DOI: 10.1111/j.1475-6773.2005.00444.x

Measuring Diagnoses: ICD Code Accuracy

Kimberly J. O'Malley, Karon F. Cook, Matt D. Price, Kimberly Raiford Wildes, John F. Hurdle, and Carol M. Ashton

Objective. To examine potential sources of errors at each step of the described inpatient International Classification of Diseases (ICD) coding process.

Data Sources/Study Setting. The use of disease codes from the ICD has expanded from classifying morbidity and mortality information for statistical purposes to diverse sets of applications in research, health care policy, and health care finance. By describing a brief history of ICD coding, detailing the process for assigning codes, identifying where errors can be introduced into the process, and reviewing methods for examining code accuracy, we help code users more systematically evaluate code accuracy for their particular applications.

Study Design/Methods. We summarize the inpatient ICD diagnostic coding process from patient admission to diagnostic code assignment. We examine potential sources of errors at each step and offer code users a tool for systematically evaluating code accuracy.

Principle Findings. Main error sources along the "patient trajectory" include amount and quality of information at admission, communication among patients and providers, the clinician's knowledge and experience with the illness, and the clinician's attention to detail. Main error sources along the "paper trail" include variance in the electronic and written records, coder training and experience, facility quality-control efforts, and unintentional and intentional coder errors, such as misspecification, unbundling, and upcoding.

Conclusions. By clearly specifying the code assignment process and heightening their awareness of potential error sources, code users can better evaluate the applicability and limitations of codes for their particular situations. ICD codes can then be used in the most appropriate ways.

Key Words. ICD codes, accuracy, error sources

Nosology (the systematic classification of diseases) has always fascinated the sick and their would-be healers. Western societies developed an interest in nosology in the seventeenth and eighteenth centuries when they began to track the causes of sickness and death among their citizens. In the twentieth

century, when medical insurance programs made payers other than patients responsible for medical care, nosology became a matter of great interest to those public and private payers. The most commonly used nosologies include International Classification of Diseases (ICD), the American Medical Association's Current Procedural Terminology, 4th Edition (CPT-4); the Health Care Financing Administration (HCFA, now known as the Centers for Medicare and Medicaid Services) Health Care Common Procedural Coding System (HCPCS); the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, 4th Revision (DSM-IV); Europe's Classification of Surgical Operations and Procedures, 4th Revision (OPCS-4); and the Agency for Healthcare Research and Quality's Clinical Classification Software (CCS).

This paper focuses on the International Classification of Diseases, now in its ninth and soon to be tenth iteration; the most widely used classification of diseases. Beginning in 1900 with the ICD-1 version, this nosology has evolved from 179 to over 120,000 total codes in ICD-10-CM (ICD-10 2003; ICD-10-CM 2003). The use of codes has expanded from classifying morbidity and mortality information for statistical purposes to diverse sets of applications, including reimbursement, administration, epidemiology, and health services research. Since October 1 1983, when Medicare's Prospective Payment System (PPS) was enacted, diagnosis-related groups (DRGs) based on ICD codes emerged as the basis for hospital reimbursement for acute-care stays of Medicare beneficiaries (U.S. Congress 1985). Today the use of ICD coding for reimbursement is a vital part of health care operations. Health care facilities use ICD codes for workload and length-of-stay tracking as well as to assess quality of care. The Veterans Health Administration uses ICD codes to set capitation rates and allocate resources to medical centers caring for its 6 million beneficiaries. Medical research uses ICD codes for many purposes. By grouping patients according to their diagnoses, clinical epidemiologists use ICD codes to study patterns of disease, patterns of care, and outcomes

---

Address correspondence to Kimberly J. O'Malley, Ph.D., Pearson Educational Measurement, 2201 Donley Drive, Suite 195, Austin, TX 78758. Karon F. Cook, Ph.D., Associate Director for Research, is with the Parkinson's Disease Research, Education and Clinical Center, Michael E. DeBakey VA Medical Center, Houston, TX. Matt D. Price, M.S., R.H.I.A., Kimberly Raiford Wildes, Dr.P.H., M.A., and Carol M. Ashton, M.D., M.P.H., are with the Houston Center for Quality of Care and Utilization Studies, Houston VA Medical Center, Baylor College of Medicine, Department of Medicine, Section of Health Services Research, Michael E. DeBakey VA Medical Center, Houston, TX. John F. Hurdle, M.D., Ph.D., is with the VA Salt Lake City Health Care System, Salt Lake City, UT.

of disease. Health services researchers use the codes to study risk-adjusted, cross-sectional, and temporal variations in access to care, quality of care, costs of care, and effectiveness of care. Medical and health services researchers commonly use ICD codes as inclusion and exclusion criteria to define sampling frames, to document the comorbidities of patients, report the incidence of complications, track utilization rates, and determine the case fatality and morbidity rates (see Calle et al. 2003 for a recent example) (Steinman, Landefeld, and Gonzales 2000; Calle et al. 2003; Charbonneau et al. 2003; Jackson et al. 2003; Martin et al. 2003; Studdert and Gresenz 2003). The widespread and diverse use of ICD codes demonstrates the central role nosology plays in health care.

Increased attention to code accuracy has occurred both as a result of the application of ICD codes for purposes other than those for which the classifications were originally designed as well as because of the widespread use for making important funding, clinical, and research decisions. Code accuracy, defined as the extent to which the ICD nosologic code reflects the underlying patient's disease, directly impacts the quality of decisions that are based on codes, and therefore code accuracy is of great importance to code users. Accuracy is a complicated issue, however, as it influences each code application differently. Using the codes for reporting case fatality rates in persons hospitalized for influenza, for example, might require a different level of accuracy than using codes as the basis for reimbursing hospitals for providing expensive surgical services to insured persons. Therefore, users of disease classifications, just as users of any measure, must consider the accuracy of the classifications within their unique situations. An appreciation of the measurement context in which disease classifications take place will improve the accuracy of those classifications and will strengthen research and health care decisions based on those classifications.

