Measuring the Information Gain of Diagnosis vs. Diagnosis Category Coding

William R. Hogan, MD, MS[1], Vergil N. Slee, MD, MPH[2]
[1]University of Arkansas for Medical Sciences, Little Rock, AR; [2] President Emeritus, Commission on Professional and Hospital Activities, Columbus, NC

Abstract

Coding categories of diseases, injuries, symptoms, findings, etc. with ICD-9-CM necessarily imparts a loss of information vs. coding such entities with a terminology or ontology—a consequence of the nature of classifications. However, to our knowledge, no one has attempted to quantify this information loss or conversely, the information to be gained by coding entities as opposed to categories. We estimated a lower bound on information gain of coding with SNOMED CT instead of ICD-9-CM, as measured by Shannon's information entropy. We found that the nation could gain more than 97 megabytes of information per year by coding diagnoses vs. diagnosis categories, an increase of 10%. This increase is more than that obtained from coding ICD-9-CM at the 5th instead of the 3rd digit level. We recommend that ICD-9-CM be removed from electronic medical record (EMR) stage 2 and later meaningful use criteria.

Introduction

The coding of *classes* or *categories* of disease, injury, symptom, etc. with the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) necessarily imparts a loss of information relative to coding such entities with a terminology or ontology. This fact is a simple consequence of the nature of ICD-9-CM as a classification, as opposed to nomenclature. For example, the ICD-9-CM category 729.1 is *Myalgia and myositits, unspecified*, and includes numerous diseases such as fibromyalgia, musculoneuralgia, intercostal myalgia, occupational myositis, myofascitis, rheumatoid myositis, and postural myositis. A researcher interested in studying fibromyalgia will retrieve all these diseases when querying 729.1, not just fibromyalgia.

Many authors have discussed the qualitative loss that occurs with ICD-9-CM coding.[1-8] Besides problems with (1) billing considerations distorting coding, (2) semantic drift, (3) the lack of formal version control, (4) the lack of resemblance of ICD-9-CM titles to normal medical terminology, and (5) the errors of coding caused thereby, they note that coding diagnosis categories instead of diagnoses loses essential information about the true conditions from which patients suffer. Restoring that information today all-too-frequently requires either revisiting the paper record—a manual and costly process—or processing narrative text for clinicians' actual diagnoses.

Also, numerous researchers have measured and discussed the accuracy and variability of ICD-9-CM coding.[3, 9-13] For example, Chen et al. measured an accuracy of 74.5% of SNOMED coding for oral diseases vs. 43.6% for ICD-9-CM.[14] The quality of ICD-9-CM data ranges from low for drug-induced liver injury[15] and hyponatremia[16] to high for acute myocardial infarction.[17]

Although ICD-10-CM replaces ICD-9-CM in 2013 in the United States and has higher diagnostic precision, it nevertheless remains a classification and imparts information loss. For example, ICD-10-CM G43.1 *Migraine with aura* includes basilar migraine, classical migraine, retinal migraine, migraine equivalent, and migraine-triggered seizures.

To our knowledge, no one has attempted to *measure* the loss of information when coding with a classification such as ICD-9-CM, or conversely, the information to be gained by coding with a disease terminology or ontology in its place. In this work, we estimate upper and lower bounds on the Shannon information gain of coding diseases with SNOMED CT vs. ICD-9-CM. This estimation requires two key resources provided by the International Health Terminology Standards Development Organization (IHTSDO) and the National Library Medicine (NLM)—the mappings from SNOMED CT to ICD-9-CM and the CORE Problem List Subset of SNOMED CT (CORE Subset), respectively.

Methods

Our measurement of information gain (of coding with SNOMED CT vs. ICD-9-CM) is Shannon's information entropy:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

We represented each ICD-9-CM code—that has more than one SNOMED CT Concept Id (SCTID) mapped to it—as the random variable X. The "outcomes" x_i of X were the SCTIDs mapped to the ICD-9-CM. For the probability of each outcome $p(x_i)$, we used the column labeled 'usage' in the CORE Subset, which is the average frequency with which an SCTID has been used at seven, large, medical institutions across the United States.[18] The usage is a percentage:

we divided it by 100 to obtain a probability. For each such ICD-9-CM code, we normalized the probabilities to one. Finally, we used b=2 (i.e., base 2 logarithm) and thus our unit of measurement was binary digits or bits, which we converted to larger units for very large values (e.g., 1 megabyte = 8,000,000 bits).

