; these choices implement "most certain" coding guidelines, proper identification of incidental versus pertinent diagnoses, appropriate treatment of equivocal language, and numerous other complexities of proper coding. At all times, the driving principle of engine coding is compliance with all applicable coding rules and guidelines, although numerous preferences driven by customer or payer may come into play at different stages in the coding process.

Human review of automatically coded output is governed by a combination of rule-based filtering and a rigorous statistical confidence assessment model, which predicts with extremely high accuracy whether the customer's own coders would approve the entire note without modifications. Notes for which this prediction meets customer-defined criteria—typically 98 percent confidence for CPT codes and 95 percent for primary ICD—are sampled in small numbers for quality assurance. Jiang et al. discuss the confidence assessment process in more detail and describe a case study formally comparing the technique to less sophisticated alternatives. They find that using the confidence assessment model, the strict 98 percent/95 percent criteria are met for 38 percent of the notes in a 10,998-note sample.[8]

Notes that go to full human review are divided into two groups. Those for which there is evidence of missing documentation are placed in the "code" category. For notes in this category, typically 10 to 15 percent of the total, the engine is unable to discern codable information from the available documentation and can be interpreted as handing the note off entirely to a human coder. Interestingly, it is typical in practice for 25 to 55 percent of notes in this category to be approved without changes; for example, for many notes, the engine reported that there was no documentation because there was, in fact, no documentation. In addition, when codes are changed in this category, it is not uncommon for the human coder to assign codes only after a phone call to the referring physician has been made to obtain a clinical indication that was missing in the note as given to the NLP engine.

The remaining group of notes comprises the "review" category. In this category, it is typical to see changes to 5 to 15 percent of CPT codes and 10 to 25 percent of primary ICDs.

Table 1 reports the percentages of CPT and primary ICD codes changed by customer coders for a large billing company, for a typical week during summer 2006. Notes in the confident queue were not reviewed exhaustively; the human coder change rates are based on a subset sampled for purposes of quality assurance.

The table confirms that for notes in the confident queue, the statistical model successfully enforced the stringent criteria defined by the customer. Furthermore, change rates in the review queue (rates of disagreement between human and machine) are generally consistent with rates of intercoder agreement

for human coding experts coding independently interface.[9] Human review agreed with some 42 percent of CPTs and 53 percent of primary ICDs even in the code queue.

Analyzing Coder Changes

In a commercial setting, where selected CAC output is reviewed by human coders during an ongoing business process, change rates like those in Table 1 reflect how much revision of autogenerated codes is taking place. Clearly the degree of revision necessary is a function of the engine's coding quality. Therefore, to the extent that human corrections conform to the correct coding for a note, change rates in an ongoing business process are a natural way that autocoding quality can be measured. In principle, this might make it possible to do without alternative evaluation methods such as formal gold-standard comparisons or customer-independent audits. [10-11]

We examined real-world customer coder behavior in the work flow of Section 2 in order to assess whether it would be appropriate to use coder changes for this purpose. On the one hand, we found that in the majority of cases, changes by human coders are in fact correcting errors, as expected, and doing so correctly. However, we also found cases in which correct codes were regularly changed. The remainder of this section presents our findings based on both systematic analysis and informal observation.

Systematic Study of Two Specific Changes

Abdominal Ultrasounds. The American Medical Association, the authoritative body for CPT-4 coding, added specific guidelines in 2005 for assignment of CPT-4 codes for abdominal and retroperitoneal ultrasounds. In order to code a complete abdominal ultrasound (76700, ultrasound, abdominal, B-scan and/or real time with image documentation; complete) the documentation must include "B mode scans of liver, gall bladder, common bile duct, pancreas, spleen, kidneys, and the upper abdominal aorta and inferior vena cava [ivc] including any demonstrated abdominal abnormality." If all of the organs listed above are not documented, the limited CPT (76705, ultrasound, abdominal, B-scan and/or real time with image documentation; limited (e.g., single organ, quadrant, follow-up)) must be assigned.

Having observed anecdotally that limited ultrasounds are often changed to complete, we investigated this change pattern more systematically. From a database of approximately 2 million recently coded notes, we randomly selected 35 reports for which a customer coder changed the engine-assigned 76705 (limited abdominal ultrasound) to 76700 (complete abdominal ultrasound). On manual review, an expert coder/auditor found that in 32 of the 35 cases (91.4 percent), the human coder had made the change incorrectly in a note missing required documentation as stated by the AMA. Although N=35 is a small sample, the pattern is unquestionably significant ($z=6.93$, $p < .00001$).