Researchers studying errors in the code assignment process have reported a wide range of errors. Studies in the 1970s found substantial errors in diagnostic and procedure coding. These error rates ranged from 20 to 80 percent (Institute of Medicine 1977; Corn 1981; Doremus and Michenzi 1983; Johnson and Appel 1984; Hsia et al. 1988). Studies in the 1980s reported slightly increased accuracy with average error rates around 20 percent, and most below 50 percent (Lloyd and Rissing 1985; Fischer et al. 1992; Jolis et al. 1993). Studies in the 1990s found rates similar to those of the 1980 studies, with error rates ranging from 0 to 70 percent (Benesch et al. 1997; Faciszewski, Broste, and Fardon 1997; Goldstein 1998). The inconsistency in the error rates and wide range of reported amounts of error is due largely to

differences across study methods (i.e., different data sets, versions of the ICD classifications, conditions studied, number of digits compared, codes examined, etc.) (Bossuyt et al. 2004). However, variation in error rates is also influenced by the many different sources of errors that influence code accuracy (Green and Wintfeld 1993). By clearly specifying the code process and the types of errors and coding inconsistencies that occur in each study, researchers can begin to understand which errors are most common and most important in their situation. They can then institute steps for reducing those errors.

If we think of the assignment of ICD codes as a common measurement process, then the person's true disease and the assigned ICD code represent true and observed variables, respectively. One approach to evaluating ICD code accuracy is to examine sources of errors that lead to the assignment of a diagnostic code that is not a fair representation of the patient's actual condition. Errors that differentiate the ICD code from the true disease include both random and systematic measurement errors. By understanding these sources of error, users can evaluate the limitations of the classifications and make better decisions based on them. In this manuscript, we (1) present the history of ICD code use, (2) summarize the general inpatient ICD coding process (from patient admission to the assignment of diagnostic codes), (3) identify potential sources of errors in the process, and (4) critique methods for assessing these errors.

BACKGROUND

History of ICD Codes

In 1893, the French physician Jacques Bertillon introduced the Bertillon Classification of Causes of Death. This first edition, had 179 causes of death. It was recommended that this classification system, subsequently known as the International Classification of Causes of Death (ICD), be revised every 10 years. With each revision, the numbers of codes increased, as did the appeal of using them for other purposes.

The World Health Organization (WHO) published the ninth revision of ICD in 1978. The ICD-9 is used to code mortality data from death certificates. To make the ICD more useful for American hospitals, the U.S. Public Health Service modified ICD-9 and called it the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). In the Clinical Modification (CM) of ICD-9, codes were intended to be more precise than those needed only for statistical groupings and trend analysis (The International Classification of Diseases, 9th Revision, Clinical Modification [ICD-9-CM], Sixth Edition 2002).

The modified version expanded to three volumes, and a fifth-digit subclassification was introduced. The fifth digit adds increased specificity and is required when available. When identifying burns with codes 941 through 954, for example, a fourth digit indicates the depth of the burn and a fifth digit specifies the exact location. ICD-9-CM is used in the United States to code and classify diagnoses and procedures from inpatient and outpatient records, as well as inpatient procedures. The ninth revision contains over 12,000 diagnostic and 3,500 procedure codes.

The development of a tenth revision, International Statistical Classification of Diseases and Related Health Problems (ICD-10), introduces alphanumeric codes and greater specificity than ICD-9, and includes over 21,800 total codes (ICD-10 2003). In January 1999, the United States began using ICD-10 to code and classify mortality data from death certificates; however, at the time of this writing, the clinical modification version (ICD-10-CM), was still under development by the National Center for Health Statistics (NCHS) and has not been released.

The Process for Assigning ICD Codes

In this paper, we focus on codes assigned during hospital stays, and whereas basic coding procedures are similar across ambulatory settings, the reader is cautioned that some errors may be setting-specific. The basic process for assigning ICD codes for inpatient stays, presented in Figure 1, can be conceptualized as the dynamic interplay between the patient as he or she progresses through the health care system (called the "patient trajectory") and the creation of the medical record (called the "paper trail"). The basic process shown in Figure 1 is typical, even though details of any given step at any given facility may vary. The left side of Figure 1 portrays the patient trajectory through the system, from admission to discharge. The right side of Figure 1 represents the paper trail, or medical record creation, from the recording of the admitting diagnosis to the assignment of the ICD codes after discharge.

The patient trajectory starts when the patient arrives at the hospital, at which time some type of precertification (insurance) based on the admitting diagnosis is performed by the admission clerk (at least for insured patients). After admission, based on the physician's admitting diagnosis and the information generated by the initial workup, the patient undergoes diagnostic tests and procedures and/or other treatment, as ordered by the medical staff. The patient and medical staff members continue to meet throughout the hospital stay to exchange information, and additional tests, procedures, and/or