This information is the information required to distinguish among the entities that ICD-9-CM groups into classes, per Shannon's metric. For example, one bit of information is necessary to distinguish between two equally probable (and mutually exclusive) diagnoses when the diagnosis category is known. It is unrelated to the lengths of identifiers or codes used by ICD-9-CM or SNOMED CT. Note that when the probabilities of the diagnoses in an ICD-9-CM category are not equal, then it is possible for 1 bit to distinguish among more than two diagnoses. For example, the code 729.5 has 12 SCTIDs mapped to it, but only 2.13 bits are required to distinguish among the 12 due to the varying probabilities of each SCTID. Were the 12 SCTIDs equally likely, then 3.58 bits would have been necessary.

Because not every SCTID that is mapped to an ICD-9-CM appears in the CORE Subset, we used two values of p for them, which result in upper and lower bounds on Shannon entropy. Specifically, for the upper bound, we assumed that any code not in the CORE Subset had a maximum usage equivalent to the minimum usage value over the entire CORE Subset. The reason was that if one of these SCTIDs had a usage greater than the CORE Subset minimum, it would have appeared there. Because the $p \log(p)$ term of the formula for entropy monotonically approaches 0 as $p \rightarrow 0$, any value smaller than this maximum for p computes a smaller value of entropy. Thus, using p = minimum CORE Subset usage/100 results in the maximum possible information gain. For the lower limit on information gain, we used p=0 (using $p \log(p) = 0$ when p=0, as is customary).

We ignored the 78 mappings (out of 2,241 total) from one SCTID to two ICD-9-CM codes with a map advice of '1' or '2' in this analysis. In addition, the minimum and maximum values of information gain are uninformative for any ICD-9-CM code none of whose mapped SCTIDs has a usage value, because they are the theoretical minimum and maximum; we excluded all such ICD-9-CM codes. The large number of such codes (1,306) also would have an extreme effect on the aggregate statistics of information gain over all ICD-9-CM codes.

We used the July, 2009 version of SNOMED CT, but used the ICD-9-CM mapping files dated Oct 31, 2009 (e.g., sct_crossmaps_icd9_20091031.txt). These mappings reflect the most recent version of ICD-9-CM, dated Oct 1, 2009. For display purposes, we used the short ICD-9-CM titles available for download from the Centers for Medicare and Medicaid Services web site.[19] We used the November, 2009 CORE Problem List Subset of SNOMED CT based on the July, 2009 Release of SNOMED CT. The minimum value of usage from this version of the CORE Subset was 0.0003.

We downloaded the materials from the UMLSKS web site, imported them into Microsoft Access tables, joined the mapping and CORE Subset tables together on the SCTID, and then exported a file where each line contained a SCTID to ICD-9-CM mapping (SCTID, ICD-9-CM code, and map advice) with the usage from the Core Subset (or null if not present). We included all mappings with a map advice of '2', meaning that the SCTID is "narrower in meaning" than the ICD-9-CM code. We excluded mappings with an advice of '1' (indicating synonymy) because there is no information loss if there is true synonymy. However, if two different SCTIDs were mapped to one ICD-9-CM with an advice of '1', then we assumed that the SCTIDs were not synonymous with either each other or the ICD-9-CM code and included them in the analysis

We report summary statistics (mean, median, standard deviation, etc) of the minimum and maximum information gain over all ICD-9-CM codes that met these inclusion criteria. We also list the top ten ICD-9-CM codes by minimum information gain as well as the information gain for several, commonly-used ICD-9-CM codes.

This procedure makes several assumptions: (1) none of the SNOMED SCTIDs mapped to a particular ICD-9-CM is an ancestor or descendant of another mapped SCTID, (2) no SCTID is mapped to more than one ICD-9-CM code, and (3) the SCTIDs are mutually exclusive and exhaustive (i.e., no two SCTIDs occur together that stand in an ancestor/descendant relationship).

