

Determination of Problematic ICD-9-CM Subcategories for Further Study of Coding Performance: Delphi Method

by Xiaoming Zeng, MD, PhD; and Paul D. Bell, PhD, RHIA, CTR

Abstract

In this study, we report on a qualitative method known as the Delphi method, used in the first part of a research study for improving the accuracy and reliability of ICD-9-CM coding. A panel of independent coding experts interacted methodically to determine that the three criteria to identify a problematic ICD-9-CM subcategory for further study were cost, volume, and level of coding confusion caused. The Medicare Provider Analysis and Review (MEDPAR) 2007 fiscal year data set as well as suggestions from the experts were used to identify coding subcategories based on cost and volume data. Next, the panelists performed two rounds of independent ranking before identifying Excisional Debridement as the subcategory that causes the most confusion among coders. As a result, they recommended it for further study aimed at improving coding accuracy and variation. This framework can be adopted at different levels for similar studies in need of a schema for determining problematic subcategories of code sets.

Keywords: coding, Delphi method, qualitative research, accuracy, consistency, ICD-9-CM, excisional debridement

Medical coding is one of the core practice areas of health information management (HIM) professionals. It involves assigning standardized classification or terminology codes (e.g., International Classification of Diseases [ICD] or Current Procedural Terminology [CPT]) to the original documents recording diseases, injuries, or procedures (e.g., medical records) pertinent to individual patient cases. The assigned codes are then used to represent the cases in ensuing healthcare transactions such as reimbursement. Although the application of medical coding is primarily for the purposes of administration and reimbursement, these codes are widely employed as the bases for additional analysis of quality improvement, health outcomes and services research, disease surveillance, and other purposes.¹⁻⁵

Because medical coding serves as an important nexus between the primary data sources and many of their secondary data usages, inaccuracy or variation present in low-quality coding will be carried forward to such secondary use. A wrongly coded case will not be included in the collection of data used for secondary analysis if the code is used as the field for querying the data. Moreover, with electronic health records, e-HIM, and health information exchanges being increasingly implemented in healthcare, such a “cascading” effect might be magnified without prompt interventions from the information system.^{6,7}

With the increased promotion and adoption of electronic health records (EHRs) in healthcare management and transactions, addressing the coding quality issue has become increasingly important. A GIGO (garbage in, garbage out) effect associated with digitization presents a unique challenge to quality

management in an electronic healthcare environment. Codes assigned through the medical coding process often are considered one of the “upstream” points for quality of data along the patient and cash flows. Therefore, guidelines directing the coding process need to be reviewed and updated consistently in order to reduce coders’ confusion and improve coding accuracy.^{8,9}

Background and Significance

Medical coding systems transform medical descriptions of patient diseases, injuries, and procedures into predefined alphanumeric codes that for the most part can be quickly entered into computer systems and processed into useful healthcare information. The codes with relevant documents are then submitted to payers to get reimbursement for the care. Additionally, these coded data can be used for many purposes other than reimbursement for medical care provided. For example, International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) data are used as a key source of information for medical research to improve the quality of patient care.¹⁰ They can substantiate that the care rendered was medically necessary, healthcare resources were properly used, and healthcare provider charges were reasonable.^{11,12} Finally, medical codes also form the basis for the vital statistics that are generated concerning morbidity and mortality in the country.¹³

The current official code set used in the United States is ICD-9-CM for medical diagnoses and certain procedures in hospitals. It will be upgraded to International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) and International Classification of Diseases, Tenth Revision, Procedure Coding System (ICD-10-PCS) after October 1, 2013. It is both generally accepted and documented that code assignment in the United States using the current classification systems, especially ICD-9-CM, remains somewhat subjective.

Interest in the study of ICD-9-CM coding quality started to grow around the time when the prospective payment system using diagnosis related groups (DRGs) was implemented in the 1980s. One study of ICD-9-CM diagnostic code assignment accuracy reported an average 20.8 percent chance of coding the discharge of a patient inaccurately in such a way that the case’s DRG would change.¹⁴ The same author did another specific study focusing on the accuracy of medical reimbursement for cardiac arrest cases. He found that cardiac arrest (DRG 129) had significantly higher rates of coding errors than other DRGs. Among the 857 medical records coded as cardiac arrest, 42.1 percent of the patients entered the hospital for other heart diseases and 55.2 percent died in the hospital from other diseases.¹⁵

Another early study looking at both code accuracy and variation reported 22 percent inaccuracy in coding items and disagreement on at least one data field in more than 50 percent of the cases from five VA hospitals.¹⁶ Three sources were identified for such error rates: physician (62 percent), coding (35 percent), and keypunch (3 percent). The author projected that there were 0.81 coding errors in the average abstract. If the errors were corrected in the abstracts, it would change 19 percent of the records for DRG purposes and substantially increase future resource allocations.