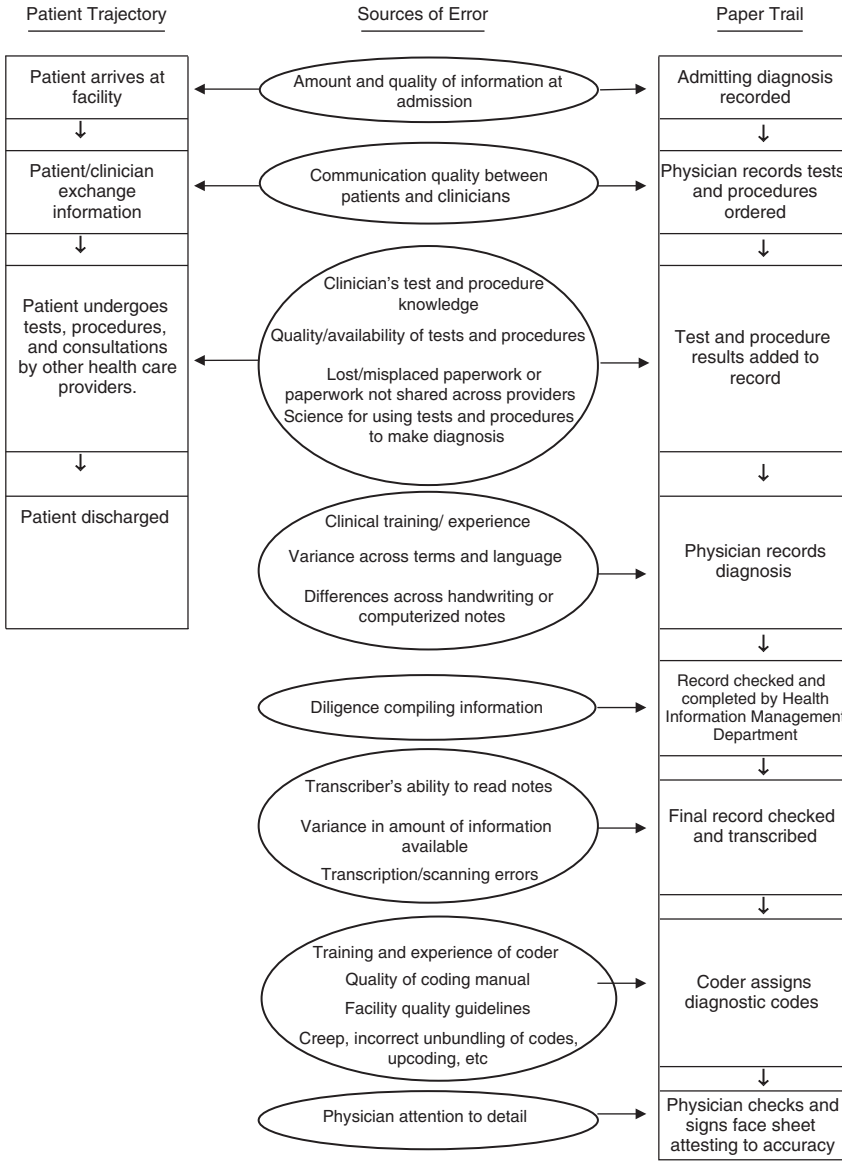

Figure 1: Overview of the Inpatient Coding Process.

treatments may be ordered. Test and procedure results are added to the medical record. The results from the tests and procedures often result in

changes in the admitting diagnosis. Furthermore, complications arising from care may also add to the list of diagnoses. The staff documents the hospital stay using either handwritten or electronic reporting. Upon discharge, the physician completes a narrative discharge summary that includes a list of primary and secondary diagnoses (word labels) and describes follow-up plans.

Upon discharge, the patient's medical record and all associated documentation are transferred to the medical record or health information management department. Concurrently, technicians check to ensure that all medical record information is accurate and complete (including the face sheet, history and physical, operative reports, radiology reports, physician's orders, progress and nursing notes, consultations, discharge summary, etc.). Coders then begin the process of classifying documentation, including diagnoses and procedures, using rigid ICD coding guidelines and conventions. Some facilities outsource medical transcription and coding.

After reviewing all pertinent medical record information, medical coders assign a code for the principal diagnosis, defined by the Uniform Hospital Discharge Data Set (UHDDS) as "that condition established after study to be chiefly responsible for occasioning the admission of the patient to the hospital care" (Uniform Hospital Discharge Data Set 1985). The principal diagnosis assignment is made based on written documentation from the providers. Coders also assign a code for the principal procedure, or one performed for definitive treatment or that was necessary for treating a complication. They assign additional diagnostic codes (the code count being determined by the facility) for diagnoses that require clinical evaluation, therapeutic interventions, diagnostic procedures, extended lengths of stay (for inpatient stays), or increased nursing care and/or monitoring. Additional procedures are coded as well. The VA, for example, allows up to 10 diagnoses and five procedure codes per inpatient day (Department of Veterans Affairs 2002). Coders may also assign V-codes (codes describing conditions that coexist during a patient's stay that influence the stay, such as history of cancer or lack of housing), E-codes (supplementary classification of factors influencing the patient's health status and contact with health services), and M-Codes (supplementary classification of the morphology of neoplasms).

After the code assignments and the sequencing of the codes have been determined, a computerized software program, called a grouper, is often used to classify or group the codes for reimbursement purposes. When the coding process is complete, the codes are transmitted to the billing department for reimbursement purposes.

Examining Coding Errors

Sources of Errors

Many sources of error are interposed between a person's disease (as it is in truth) and the word label (the diagnosis) applied to it by a clinician, and between the diagnosis and the nosologic code applied to it by a medical coder. A summary of these errors, organized according to the patient trajectory and the paper trail depicted in Figure 1.

Errors Along the Patient Trajectory

A "diagnosis" is a word label applied to the disordered anatomy and physiology (the disease) presumed to be causing a person's constellation of symptoms and signs. Several sources of error influence the diagnostic process from patient admission to discharge. "Diagnosis, is in the end, an expression of probability" (author of quotation is unknown). Hardly any diagnosis—even one made at autopsy—is certain (Ornelas-Aguirre et al. 2003; Silfvast et al. 2003). The certainty, or accuracy, of a diagnosis depends upon multiple factors such as the participants (e.g., patient, clinician, medical staff), disease type, current state of medical knowledge and technology, context within which the diagnosis is made, and translation of coding changes into practice.