Finally, we used data from the National Hospital Discharge Survey (NHDS), National Hospital Ambulatory Medical Care Survey (NHAMCS), and National Ambulatory Medical Care Survey (NAMCS) to compute an estimate of the national information gain of coding diagnoses with SNOMED-CT. For each ICD-9-CM included in this study, we multiplied its information gain by the record weight for every record where it appeared, and summed this product over all records. We used the latest available surveys: NHDS 2006, NHAMCS 2007, and NAMCS 2007. Each survey weights its

records to adjust data for various biases in sampling and response.

For a basis of comparison for the gain from ICD-9-CM to SNOMED CT, we also measured the total information of ICD-9-CM in the surveys and the gain of full vs. truncated-at-3-digit ICD-9-CM coding. This measurement involves treating the "diagnosis fields" in the survey data as one field and thus the random variable X in Shannon's formula, and each ICD-9-CM code is an outcome x_i of this variable. For the probability of each ICD-9-CM code, we summed the weights of all the diagnosis records in which it appeared and divided this sum by the total number of diagnoses calculated similarly. For this measurement, we only included the ICD-9-CM codes that met the inclusion criteria.

Results

Overall, 857 ICD-9-CM codes met our inclusion criteria, which is just over 6% of all ICD-9-CM codes. There was a mean of 21.6 SCTIDs mapped to each ICD-9-CM (range: 1–682, std dev: 40.6, median: 12).

With respect to minimum information gain of coding with SNOMED CT vs. ICD-9-CM (or conversely, the loss of information by coding with ICD-9-CM in place of SNOMED CT), the mean was 0.82 bits (range: 0–2.57, std dev: 0.50, median: 0.86). Of the 857 ICD-9-CM codes, 71 had just one SCTID with a usage value from the CORE Subset. This situation led to a minimum loss of 0, as the lone SCTID has a probability of 1 due to normalization.

With respect to maximum information gain of SNOMED CT vs. ICD-9-CM coding, the mean was 1.92 bits (range: 0–9.03, std dev: 1.47, median: 1.62).

The ICD-9-CM codes with the highest loss relative to SNOMED CT were 282.49 *Thalassemia* at 2.57 bits, 709.8 *Skin disorders* at 2.47 bits, and 239.0 *Digestive neoplasm* at 2.44 bits (Table 1).

Table 1. Top Ten ICD-9-CM by Information Gain.

ICD-9-CM	Short Title	Min gain (bits)
282.49	Thalassemia NEC	2.57
709.8	Skin disorders NEC	2.47
239.0	Digestive neoplasm NOS	2.44
727.05	Tenosynov hand/wrist NEC	2.34
686.9	Local skin infection NOS	2.32
729.5	Pain in limb	2.31
709.09	Other dyschromia	2.24
576.1	Cholangitis	2.23
V12.5	Hx-circulatory dis NOS	2.17
478.19	Nasal & sinus dis NEC	2.14

The gain for selected ICD-9-CM codes for common diagnoses ranged from 0.035 bits for *401.9 Hypertension* to 1.48 bits for *250.00 Type 2 DM w/o complic* (Table 2).

From the three surveys, the total estimated number of diagnoses per year in the United States was 2,682,392,162. The 857 ICD-9-CM codes accounted for 39.4% of these (1,056,784,292).

For the 857 ICD-9-CM codes, the total, estimated minimum information gain (of coding with SNOMED CT vs. ICD-9-CM) from the three national surveys was 97.7 megabytes (MB). The gain estimated from the NHDS was 6.90 MB; from NAMCS 80.7 MB; and from NHAMCS 10.1 MB. The estimated maximum gain was 159 MB.

The total information of the 857 ICD-9-CM diagnosis categories in the surveys was 962MB, so the gain of 97.7MB represents a gain of just over 10%. The total information of the 857 categories at the truncated 3-digit level was 862MB. Thus, the gain of 97MB of SNOMED-CT vs. full ICD-9-CM coding exceeds the gain of 92MB of full vs. truncated-at-3-digit ICD-9-CM coding.

Table 2. Information Gain of Selected, Commonly-used ICD-9-CM.