Retroperitoneal Ultrasounds. As with abdominal ultrasounds, retroperitoneal ultrasounds are also subject to detailed documentation requirements as defined by the AMA. In order to assign 76770 (ultrasound, retroperitoneal (eg, renal, aorta, nodes), B-scan and/or real time with image documentation; complete), the guidelines state, "A complete ultrasound examination of the retroperitoneum (76770) consists of B mode scans of kidneys, abdominal aorta, common iliac artery origins, and inferior vena cava, including any demonstrated retroperitoneal abnormality. Alternatively, if clinical history suggests urinary tract pathology, complete evaluation of the kidneys and urinary bladder also comprises a complete retroperitoneal ultrasound." If all of the organs listed above are not documented, the "limited" CPT (76775) must be assigned, unless urinary tract pathology is documented, in which case just the bladder and kidneys are required for 76770.[12]

As was done for abdominal ultrasounds, we randomly selected 35 retroperitoneal ultrasound reports where the coder changed the engine-assigned 76770 (complete retroperitoneal ultrasound) to 76775 (limited retroperitoneal ultrasound). The expert coder/auditor found that of these "corrections," 28 (80 percent) changed the engine's correct output to the incorrect code. Again, despite the small sample, statistical significance is confirmed ($z=5.02$, $p < .00001$).

Although conducted on a small scale, these observations highlight the fact that one of the strong points of CAC is fast and reliable counting of documentation elements. For human coders, in contrast, this is a very time-consuming and often inaccurate process.[13]

Additional Observations

Because analyzing coder changes is a laborious process, the brief studies in Section 3.1 suffice for the moment as formal evidence that there are systematic cases of coder miscorrection. (Note, however, that large-scale, fully automatic analysis of coder change behavior can be used for a different purpose, to automatically learn which engine outputs should and should not be considered most reliable.)[14] Informally, however, coding experts familiar with workflow and client coder changes add the following observations.

Screening mammograms. Ambiguity in language can lead to conservative and less conservative interpretations. For example, if "routine" is the only reason provided for a mammogram, a conservative interpretation assumes that the procedure is a routine or screening mammogram, in which case the appropriate CPT code is 76092 (screening mammography, bilateral (two view film study of each breast)). Sometimes, however, a documenting physician dictates "routine" to indicate "routine views," without providing any actual documentation of the fact that the mammogram is diagnostic. In such cases we see human coders modifying 76092 (screening mammogram) to 76091 (diagnostic mammogram), overriding the engine's correctly conservative choice in the absence of any explicit evidence that would make it appropriate to do so.

Renal cysts. Although the official ICD-9-CM coding guidelines require assigning 753.11 (congenital single renal cyst) for "renal cysts," many coders prefer to assign 593.2 (cyst of kidney, acquired) because in most cases, kidney cysts are acquired and not congenital. Regardless of "common sense," however, a coder must follow the guidelines promulgated by the ICD-9-CM book and the Coding Clinic. CodeRyte subject matter experts report often seeing clients' coders incorrectly changing 753.11 to 593.2.

Preoperative chest x-rays. The Coding Clinic has clarified that whenever a patient is having a chest x-ray to clear the patient for surgery, V72.83 (special investigations and examinations, other specified preoperative examination) must be assigned; however, some coders believe that V72.84 (special investigations and examinations, preoperative examination, unspecified) is the most suitable ICD-9. CodeRyte always follows the guidelines provided by the ICD-9 coding book, unless there is more specific guidance from the Coding Clinic; hence coder changes from V72.83 to V72.84 are instances of altering a correct code.

Accidents coded as injury. If the reason for an exam is listed as an accident (e.g., "fell down stairs" or "MVA"), and no specific injury or pain is documented, nor are there documented any positive findings, then V71.4 (observation following other accident) must be assigned. Many coders assume that an injury must have occurred, and rather than assign V71.4, they assign an ICD-9 from the 959 series (injury). Unfortunately, this is not accurate coding, since it assumes there is an injury. As a result, coders are often seen incorrectly changing engine output from V71.4 to a code in the 959 series.

E-codes. Despite stated customer guidelines as well as correct coding guidelines, coders sometimes delete valid ICD-9 codes supported by documentation. The codes deleted span a wide range, though observed cases in an informal sample tended to be dominated by E-codes (e.g. E888.9, fall, when documentation states "fell on left shoulder two months ago;" E819.9, motor vehicle accident, when the note clearly reports "MVA a few days ago").

Discussion and Conclusions

In the day-to-day process of CAC, human coders implicitly perform an evaluation task on every note they review. However, medical coding includes many gray areas, and the definition of the "correct coding" for a note can depend on the individual coder. In this study, we identified a number of regularly

occurring cases where coder changes made in the course of day-to-day customer review cannot be viewed simply as correcting errors made by the coding engine.