The first potential sources of error along the trajectory in Figure 1 relate to communication. The quality and quantity of communication between the patient and his or her admitting clerk and treating clinicians are critical determinants of the accuracy of the admitting diagnosis. If the patient describes only a subset of symptoms or withholds important information, for example, the accuracy of the diagnosis may be compromised, leading the clinician down the wrong line of diagnostic reasoning. Clinicians can be poor communicators as well. The clinician may fail to ask the right questions, may not fully elicit the patient's history, or may misunderstand the patient, resulting in another source of error in the process.

When the clinician orders diagnostic tests and procedures, potential sources of error that may be introduced include the clinician's knowledge about the best diagnostic tests and procedures, the availability of these tests and procedures, and the clinician's ability to interpret the results. Diagnostic accuracy depends upon the state of scientific understanding regarding various presentations and etiologies of the disease. Further, the utility (sensitivity, specificity, and predictive value) of the tests and procedures available to the clinician impacts the certainty of the diagnosis. A disease for which tests have high sensitivity and specificity will result in higher diagnostic accuracy

compared with a disease with vague manifestations and poor diagnostic tests. The accuracy of cancer diagnoses, for example, is typically higher than of schizophrenia diagnoses, in part because tumor histopathology and serum markers are less ambiguous than the behavioral diagnostic criteria for schizophrenia. Errors also occur when the physician records the diagnosis. Variance in the clinician's description of the diagnosis—often 5–10 synonyms exist for the same clinical entity—and clarity in the recording of the diagnosis, especially if handwritten, also introduce error into the coding process. Clinicians are notorious for undecipherable handwriting.

Consider how errors in the patient trajectory may influence the diagnosis of a patient with a stroke (a disruption of blood supply to the brain). The warning symptom for stroke, a transient ischemic attack (TIA), is a set of transitory neurological symptoms (in medical parlance, symptoms are subjective and not directly observable by others) and/or signs (signs are observable by others) thought to result from a temporary interference with arterial circulation to a discrete part of the brain. Some TIAs are over in seconds; by definition, all TIAs resolve within 24 hours or they are given a different label (Johnston et al. 2003). The signs and symptoms of TIA are nonspecific; that is, they can result from several other conditions besides a temporary interruption of blood flow to a part of the brain. Because the symptoms are nonspecific, the patient at admission might choose to share only a few (e.g., the patient reports headache but not dizziness), or the patient might not notice the more subtle symptoms (e.g., subtle visual field disturbances) and therefore not share them with the clinician. During the patient–clinician interaction, the clinician might make a decision based only on the symptoms reported by the patient and only on the most obvious signs. Furthermore, no blood or imaging test at present can confirm or disconfirm the occurrence of a TIA. Therefore, the diagnosis of TIA rests on a clinician's acumen, and acumen depends on training, experience, attention, thoroughness, and the ability to elicit information from the patient and/or available informants. Consequently, the interrater reliability of the diagnosis of TIA is very low, such as κ values just over 0.40 and overall agreement of 57 percent (Dewey et al. 1999; Goldstein et al. 2001; Wilson et al. 2002). Further, new diagnostic criteria for TIA have recently been proposed (Albers et al. 2002). If adopted, these criteria will compound the difficulty in monitoring the incidence and prevalence of TIAs over time.

As criteria for the diagnosis of diseases are constantly in flux because of the evolving nature of medical knowledge, new types of errors (or coding inconsistencies) are introduced into the process, and other errors may decrease as diagnostic accuracy increases. New errors may evolve from

clinicians' delay in learning about medical advances or new diagnostic tools. This is especially true for conditions for which no laboratory or imaging tests are available. A case in point is mental illness, the diagnosis of which is based on the DSM published by the American Psychiatric Association, now in its fourth major iteration (American Psychiatric Association 1994). Consider, for example, how "homosexuality" as a diagnosis evolved from being a nondiagnosis to a mental illness diagnosis to a "life style."

Even when laboratory or imaging tests are available for confirming or ruling out a diagnosis, medical technology evolves and the tests improve. For example, the diagnosis of acute stroke no longer requires a spinal tap and direct arteriography of cerebral vessels; the diagnosis can now be made noninvasively with a magnetic resonance imaging study of the brain (Provenzale et al. 2003).

Errors Along the Paper Trail

The first three error sources, those of communication and those related to tests and procedures the patient undergoes, that are listed in Figure 1 affecting the paper trail were described in the previous section. Another potential set of errors can be found in the record itself. Clinicians do not generally assign codes; coders assign them based on the labels recorded by clinicians in charts or on death certificates. Errors in this phase have been reported to range from 17.1 to 76.9 percent (Hsia et al. 1988). The variability in the error rates is best understood by considering the measurement contexts in which the code assignments take place.