Campos-Outcalt (1990) studied the accuracy of ICD-9-CM codes in identifying reportable communicable diseases and found that 33 percent of the cases identified as one of 20 communicable diseases, using the ICD-9-CM codes, were incorrectly coded.¹⁷ The high rate of false positives posed a challenge when using encounter form data and ICD-9-CM codes to identify communicable diseases in an outpatient setting.

The most systematic analysis of the reasons for ICD-9-CM coding accuracy is by O’Malley and colleagues (2005).¹⁸ They examined potential sources of errors at each step of the inpatient ICD-9-CM coding process. Many factors contribute to the inaccuracy of ICD-9-CM coding: amount and quality of information at admission, communication among patients and providers, the clinician’s knowledge and experiences with the illness, the clinician’s attention to details, variance in the electronic and written records, coder training and experience, facility quality-control efforts, and unintentional and intentional coder errors (e.g., misspecification, unbundling, and upcoding).

The problem of coding inaccuracy and variation will not disappear even with the adoption of more detailed ICD-10 codes. A study in the United Kingdom assessed the accuracy of diagnostic coding using

ICD-10 in general surgery in a district general hospital by comparing codes ascribed by hospital coders to codes ascribed by expert external coders.¹⁹ They found errors of coding in 87 of 298 records (29 percent), among which 25 of 298 (8 percent) were at the most serious level (i.e., wrong ICD-10 chapter). One conclusion they drew from their study was that coding accuracy is determined by the type and quality of clinical documents. Coding should be carried out from the medical records rather than the admission form.

Another recent study from Norway looked at the diagnostic coding accuracy for traumatic spinal cord injuries (SCIs). ICD-10 codes were used to identify all hospital admissions with discharge codes suggesting a traumatic SCI. These cases were later verified using electronic health records. Among the 1,080 patients with an ICD-10 diagnostic code suggesting a traumatic SCI, only 260 SCIs were verified after reviewing the patients' hospital records. The positive predictive value (PPV) for identifying traumatic spinal cord injuries using ICD-10 codes is about .88. They concluded that ICD data cannot be trusted without extensive validity checks for either research or health planning and administration.²⁰

Given the need for high-quality data for clinical and administrative purposes, these rates are unacceptable whether achieved by a human or by using a software application. There are many different causes of code assignment inaccuracy and disagreement.²¹⁻²⁷ They include factors related to quality of documentation by the physicians, factors related to coder training and experience, and factors that are less easily addressed, such as the quality of communication between the patient and physician.

It needs to be pointed out that not all ICD-9-CM coding subcategories have the same rate of inaccuracy and variation. Despite previous studies that have broadly raised the issue of inaccuracy and variation in ICD-9-CM coding for certain subcategories of the code set,²⁸⁻⁴⁰ no studies to date have systematically identified specific high-volume, high-impact ICD-9-CM codes that are difficult to assign accurately and consistently. Moreover, there is a need for a methodological framework that can not only identify particular causes of inaccurate and inconsistent code assignment but also lead to recommendations for improvement in how codes are assigned.

This study is the first part of a two-step study to design, develop, and assess ICD-9-CM coding guidelines in order to improve coding performance. In this step, an ICD-9-CM code set subcategory needs to be identified for subsequent study and improvement. We report here a methodological framework that could allow researchers to identify problematic code subcategories in a systematic and objective manner.

Methods

The entire ICD-9-CM code set contains more than 14,000 codes used for diagnoses in both inpatient and outpatient settings as well as for inpatient procedures. Due to the limited funding for this study, the large number of codes and code subcategories in ICD-9-CM, and variations related to individual code subcategories, we have to focus on a specific subcategory of the ICD-9-CM code set for improving the guidelines pertinent to it. In this study, we report a qualitative method used to determine the subcategory to use for further research on improving the accuracy and consistency of ICD-9-CM coding. The method can be applied to other similar studies targeting the improvement of medical coding quality.

