Using Intrinsic and Extrinsic Metrics to Evaluate Accuracy and Facilitation in Computer-assisted Coding

by Philip Resnik, PhD; Michael Niv, PhD; Michael Nossal, MA; Gregory Schnitzer, RN, CCS, CCS-P, CPC, CPC-H, RCC, CHC; Jean Stoner, CPC, RCC; Andrew Kapit; and Richard Toren

Abstract

Computer-assisted coding (CAC) evaluation can be viewed from two related but distinct perspectives: accurately identifying codes, and more generally facilitating the process of human coding and information flow within and among healthcare organizations. The first perspective calls for intrinsic metrics, grounded in comparison with a gold standard. The second perspective calls for extrinsic techniques, grounded in the real-world task. We discuss this distinction and present an experimental evaluation of CAC using both intrinsic and extrinsic measures.

Introduction

Over the last 20 years, the state of language technology evaluation has changed almost as dramatically as the technology itself. In the 1980s, it was common for research and development in natural language processing (NLP) to produce handcrafted, knowledge-based systems that were evaluated by means of scripted demonstrations using a handful of carefully selected examples. Today, progress in NLP is driven forward by combining linguistic and domain knowledge with statistical analysis of large quantities of data. Evaluation takes place formally within a set of widely accepted paradigms involving careful experimental design, well-defined metrics, and large samples of previously unseen test data.[1] There are numerous advantages to formal evaluations of this kind. Hirschman and Thompson point out that formal evaluation facilitates rapid technical progress, as well as the ability to track progress over time. They also observe that large advances take place when a "community of effort" emerges through work on a shared task.[2]

A distinction between intrinsic and extrinsic measures is widely accepted in the language technology community.[3-5] Intrinsic evaluations measure the performance of an NLP component on its defined subtask, usually against a defined standard in a reproducible laboratory setting. Extrinsic evaluations focus on the component's contribution to the performance of a complete application, which often involves the participation of a human in the loop.

Information about evaluation in the coding industry comes primarily from review or auditing of notes coded by human coders. However, post hoc reviews can overestimate levels of agreement when complex or subjective judgments are involved, since it is more likely that a reviewer will approve of a choice than it is that they would have made exactly the same choice independently. In the experimental literature, evaluation of agreement among independently assigned medical codes has focused almost entirely on

human coding without computer assistance.[6-8] A notable exception is a 2000 study by Morris et al investigating automatic coding for emergency medicine, in which the team performs an intrinsic evaluation comparing system output to codes assigned independently by production coders and coding experts. However, their study is limited to the five evaluation and management (E&M) level-of-service codes, despite the fact that procedure and diagnosis codes apparently were also obtained and compared. We have found little experimental literature involving extrinsic measures for CAC, although Morsch et al present a case study documenting business-level benefits of CAC, such as increased coder productivity, in a large physician practice.[9]

Evaluating CAC requires understanding both the performance of the automatic coding component and the value of the technology as a whole in the real-world task. In this paper, we describe an experimental study looking at both—via intrinsic and extrinsic metrics. To our knowledge, this is the first time formal evaluation measures have been reported for coding of procedures and diagnoses, as well as the first evaluation involving extrinsic measures in a formal experimental setting. The intrinsic results demonstrate performance comparable to human coders, and the extrinsic results empirically establish significant facilitation for inter-coder agreement and intra-coder consistency when the technology is used to assist human coders at their task.

Methods

When evaluating a system intended to match human expert performance, issues to address include defining test data, selecting performance measures, determining what responses the system should produce, and deciding whether particular levels of performance are "good enough." Here we present our approach to these issues.

Test Data

The test collection for this study comprised a random sample of 720 radiology notes from a single week in summer 2006, from a large teaching hospital.

Intrinsic Agreement Measures

Automatic coding can be viewed as a task that involves assigning an annotation or set of annotations to each medical note. In other NLP scenarios, documents might be news articles to be auto-annotated with labels that identify the article's topic (such as politics or sports). For computer-assisted coding, documents are dictated medical reports and the annotations identify CPT codes for procedures and ICD codes for symptoms and diagnoses. Current U.S. standards for procedure coding use CPT-4, a set of approximately 8,500 codes, and ICD-9, a set of approximately 16,000 codes. In settings where multiple valid annotations are assigned, for example multiple CPT codes for a single note, it is common to evaluate precision and recall. Expressed in terms of true (t) and false (f) positives and negatives (p and n), precision $P = tp/(tp+fp)$, and recall $R = tp/(tp+tn)$. That is, P focuses on the quality of what did get reported, and recall focuses on whether anything was missed. Precision and recall are usually combined into a single figure of merit by computing their weighted harmonic mean, or F-measure[10]. When the two are weighted equally, $F = 2PR/(P+R)$.[11] We report F-measure for CPT and ICD codes, as well as simple accuracy (proportion correct) for primary ICD codes.