One potential set of errors can be found in the record itself. In Figure 1, these are the fourth set of errors listed. In their written or electronically entered record, clinicians often use synonyms and abbreviations to describe the same condition. For example, synonyms for "stroke" include cerebrovascular accident, cerebral occlusion, cerebral infarction, and apoplexy, among others. The variance in terms is problematic, as each diagnostic code should represent one and only one disease entity. From the clinician's recorded diagnosis label, the coder must select the ICD code that best seems to match the clinician's terminology. The use of synonyms leads to imprecision. For example, a patient who had a stroke can be described by one doctor as having had an intracerebral hemorrhage (Code 431) and by another doctor as having had a cerebrovascular accident (Code 436) and both doctors would be technically correct. Errors can also occur because of physicians' and other staff's omissions in the medical record. In 1985, a study of 1,829 medical records in the Veterans Administration indicated that over 40 percent of physician errors

were attributed to omissions (Lloyd and Rissing 1985). Another source of recorder error is in the transcription of the medical record. Transcription can be defined as the process of converting medical record information from voice (dictation) to hardcopy report or electronic format. Transcription or scanning errors are additional threats to code accuracy. The extent to which the chart information is complete influences the accuracy of the codes as well. Concurrent coding, coding that is performed before patient discharge, is implemented in some facilities in order to expedite the coding and billing processes. With the considerable pressure on hospitals to discharge patients these days (at least, in the U.S.), coders may have incomplete clinical information when they receive the chart. As a result, coders are required to assign codes with varying amounts of information, which impacts code accuracy.

Many potential errors originate with the coder. The fact that coders must pour through sometimes voluminous records to extract diagnoses can lead to several types of errors. One study examining coding variation found that when 11 experienced, active medical coders reviewed 471 medical records and were told they would be reevaluated, all of the coders differed in one or more data fields for more than half of the records (Lloyd and Rissing 1985). The adequacy of training the coder receives influences her or his ability to synthesize large amounts of information and assign precise codes. The American Health Information Management Association (AHIMA), the governing body for health information professionals, designates two types of certification: a 2-year certification (Registered Health Information Technician, previously Accredited Record Technician) and a 4-year certification (Registered Health Information Administrator, previously Registered Record Administrator). The existence of two levels of certification, based on length of academic programs and course content, may contribute to coding inconsistencies. In addition to these two professional credentials, AHIMA also offers multiple coder certification opportunities and credentials. Continuing education of coders, or lack thereof, also influences coding accuracy, as the codes and coding rules expand and change annually. For example, on October 1 2003, a new ICD-9 procedure code (00.15) was created to identify patients who receive high dose interleukin-2 treatment. As another example, a comparison of ICD-9-CM with ICD-10-CM indicates that the number of categories doubled from 4,000 to 8,000 and the number of death causes increased from 72 to 113 (Colorado Department of Public Health and Environment 2001). Changes in codes include reclassification of codes, such as moving of hemorrhage from the "circulatory" chapter to the "signs and symptoms" chapter, and changing of the four-digit numeric codes of ICD-9 to the four-digit alphanumeric codes of

ICD-10. Diabetes mellitus, for example, was coded 250 in ICD-9 and is coded E10-E14 in ICD-10 (ICD-10 2003). Without continuing education on code changes and additions, hospitals can lose reimbursement funds and researchers can lose data accuracy.

The coders' experience, attention, and persistence also affect the accuracy of coding. These errors are the fifth and sixth errors shown in Figure1. When a patient is admitted with renal failure and hypertension, a novice coder may code each condition separately, whereas an experienced coder will look to see if there is a connection between the two conditions, and if so, will use the specific combination code. If coders are unsure of a diagnosis or which diagnosis constitutes the principal diagnosis, they are expected to contact the physician or gather the necessary information to record the correct diagnosis. If coders fail to recognize when they need additional information or if they are not persistent in collecting it, additional error is imposed into the coding system.

At the phase of the paper trail in which diagnostic labels are translated into ICD codes, some specific types of coder-level errors can be identified. These errors, the next to last set in Figure 1, include creep, upcoding, and unbundling, to name a few. Creep includes diagnostic assignments that deviate from the governing rules of coding. Creep errors have also been labeled as misspecification, miscoding, and resequencing errors (Hsia et al. 1988).

Misspecification occurs when the primary diagnosis or order for tests and procedures is misaligned with the evidence found in the medical record. Miscoding includes assignment of generic codes when information exists for assigning more specific codes, assignment of incorrect codes according to the governing rules, or assignment of codes without the physician attesting to their accuracy (Hsia et al. 1988). An example of miscoding for an ischemic stroke might involve using the more generic ICD-9 code of 436 (acute but ill-defined cerebrovascular disease) in place of the more specific ICD-9 codes of 433 (occlusion and stenosis of precerebral arteries) or 434 (occlusion of cerebral arteries) (Goldstein 1998).

Resequencing codes, or changing the order of them, comprises another potential error source. Take as an example the patient who had respiratory failure as a manifestation of congestive heart failure. The congestive heart failure should be the principal diagnosis and the respiratory failure the secondary diagnosis. Resequencing errors occur when these diagnoses are reversed (Osborn 1999). Most sequencing errors are not deliberate. Sequencing errors may comprise the commonest kind of errors in hospital discharge abstracts (Lloyd and Rissing 1985).

Upcoding, assigning codes of higher reimbursement value over codes with lesser reimbursement value, is an additional source of error at the coder level. For example, upcoding a urinary tract infection to the more serious condition of septicemia results in an increase of over $2,000 in reimbursement. Because upcoding misrepresents the true condition of the patient, it constitutes falsification of medical records and can often be detected by comparing the medical record to the codes listed in the discharge abstract. When coders assign codes for all the separate parts of a diagnosis instead of assigning a code for the overall diagnosis, the practice is called unbundling. Whether done in error or intentionally for gain, unbundling constitutes coding error. Although we label these errors as coder-level errors, systematic coding variation may be apparent at the hospital-level, as shown by some evidence from at least one study (Romano et al. 2002).

The final potential error listed in Figure 1 occurs when the physician attests to the accuracy of the coding information. As physicians treat many patients simultaneously and support heavy workloads, the time and attention physicians dedicate to checking the accuracy of the codes varies tremendously. Errors at the point of attestation include reviewing too quickly or not reviewing the face sheet and supporting documents, poor recall of the details of the patient's conditions, and incorrect recording on the attestation sheet.