ICD-9-CM	Short Title	Min gain (bits)
401.9	Hypertension NOS	0.035
250.00	Type 2 DM w/o complic	1.48
715.90	Osteoarthros NOS-unspec	0.24
536.8	Stomach function dis NEC	0.27
493.90	Asthma NOS	1.24
244.9	Hypothyroidism NOS	0.22
465.9	Acute uri NOS	0.23
307.81	Tension headache	0.24
305.1	Tobacco use disorder	1.46
486	Pneumonia, organism NOS	0.22

Discussion

We estimated a minimum, annual gain of diagnostic information of 97.7 MB were the nation to code diagnoses with SNOMED CT in place of ICD-9-CM. This represents a 10% increase in information, and exceeds the information gain of full vs. 3-digit ICD-9-CM coding. This information gain is large.

This estimate is based on only the 857 ICD-9-CM codes for which we could measure gain. The overall loss of information with ICD-9-CM is likely to be higher. These codes represent ~6% of all codes, but 39% of diagnosis records in the surveys. Based on the assumptions of this study, it would be difficult to extrapolate to all of ICD-9-CM.

Given the magnitude of the gain, it is worth reconsidering the final rule on meaningful use of EMRs. The American Recovery and Reinvestment Act of 2009 allows the Centers for Medicare & Medicaid Services (CMS) to incent eligible hospitals and healthcare providers to become "meaningful users" of certified EMRs. The Secretary of Heatlh and Human Services is tasked with developing a regulation that defines 'meaningful use'. The final rule defining meaningful use became available on July 13, 2010. This rule states that either ICD-9-CM or SNOMED CT is acceptable for maintaining problem lists. Given the information to be gained and its importance for patient care as well as numerous secondary uses of EMR data, we recommend that ICD-9-CM be dropped from Stage 2 and subsequent meaningful use criteria for EMRs. Delaying the planned conversion to ICD-10-CM may also be worthwhile.

The information gained—according to Shannon entropy—is that which is necessary to distinguish among the entities that ICD-9-CM groups into categories. Note that although 2 bits is *necessary* to differentiate among four, equally probable diagnoses, it is also *sufficient* to differentiate among 10 or even 100+ diagnoses of widely varying probabilities.

Our study did not estimate the information gain of clinicians' actual diagnoses relative to ICD-9-CM or to SNOMED CT. We used SNOMED CT because (1) there was available probability information from the CORE Subset, (2) it is close to clinicians' language, and (3) the resources we used were freely available. We note that post-coordination of SNOMED CT codes is often required to capture clinicians' expressions fully,[20] and thus there is also likely a loss of information with pre-coordinated SNOMED CT codes.

The limitations of this study include (1) it is subject to various criticisms of Shannon entropy, (2) it makes assumptions including exclusivity and exhaustiveness of SCTIDs mapped to a single ICD-9-CM, (3) not all SCTIDs are mapped to ICD-9-CM nor do all ICD-9-CM codes have an SCTID mapping, (4) it ignores "broad to narrow" SNOMED CT to ICD-9-CM mappings, (5) it ignores the small number of mappings from an SCTID to >1 ICD-9-CM codes, (6) the set of SNOMED CT to ICD-9-CM mappings may not be complete, and (7) usage values were not available for all mapped SCTIDs.

The main criticism of Shannon's information entropy is that it does not account for the context, meaning, utility, or truthfulness of information.[21] However, information about actual diagnoses vs. diagnosis categories is likely of high utility for numerous secondary uses of data, and is certainly essential to patient care.

The limitations of incomplete mappings and lack of usage values (#6 and #7 above) lead to underestimation of the information loss incurred by coding with ICD-9-CM. Were there more SCTIDs mapped to the 857 ICD-9-CM codes; or multiple SCTIDs mapped to ICD-9-CM codes not included in this study, the information loss per ICD-9-CM code, and per usage of ICD-9-CM code, would be higher.

Future work includes addressing the limitations of this study. In particular, it will be important to (1) relax the assumption that diseases in a category do not co-occur, and (2) find or generate sources of data with 'usage' information about actual diagnoses.