Extrinsic Measures

We use the three agreement measures discussed above to assess human coders' output with and without computer coding technology, providing assessments of agreement with a gold standard, inter-coder agreement, and intra-coder consistency.

Gold Standard

In order to quantify performance on a task, it is common to create a "gold" (reference) standard by having independent experts perform the task and resolving cases where they differed.[12] We established gold standard coding for our test set by having two experts, Expert1 and Expert2, independently code the 720 notes. Notes where they differed on CPT or primary ICD were coded independently by a third expert, Expert3, and the third set of judgments used in a voting-based arbitration procedure to resolve the differences.[13] In a relatively small number of cases Expert3 disagreed with both Expert1 and Expert2.[14] We therefore constructed two gold standard variants, one where three-way ties were broken by preferring Expert1, and the other where Expert2 was preferred. Evaluations with both variants led to the same patterns of results, and therefore we report measures using the Expert1 variant.[15]

Upper Bounds

Upper bounds define the best performance one can expect from a system. Especially for tasks involving complex or subjective judgments, standard practice in NLP is not to simply fix "perfect" performance expectations at 100 percent, but rather to "estimate an upper bound on performance by estimating the ability for human judges to agree with one another."[16] Accordingly, we will define upper bounds on our evaluation measures by comparing the agreement of Expert1 and Expert2 using those measures. (See Table 1.)

Our main comparison is between two conditions: coding on paper ("Paper") and coding using CodeRyte's CAC technology ("NLP"), which makes codes provided by the NLP engine available for review or modification by the coder. Twelve professional production coders each coded online batches of notes containing subsets of the 720-note corpus in a balanced design to avoid ordering effects. In order to permit the measurement of intra-coder consistency, coding was done in two sessions, four days apart. In the second session, coders saw a small random subset of notes they had coded previously in the first session, mixed in randomly with the new notes to code.

Coders were assured that their identities would be replaced with anonymous identifiers for evaluation purposes. They were instructed to code as normally as possible (using whatever resources they usually use, for example), without a time limit, with the exceptions that they should code as continuously and with as few interruptions and distractions as possible, should not communicate with other coders or their coding supervisors, and should not skip notes. All the coders code regularly in adherence to CodeRyte's general and radiology-specific coding guidelines, and they were instructed to carefully follow those guidelines in this study.

Results

For each of the agreement measures, Table 1 shows agreement between our expert "gold standard" coders, which constitutes an upper bound on the measure. It also shows NLP engine agreement with the gold standard, and, for comparison, mean production coder agreement with the gold standard when coding on paper. If we define "human performance" as the agreement of production coders with the gold standard, then the engine performs at 97.8 percent of human performance for CPT F-measure (i.e. 83.3 compared to coders' 85.7), 92.7 percent for correctly coding primary ICDs, and 85.0 percent of human performance for ICD F-measure.[17] As predicted, formal comparisons of fully independent coding yield lower human coder performance figures, in absolute terms, than those commonly (and often anecdotally) reported for post hoc code review.[18, 19]

It is important to note that our results tables show performance aggregated over the entire set of notes. In production, a rigorous statistical confidence assessment model identifies a substantial portion of notes for which the engine's coding can be regarded as confident. Elsewhere, we report on a formal study in which the model identifies 38 percent of the test set as confident. In confident notes, CPTs are 100 percent correct in 98.36 percent of the notes and primary ICDs are all correct in 95.1 percent of the notes, as measured via (post hoc) customer review.[20]

Table 2, table 3, and table 4 compare on-paper versus computer-assisted coding.

Table 2 shows a comparison for production coder agreement with the gold standard. Coding with the benefit of NLP is consistently better than on-paper coding, with results all achieving statistical significance. (Assume $p < .05$ throughout unless stated otherwise. For CPT F and ICD F, comparisons use a t test. For ICD-1, we use a z-score test for significant difference of two proportions.)

Table 3 shows means of pair-wise inter-coder agreement measures among the production coders. Again, coding with NLP significantly outperforms coding on paper ($p < .005$).

Table 4 shows the same pattern of strong improvement for intra-coder consistency. Despite a small sample size, the difference in CPT is significant.