Ways to Measure Code Accuracy

Five statistics are commonly used to summarize the amount of error in ICD coding: sensitivity, specificity, positive predictive value, negative predictive value, and κ coefficient. These statistics are simple to compute; that is, it is easy to come up with the "right answer." It is more challenging to state precisely what questions are answered by each of these statistics. In the current context, it is helpful to remember that the reliability of ICD coding is with respect to some other method of obtaining a diagnostic label. Sensitivity and specificity are statistics often used when some "gold standard" is available. As discussed above, however, there is no gold standard for diagnostic labeling.

A researcher whose question is, "How accurate are the diagnoses?" might compare the diagnostic labels assigned by two or more experts (e.g., physicians) evaluating the same sample of patients. A good choice of statistic for this research design might be the κ coefficient. This statistic quantifies beyond-chance agreement among experts; therefore, it would be an appropriate estimator of the reliability of diagnoses made by experts. However, if the research question is, "In medical chart reviews, how well do medical

coders' ICD code assignments match those of physicians?" then a true gold standard exists. In such a case, the researcher might prefer to calculate specificity, sensitivity, and predictive values using the physicians' reviews as the gold standard. What must be kept clearly in mind, however, is that the values of the statistics obtained in this scenario express nothing about the reliability of medical diagnosis. They estimate, in the context of medical chart review, the corroboration between physician and medical coders' ICD classifications.

Discussion

The process of assigning ICD codes is complicated. The many steps and participants in the process introduce numerous opportunities for error. By describing a brief history of ICD coding, detailing the process for assigning codes, identifying places where errors can be introduced into the process, and reviewing methods for examining code accuracy, we hope to demystify the ICD code assignment process and help code users more systematically evaluate code accuracy for their particular applications. Consideration of code accuracy within the specific context of code use ultimately will improve measurement accuracy and, subsequently, health care decisions based on that measurement.

Although this paper focused on errors influencing code accuracy, the goal was not to disparage ICD codes in general. ICD codes have proven incredibly helpful for research, reimbursement, policymaking, etc. In fact, without ICD codes, health care research, policy, and practice could not have advanced as far as they have. However, code use and decision making on the bases of codes is improved when code accuracy is well understood and taken into account. By heightening their awareness of potential error sources, users can better evaluate the applicability and limitations of codes in their own context, and thus use ICD codes in optimal ways.

One way to heighten code users' awareness of potential error sources is to create a tool for their use when evaluating ICD codes. Based on our evaluation of the code assignment process, we created Figure 1, which summarizes the basic inpatient process for code assignment. This flowchart is designed to focus code users' attention on key aspects of the code assignment process and facilitate their critique of codes. By identifying potential code errors, users may be able to specify bias that might influence data accuracy. Instead of weakening a study, the recognition of potential sources of code bias will strengthen researchers' interpretations of data analyses using the codes.

A few practical recommendations can be made for code users. First, codes are likely to be most accurate under the following conditions: the disease has a clear definition with observable signs and symptoms, highly qualified physicians document information on the patient, experienced coders with full access to information assign the codes, and the codes are not new. Furthermore, codes are more likely to be accurate in calculating disease prevalence than in calculating disease incidence, as incidence requires identification of new cases, or cases without previous documentation. When the accuracy of a specified code is high, it would be appropriate to identify individuals for inclusion in patient registries or intervention studies. Codes considered less accurate are better suited for screening for potential study participants and for identifying pools of recruits. As code accuracy decreases, or becomes more questionable, researchers will want to use codes in combination with other measures. For example, codes from one occasion could be combined with test results or with codes from other occasions to improve the accuracy of disease classification. Suppose researchers wish to identify patients with stroke for an expensive intervention study. Given that past studies have found that using ICD-9 diagnostic codes of 433 through 436 from administrative databases are not very accurate for diagnosing stroke (Leibson et al. 1994; Benesch et al. 1997), researchers may decide to use codes 433 through 436 as an initial screener for including patients in the study. Then, to increase the accuracy of the diagnosis, researchers may wish to review these patients' hospitalization charts and outpatient records. Researchers can make study inclusion decisions based on the combination of ICD codes and clinical evidence, such as prior history of cerebral ischemic events, cerebrovascular risk factors, related procedures (carotid endarterectomy or angiography), and functional abilities. Although collecting the clinical evidence will take additional time and resources, it will improve the accuracy of the diagnoses and will likely lead to more appropriate study results.

Because the use of ICD codes is commonplace, and studies on code accuracy can be found in a wide variety of disease- and discipline-specific journals, code users need easy access to a resource for reviewing code accuracy studies. To meet this need, the Measurement Excellence and Training Resource Information Center (METRIC), a VA initiative for improving measurement in health care, has created a repository of abstracts from code accuracy studies. The repository is located at http://www.measurementexperts.org/icd9ab.htm. The METRIC started the repository, but recognizes many important study references are missing. METRIC encourages code users to visit the site, review the repository, and recommend other studies and

documents that can be added. With input from the many types of code users, this resource can become a valuable tool for evaluating ICD code accuracy. METRIC envisions this as a dynamic resource that will facilitate ICD code users' ability to access code accuracy information in an efficient and timely manner.

Although many studies have examined ICD code accuracy, knowledge in several areas is underdeveloped. Two important areas are the reliability of physician diagnoses and the factors that influence that reliability. Many ICD code accuracy studies consider the physician diagnosis as recorded in the medical record as the gold standard for measuring diagnoses (Lloyd and Rissing 1985; Hsia et al. 1988; Fischer et al. 1992). In at least one study, researchers demonstrated that the medical record cannot be considered a gold standard, as measured against standardized patients, for example Peabody et al. (2000).