One category we found included distinctions between complete and limited ultrasounds, where counting imaged organs may be error-prone for human coders. Another category involves situations where real-world contingencies favor the likelihood of a particular inference (such as the fact that exams after accidents usually imply there were injuries), but strict application of coding guidelines forbids the inference and requires explicit documentation (the injury has to be documented explicitly). A third category involves deletion of valid codes supported by documentation, possibly because the coder judges them to be incidental and therefore irrelevant to billing, or as a result of reimbursement/payer requirements—for example, many insurance companies do not accept ICD-9 E-codes (external causes of injury and poisoning), even though they are valid ICD-9 codes and should be included according to correct coding guidelines.

Taken together, our findings suggest that despite its practical appeal, simple change rate analysis is unlikely to be an adequate substitute for rigorous, customer-independent comparison against a gold standard, if the goal is evaluation of coding with respect to correctness. More generally, when looking beyond the correct codes to post hoc changes of coding engine output, understanding why coders make changes is as important an evaluation issue as the quantitative summary of the changes themselves.

Jean Stoner, CPC, RCC, is a Coding Analyst for NLP at CodeRyte.

Michael Nossal, MA, is a Senior NLP Engineer at CodeRyte, Inc., in Bethesda, MD.

Philip Resnik, PhD, is a Strategic Technology Advisor for CodeRyte, Inc., in Bethesda, MD, and an associate professor at the University of Maryland in the Department of Linguistics and the Institute for Advanced Computer Studies.

Andrew Kapit is CEO of CodeRyte, Inc., in Bethesda, MD.

Richard Toren is co-founder and president of CodeRyte, Inc.

Notes

1. Allen, Jeffrey. “Post-editing.” *Computers and Translation: A Translators Guide*. Somers, Harold, Editor. Benjamin’s Translation Library, 35. Amsterdam: John Benjamins, 2003.

2. Atwell, E. *LOB Corpus Tagging Project: Manual Postedit Handbook.* Department of Linguistics and Modern English Language and the Department of Computer Studies, University of Lancaster, 1982.

3. Marcus, Mitchell, et al. “Building a Large Annotated Corpus of English: The Penn Treebank.” *Computational Linguistics* 19, no. 2 (1993); 313-330. Reprinted in Armstrong, Susan, Editor. *Using Large Corpora.* Cambridge, MA: MIT Press, 1994, p. 273-290.

4. Ibid.

5. Abeillé, Anne. *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers, 2003.

6. Bird, Steven and Jonathan Harrington, Editors. “Speech Annotation and Corpus Tools.” Special Issue of *Speech Communication* Volume 33, numbers 1-2 (2001).

7. Webber, Bonnie and Donna Byron. *Proceedings of the ACL 2004 Workshop on Discourse Annotation*. Barcelona, Spain, July 2004. Available at http://acl.ldc.upenn.edu/acl2004/discourseannotation/.

8. Jiang, Yuankai, et al. “How Does the System Know It’s Right? Automated Confidence Assessment for Compliant Coding.” *Computer-assisted Coding.* Presented at AHIMA/FORE Computer-assisted Coding Software Standards Workshop, Arlington, VA, September 2006.

9. Nossal, Michael, et al. “Using Intrinsic and Extrinsic Metrics to Evaluate Accuracy and Facilitation.” *Computer-assisted Coding*. Presented at AHIMA/FORE Computer-assisted Coding Software Standards Workshop, Arlington, VA, September 2006.

10. Ibid.

11. Heinze, Daniel, et al. “Computer-assisted Auditing for High Volume Medical Coding.” Presented at AHIMA/FORE Computer-assisted Coding Software Standards Workshop, Arlington, VA, September 2006.

12. The AMA has not defined what urinary tract pathology is, but CodeRyte has created a list of ICD-9 codes that accurately captures most urinary tract pathology diagnoses.

13. In 2004, Medicare paid out $87,000,000 for 1,238,668 claims coded as 76700 and $71,000,000 on 1,049,135 claims coded as 76770. (Source: CMS Web site.) With a difference in reimbursement of $20 between the complete and limited codes, if a substantial portion of the human-coded complete ultrasounds should really be reimbursed as limited, the financial implications are potentially quite significant.

14. Jiang, Yuankai, et al. “How Does the System Know It’s Right? Automated Confidence Assessment for Compliant Coding.”

Table 1

Change rates by queue (percentage), anonymous CodeRyte customer during a representative week in summer 2006

	Code	**Review**	**Confident**
CPT	57.72	9.6	0.12
Primary ICD	46.51	24.1	0.84
All ICD	69.98	39.44	1.46