Although we are focusing here on correct coding and on consistency within and between coders, an additional extrinsic factor that clearly matters is coder productivity. In order to include some discussion of that issue, we briefly present production figures. Within CodeRyte's CAC workflow, coded notes are segregated into work queues. For the queue containing confident notes (38 percent of the test set in our confidence assessment study), only a small fraction need to be reviewed at all, for quality assurance. [21] Examining our mean-notes-per-hour reports for a set of 17 customers, for a representative week during summer 2006, we found that the average over the 17 was about 276 notes per hour, for the notes in the "confident" queues that received review, with corresponding CPT and primary ICD post hoc change rate averages of 2.3 percent and 3.9 percent, respectively. In the slowest queue (containing notes which the engine either found no codes, or found relevant language that it did not have enough information to code) that same average was 97 notes per hour, with CPT and primary ICD change rate averages of 41.9 percent and 64.2 percent. (For a "middle" queue, containing coded notes where review is needed, coding speed averaged 136.8 notes per hour with change rates of 8.7 percent for CPT and 28.7 percent for primary ICDs.) Overall, we have found consistently that the use of the CodeRyte CAC workflow leads to large productivity gains over traditional coding; studying this more formally is a topic for future work.

Conclusions

One central question in CAC evaluation is how to assess the accuracy of the underlying coding engine. The NLP community has found significant advantages in assessment methods that are rigorous, reproducible, and that permit fair comparisons of alternative systems on the same system as it improves over time. For CAC evaluation, this means creating and using a gold standard. Along with others in the CAC industry, we have found that true rates of inter-coder agreement may not be as high as has sometimes been claimed.[22] The way to find out is rigorous experimentation, especially targeted at extrinsic measures.

We have shown that creating a gold standard is practical, and we showed that our coding engine performs strongly relative to human performance on relevant agreement measures. We would argue that agreement on independently coded notes provides a more realistic and appropriate picture of both human and system performance than small demonstrations or post hoc measures of agreement.

Another central question is how to assess the value of the technology in the context of the larger task. We have shown experimentally that compared to traditional on-paper coding, using CAC technology yields greater agreement with a gold standard, greater inter-coder agreement, and greater consistency as measured by intra-coder agreement. Our formal results, documenting significant gains in correctness and consistency, complement our production-level findings concerning increased productivity and other returns on investment we see when our system is deployed.

As an application of our results, suppose one wanted to improve inter-coder agreement on a team of production coders. What is the best approach? One could find expert coders, or one could deploy CAC. Table 5 compares these approaches, using inter-coder agreement figures from our study to hypothesize relative gains afforded by these approaches over the status quo. The Use Expert Coders line assumes all coders on the team are performing at an expert level. We suspect that equipped with reliable evaluation figures, a comparison of costs, risks, and benefits would make CAC a very attractive option, indeed.

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.

Michael Niv, PhD, is Lead Software Architect at CodeRyte, Inc., in Bethesda, MD.

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

Gregory Schnitzer, RN, CCS, CCS-P, CPC, CPC-H, RCC, CHC, is Director of Product Development at CodeRyte, Inc., in Bethesda, MD.

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

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

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

Notes

1. Manning, C. and H. Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, MIT Press, 1999.

2. Hirschmann, Lynette and Henry S. Thompson. "Overview of Evaluation in Speech and Natural Language Processing" *Survey of the State of the Art in Human Language Technology*. Cole, Ronald, Editor. Cambridge, England, 409-414, Cambridge University Press, 1998.

3. Goldstein, Jade, et al (Editors). *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Ann Arbor, Association for Computational Linguistics, June 2005.

4. Molla, D. and B. Hutchinson. "Intrinsic versus Extrinsic Evaluations of Parsing Systems." Workshop on Evaluation Initiatives in Natural Language Processing, 11th Conference European Chapter of the Association for Computational Linguistics, Budapest, April 2003. Association for Computational Linguistics.

5. Resnik, Philip. "Word Sense Disambiguation in NLP Applications." Agirre, Eneko and Philip Edmonds (Editors). *Word Sense Disambiguation: Algorithms and Applications*. Springer, 2006.

6. King, M., M. Lipsky, and L. Sharp. "Expert Agreement in Current Procedural Terminology Evaluation and Management Coding." *Archives of Internal Medicine*. 162 (2002): 316-320.

7. Nilsson, G, et al. "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.

8. Stausberg, J., et al. "Comparing Paper-based with Electronic Patient Records: Lessons Learned during a Study on Diagnosis and Procedure Codes" *Journal of the American Medical Informatics Association*.10, no. 5 (Sep-Oct 2003): 470-477.

9. Morsch, M., D. Byrd, and D. Heinze. "Factors in Deploying Automated Tools for Clinical Abstraction and Coding." ITHC (IT in Health Care) Conference, Portland, 13-14, September 2004.

10. van Rijsbergen, C. J. *Information Retrieval.* London: Butterworths, 1979.