Little consideration is given to the process leading to the physician's diagnosis. Certainly the quality of the gold standard varies based on disease factors (type, knowledge, and progression) and physician factors (experience with the disease and knowledge of diagnostic tools for the disease). Further research examining which factors influence the quality of the physician's diagnosis and the extent to which these factors affect the gold standard is greatly needed.

As researchers, policy makers, insurers, and others strive to impose some organization on the complicated health care field, disease and procedure classification systems will receive increased attention. Although no system of classification will ever be perfect, our ability to improve taxonomies rests in our dedication to understanding the code assignment process and to sharing information about its strengths and weaknesses.

ACKNOWLEDGMENTS

The authors would like to acknowledge and thank University of Kansas Medical Center staff members Richard Sahlfeld, RHIA, director and corporate privacy officer; Theresa Jackson, RHIA, assistant director; and Vicky Koehly, RHIT, coding supervisor, medical record department, for their input during our interviews and for their time spent in reviewing our manuscript. The authors would also like to thank Sandra Johnston, MA, RHIA, clinical coordinator, health information management department, School of Allied Health, University of Kansas, Kansas City, Kansas, for her contributions to the

manuscript. The expertise of these professionals in health information management and the coding process was invaluable.

Funding for the development of this manuscript came from the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service, Measurement Excellence and Training Resource Information Center (METRIC; RES 02-235).

This material is the result of work supported with resources and the use of facilities at the Houston Center for Quality of Care & Utilization Studies, Houston Veterans Affairs Medical Center.

The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or Baylor College of Medicine.

REFERENCES

Albers, G. W., L. R. Caplan, J. D. Easton, P. B. Fayad, J. P. Mohr, J. L. Saver, and D. G. Sherman, TIA Working Group. 2002. "Transient Ischemic Attack—Proposal for a New Definition." *New England Journal of Medicine* 347 (21): 1713–6.

American Psychiatric Association. 1994. *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition. Washington, DC: American Psychiatric Association.

Benesch, C., D. M. Witter Jr., A. L. Wilder, P. W. Duncan, G. P. Samsa, and D. B. Matchar. 1997. "Inaccuracy of the International Classification of Diseases (ICD-9-CM) in Identifying the Diagnosis of Ischemic Cerebrovascular Disease." *Neurology* 49: 660–4.

Bossuyt, P. M., J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, J. G. Lijmer, D. Moher, and D. Rennie, H. C. de Vet, and the STARD Group. 2004. "Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative." *Family Practice* 21: 4–10.

Calle, E. E., C. Rodriguez, K. Walker-Thurmond, and M. J. Thun. 2003. "Overweight, Obesity, and Mortality from Cancer in a Prospectively Studied Cohort of U.S. Adults." *New England Journal of Medicine* 348 (17): 1625–38.

Charbonneau, A., A. K. Rosen, A. S. Ash, R. R. Owen, B. Kader, A. Spiro, C. Hankin, L. R. Herz, M. J. V. Pugh, L. Kazis, D. R. Miller, and D. R. Berlowitz. 2003. "Measuring the Quality of Depression in a Large Integrated Health System." *Medical Care* 41: 669–80.

Colorado Department of Public Health and Environment. 2001. "New International Classification of Diseases (ICD-10): The History and Impact." Brief Health Statistics Section, March 2001, No. 41. Available at http://www.cdphe.state.co.us/hs/Briefs/icd10brief.pdf

Corn, R. F. 1981. "The Sensitivity of Prospective Hospital Reimbursement to Errors in Patient Data." *Inquiry* 18: 351–60.

Department of Veterans Affairs. 2002. *Handbook for Coding Guidelines, Version 2.0.* Health Information Management. Available at http://www.virec.research.med.va.gov/References/VHACodingHandbook/CodingGuidelines.htm

Dewey, H. M., G. A. Donnan, E. J. Freeman, C. M. Sharples, R. A. Macdonell, J. J. McNeil, and A. G. Thrift. 1999. "Interrater Reliability of the National Institutes of Health Stroke Scale: Rating by Neurologists and Nurses in a Community-Based Stroke Incidence Study." *Cerebrovascular Disease* 9 (6): 323–7.

Doremus, H. D., and E. M. Michenzi. 1983. "Data Quality: An Illustration of Its Potential Impact upon Diagnosis-Related Group's Case Mix Index and Reimbursement." *Medical Care* 21: 1001–11.

Faciszewski, T., S. K. Broste, and D. Fardon. 1997. "Quality of Data Regarding Diagnoses of Spinal Disorders in Administrative Databases. A Multicenter Study." *Journal of Bone Joint Surgery America* 79: 1481–8.

Fischer, E. D., F. S. Whaley, W. M. Krushat, D. J. Malenka, C. Fleming, J. A. Baron, and D. C. Hsia. 1992. "The Accuracy of Medicare's Hospital Claims Data: Progress Has Been Made, but Problems Remain." *American Journal of Public Health* 82: 243–8.

Goldstein, L. B. 1998. "Accuracy of ICD-9-CM Coding for the Identification of Patients with Acute Ischemic Stroke: Effect of Modifier Codes." *Stroke* 29 (8): 1602–4.

Goldstein, L. B., M. R. Jones, D. B. Matchar, L. J. Edwards, J. Hoff, V. Chilukuri, S. B. Armstrong, and R. D. Horner. 2001. "Improving the Reliability of Stroke Subgroup Classification Using the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) Criteria." *Stroke* 32 (5): 1091–8.