Conclusion

The United States has the potential to gain substantial information about patients' diagnoses by coding individual diagnoses vs. diagnosis categories. Meaningful use criteria for EMRs should drop provisions for ICD-9-CM and perhaps ICD-10-CM.

Acknowledgements

This work was supported by award numbers 1UL1RR029884 and 3 P20 RR016460-08S1 from the National Center for Research Resources. The content is solely the responsibility of the author and does not necessarily represent the official views of the NCRR or NIH.

References

1. Slee VN, Slee D, Schmidt HJ. The tyranny of the diagnosis code. *N C Med J.* Sep-Oct 2005;66(5):331-337.
2. Slee VN, Slee DA, Schmidt HJ. *The endangered medical record : ensuring its integrity in the age of informatics*. St. Paul: Tringa Press; 2000.
3. Hsia DC, Krushat WM, Fagan AB, et al. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *N Engl J Med.* Feb 11 1988;318(6):352-5.
4. Cimino JJ. Formal descriptions and adaptive mechanisms for changes in controlled medical vocabularies. *Methods Inf Med.* Sep 1996;35(3):202-210.
5. Vardy DA, Gill RP, Israeli A. Coding medical information: classification versus nomenclature and implications to the Israeli medical system. *J Med Syst.* Aug 1998;22(4):203-210.
6. Feinstein AR. ICD, POR, and DRG. Unsolved scientific problems in the nosology of clinical medicine. *Arch Intern Med.* Oct 1988;148(10):2269-2274.

7. Cimino JJ. An approach to coping with the annual changes in ICD9-CM. *Methods Inf Med.* Sep 1996;35(3):220.
8. deBronkart D. e-Patients Can Help. Let Us. Our Families' Lives Are At Stake. Testimony to HIT Policy Committee 2010: http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_11673_910712_0_0_18/2DeBronkart_testimony022510.pdf. Accessed Mar 1, 2010.
9. Dixon J, Sanderson C, Elliott P, et al. Assessment of the reproducibility of clinical coding in routinely collected hospital activity data. *J Public Health Med.* Mar 1998;20(1):63-69.
10. Fisher ES, Whaley FS, Krushat WM, et al. The accuracy of Medicare's hospital claims data: progress has been made, but problems remain. *Am J Public Health.* Feb 1992;82(2):243-248.
11. Hsia DC. Diagnosis related group coding accuracy of the peer review organizations. *J AHIMA.* Sep 1992;63(9):56-64.
12. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res.* Oct 2005;40(5 Pt 2):1620-1639.
13. Surjan G. Questions on validity of International Classification of Diseases-coded diagnoses. *Int J Med Inform.* May 1999;54(2):77-95.
14. Chen JW, Flaitz C, Johnson T. Comparison of accuracy captured by different controlled languages in oral pathology diagnoses. *AMIA Annu Symp Proc.* 2005:918.
15. Jinjuvadia K, Kwan W, Fontana RJ. Searching for a needle in a haystack: Use of ICD-9-CM codes in drug-induced liver injury. *Am J Gastroenterology.* 2007;102(11):2437-2443.
16. Shea A, Curtis L, Szczech L, Schulman K. Sensitivity of International Classification of Diseases codes for hyponatremia among commercially insured outpatients in the United States. *BMC Nephrology.* 2008;9(1):5.
17. Kiyota Y, Schneeweiss S, Glynn RJ, et al. Accuracy of Medicare claims-based diagnosis of acute myocardial infarction: estimating positive predictive value on the basis of review of hospital records. *Am Heart J.* Jul 2004;148(1):99-104.
18. The CORE Problem List Subset of SNOMED CT. 2010; http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html. Accessed Mar 10, 2010.
19. Diagnosis and procedure codes: Abbreviated and full code titles. 2009; http://www.cms.hhs.gov/ICD9ProviderDiagnosticCodes/06_codes.asp#TopOfPage. Accessed Mar 8, 2010.
20. Wade G, Rosenbloom ST. Experiences mapping a legacy interface terminology to SNOMED CT. *BMC Med Inform Decis Mak.* 2008;8 Suppl 1:S3.
21. Blois MS. *Information and medicine: The nature of medical descriptions*. Berkley: University of California Press; 1984.