11. A useful property of the F-measure is that it incurs a large penalty when P and R are far apart, discouraging attempts to perform well on only one at the expense of the other. Hripcsak and Rothschild (2005) show that κ, another widely used agreement measure, approaches the F-measure when a large number of annotations is assigned by neither annotator.

12. Hripcsak, G. and A. Wilcox. "Reference Standards, Judges, Comparison Subjects: Roles for Experts in Evaluating System Performance." *Journal of the American Medical Informatics Association.* 9 (2002): 1-15.
13. We used a modified voting scheme in which Expert3's codes were compared against Expert1 and Expert2's codes on a note-by-note basis, using backoff through an ordered succession of quality metrics to determine whether there was greater agreement for one of the experts. If so, this constituted a "vote" and broke the tie.
14. Two of the three agreed on all CPT codes 94 percent of the time.
15. Hripcsak and Wilcox (2002) discuss a number of resolution strategies, including breaking ties randomly. We preferred to have a single expert to whom decisions could be attributed in the subset of cases with unresolved disagreement.
16. Gale, W., K. Church, and D. Yarowsky. "Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs," In Proceedings of the 30th Annual Meeting of the ACL, pages 249-256, Newark DE, June 1992.
17. Expressing intrinsic metrics as a percentage of human performance has recent precedent in machine translation, another area where human output has high variability.
18. Morris, W., et al. "Assessing the Accuracy of an Automated Coding System in Emergency Medicine." Proceedings of the AMIA 2000 Annual Symposium. American Medical Informatics Association, November 2000.
19. Nilsson, G, et al. "Evaluation of Three Swedish ICD-10 Primary Care Versions: Reliability and Ease of Use in Diagnostic Coding."
20. Jiang, Yuankai Michael Nossal, and Philip Resnik. "How Does the System Know It's Right? Automated Confidence Assessment for Compliant Coding." AHIMA/FORE Computer-assisted Coding Software Standards Workshop, Arlington, VA, September 2006.
21. Ibid.
22. Morris, W., et al. "Assessing the Accuracy of an Automated Coding System in Emergency Medicine."

References

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

Allen, Jeffrey. "Post-editing." *Computers and Translation: A Translator's Guide.* Harold Somers, Editor. Benjamins Translation Library, 35. Amsterdam: John Benjamins, 2003.

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.

Bird, Steven and Jonathan Harrington, Editors. "Speech Annotation and Corpus Tools." Special issue of *Speech Communication*, 33, no. 1-2, 2001.

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

Hripcsak, G. and A. Rothschild. "Agreement, the F-Measure, and Reliability in Information Retrieval." *Journal of the American Medical Informatics Association*. 12, no. 3 (May-June 2005): 296-298.

Hripcsak, G. and A. Wilcox. "Reference Standards, Judges, Comparison Subjects: Roles for Experts in Evaluating System Performance." *Journal of the American Medical Informatics Association.* 9 (2002): 1-15.

Marcus, Mitchell, Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics*. 19, no. 2 (1993): 313-330.

Webber, Bonnie and Donna Byron. "Proceedings of the ACL 2004 Workshop on Discourse Annotation" Barcelona, Spain, July 2004.

Table 1

Intrinsic Evaluation: Gold Standard Agreement Measures for NLP Engine, Compared with Human Coders and Upper Bounds.

	CPT F	ICD F	ICD-1
Engine	0.838	0.435	0.486
Coders (Paper)	0.872	0.540	0.516
Upper Bound	0.910	0.592	0.572

Figures are aggregated over all engine confidence levels.

Table 2

Extrinsic Evaluation: Production Coders' Agreement with Gold Standard

	CPT F	**ICD F**	**ICD-1**
Paper	0.872	0.540	0.516
NLP	0.897	0.586	0.587

Table 3

Extrinsic Evaluation: Inter-coder Agreement

	CPT F	ICD F	ICD-1
Paper	0.852	0.508	0.470
NLP	0.904	0.616	0.581

Table 4

Extrinsic Evaluation: Intra-coder Agreement

	CPT F	ICD F	ICD-1
Paper	0.875	0.716	0.643
NLP	0.959	0.768	0.675

Table 5

Scenarios Improving Inter-coder Agreement

	CPT F	*gain*	**ICD-F**	*gain*
Baseline	0.852		0.508	
Use expert coders (no CAC)	0.910	*7%*	0.592	*17%*
Use production coders with CAC	0.904	*6%*	0.616	*21%*
	ICD-1	*gain*		
Baseline	0.470			
Use expert coders (no CAC)	0.572	*22%*		
Use production coders with CAC	0.581	*24%*		