Green, J., and N. Wintfeld. 1993. "How Accurate Are Hospital Discharge Data for Evaluating Effectiveness of Care?" *Medical Care* 31: 719–31.

Hsia, D. C., W. M. Krushat, A. B. Fagan, J. A. Tebbutt, and R. P. Kusserow. 1988. "Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-Payment System." *New England Journal of Medicine* 318 (6): 352–25.

Institute of Medicine. 1977. *Reliability of Hospital Discharge Records.* Washington, DC: National Academy of Sciences.

International Classification of Diseases, 10th Revision (ICD-10). 2003. Department of Health and Human Services. Centers for Disease, Control and Prevention. National Center for Health Statistics. Available at http://www.cdc.gov/nchs/data/dvs/icd10fct.pdf

International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM), Sixth Edition. 2002. Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. Available at ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD9-CM/2002/

International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM). 2003. National Center for Health Statistics. Pre-release Draft, June 2003. Centers for Disease Control and Prevention. Available at http://www.cdc.gov/nchs/about/otheract/icd9/icd10cm.htm

Jackson, L. A., K. M. Neuzil, O. Yu, W. E. Barlow, A. L. Adams, C. A. Hanson, L. D. Mahoney, D. K. Shay, and W. W. Thompson. 2003. "Effetiveness of

Pneumococcal Polysaccharide Vaccine in Older Adults." *New England Journal of Medicine* 348 (18): 1747–55.

Johnson, A. N., and G. L. Appel. 1984. "DRGs and Hospital Case Records: Implications for Medicare Casemix Accuracy." *Inquiry* 21: 128–34.

Johnston, S. C., P. B. Fayad, P. B. Gorelick, D. F. Hanley, P. Shwayder, D. Van Husen, and T. Weiskopf. 2003. "Prevalence and Knowledge of Transient Ischemic Attack among US Adults." *Neurology* 60 (9): 1429–34.

Jolis, J. G., M. Ancukiewics, E. R. DeLong, D. B. Pryor, L. H. Muhlbaier, and D. B. Mark. "Discordance of Databases Designed for Claims Payment versus Clinical Information Systems." *Annals of Internal Medicine* 119: 844–50.

Leibson, C. L., J. M. Naessens, R. D. Brown, and J. P. Whisnant. 1994. "Accuracy of Hospital Discharge Abstracts for Identifying Stroke." *Stroke* 25: 2348–55.

Lloyd, S. S., and J. P. Rissing. 1985. "Physician and Coding Errors in Patient Records." *Journal of the American Medical Association* 254 (10): 1330–6.

Martin, G. S., D. M. Mannino, S. Eaton, and M. Moss. 2003. "The Epidemiology of Sepsis in the United States from 1979 through 2000." *New England Journal of Medicine* 348 (16): 1546–54.

Ornelas-Aguirre, J. M., G. Vazquez-Camacho, L. Gonzalez-Lopez, A. Garcia-Gonzalez, and J. I. Gamez-Nava. 2003. "Concordance between Premortem and Postmortem Diagnosis in the Autopsy: Results of a 10-Year Study in a Tertiary Care Center." *Annals of Diagnostic Pathology* 7 (4): 223–30.

Osborn, C. E. 1999. "Benchmarking with National ICD-9-CM Coded Data." *Journal of the American Health Information Management Association* 70 (3): 59–69.

Peabody, J. W., J. Luck, P. Glassman, T. R. Dresselhaus, and M. Lee. 2000. "Comparison of Vignettes, Standardized Patients, and Chart Abstraction: A Prospective Validation Study of 3 Methods for Measuring Quality." *Journal of the American Medical Association* 283: 1715–22.

Provenzale, J. M., R. Jahan, T. P. Naidich, and A. J. Fox. 2003. "Assessment of the Patient with Hyperacute Stroke: Imaging and Therapy." *Radiology* 229 (2): 347–59.

Romano, P. S., B. K. Chan, M. E. Schembri, and J. A. Rainwater. 2002. "Can Administrative Data Be Used to Compare Postoperative Complication Rates across Hospitals?" *Medical Care* 40: 847–50.

Silfvast, T., O. Takkunen, E. Kolho, L. C. Andersson, and P. Rosenberg. 2003. "Characteristics of Discrepancies between Clinical and Autopsy Diagnoses in the Intensive Care Unit: A 5-Year Review." *Intensive Care Medicine* 29 (2): 321–4.

Steinman, M. A., C. S. Landefeld, and R. Gonzales. 2003. "Predictors of Broad-Spectrum Antibiotic Prescribing for Acute Respiratory Tract Infections in Adult Primary Care." *Journal of the American Medical Association* 289 (6): 719–25.

Studdert, D. M., and C. R. Gresenz. 2003. "Enrollee Appeals of Preservice Coverage Denials at 2 Health Maintenance Organizations." *Journal of the American Medical Association* 289 (7): 864–70.

U.S. Congress. 1985. *Office of Technology Assessment, "Medicare's Prospective Payment System: Strategies for Evaluating Cost, Quality, and Medical Technology," OTA-H-262.* Washington, DC: U.S. Government Printing Office.

Uniform Hospital Discharge Data Set (UHDDS). 1992. "Definition of Principal and Other [Secondary] Diagnoses." 50 Federal Register 31039; adopted 1986, revised 1992.

Wilson, J. T., A. Hareendran, M. Grant, T. Baird, U. G. Schulz, K. W. Muir, and I. Bone. 2002. "Improving the Assessment of Outcomes in Stroke: Use of a Structured Interview to Assign Grades on the Modified Rankin Scale." *Stroke* 33 (9): 2243–6.