The study used a qualitative research method called the Delphi method. In the literature it is sometimes referred to as the Delphi technique. It is a qualitative consensus-building method that solicits predictions from a group of independent experts in a systematic and interactive way. The Delphi method is based on the principle that prediction from a structured group of experts is more accurate and reliable than prediction from an unstructured group of experts. It was originally developed at RAND Corporation during the Cold War to forecast the impact of technology on warfare.⁴¹ It was, then, widely used in many fields where consensus prediction needed to be built among a group of experts. The group usually consists of seven to eight members. Several rounds of individual assessment and group synthesis are usually needed before predefined criteria for consensus are met. The method can prevent a more vocal expert from dominating the discussion and consequently skewing the final prediction. Because the individual assessments are shared only within the group and the only measure used is the average score of the assessment, participating experts are not influenced by any one individual in the group.⁴²

The study was approved by the Institutional Review Board (IRB) at East Carolina University. With help from AHIMA, a group of seven specialists in medical coding were identified. They included recognized industry experts who either had a professional credential such as RHIA, CCS, or RN or had extensive practical experience, greater than five years of experience as coding consultants/advisors, administrators of coding classification systems, or directors of research/policy making in coding and classification systems. Their job titles include developer of the DRG patient classification scheme and senior vice president of clinical and economic research for 3M Health Information Systems, medical systems administrator for the National Center for Health Statistics, director of coding and classification at the American Hospital Association (AHA), senior coding technical advisor for the Hospital Ambulatory Policy Group of the Centers for Medicare and Medicaid Services (CMS), director of coding policy and compliance at AHIMA, and nurse MBA serving as ICD-10 program manager at 3M.

The experts agreed to participate as panelists for the study, and consent forms were obtained from them at the beginning of the study. The study was conducted in 2008. All communications in this study were through teleconference calls and e-mails with the expert panelists.

Determining the Criteria and Weights

At the first meeting with all the panelists, three criteria were proposed to choose the final ICD-9-CM subcategory for consideration in the study: cost, volume, and degree of confusion. Cost (*c*) is determined by the monetary value of the cases with the specific codes for the healthcare system in the United States. In other words, the cost criterion is based on the amount of money the cases with specific ICD-9-CM code(s) cost the U.S. healthcare system. Volume (*v*) is defined as the total number of cases that happen each year in the United States. The last criterion, the degree of confusion (*cf*) when coding cases in certain subcategories of ICD-9-CM, is a subjective measure that would be estimated by the panelists according to their experience.

Each criterion needed an associated weight (percentage) in order to calculate the weighted total rank order of the candidate ICD-9-CM subcategories (see Figure 1). The weights of the three criteria had to add up to one. For example, cost could be weighted 40 percent, volume 30 percent, and level of confusion 30 percent when calculating the weighted rank for a specific ICD-9-CM subcategory. If the category is ranked first in cost, second in volume, and third in level of confusion, then its final weighted rank is 1.9 ($1 \times 0.40 + 2 \times 0.30 + 3 \times 0.30$). In this way, we included all three criteria in the decision-making process while also incorporating their individual impact in the selection process. (See Figure 1.)

The study consisted of three steps. First, the weights associated with each criterion were determined by soliciting estimates from the experts. We used the Delphi method in this step to obtain the final weights. Second, we asked the panelists to provide the candidate ICD-9-CM code subcategories that they deemed confusing based on their experience and that they felt should be considered for future study to improve the performance of coding. We combined all the recommended subcategories to compile a list of unique candidate subcategories for consideration in the study. Third, the ranked order of the degree of confusion for each individual code subcategory was estimated after aggregating the ranked orders offered by the individual expert panelists. The Delphi method was also used in this step to get the converged estimate of rank order. Then we calculated the weighted total scores of the subcategories as the basis for the final ranking.

The panelists were given an explanation of the purpose and protocol of the study in the first meeting. Three criteria—cost, volume, and confusion—were proposed as the three criteria for determining the final overall rank order of the ICD-9-CM code subcategories. All panelists agreed that the three proposed criteria should be the ones used as the criteria for the final rank order. Each panelist was asked to assign a weight between 0 and 100 percent to each criterion resulting in a total of 100 percent for all three weights. The panelists were asked to work on the weights independently. The results of their estimated weights were sent to the researchers via e-mail. We calculated the averages and standard deviations of each weight estimate and then sent them back to the panelists for any updates after they reviewed the statistics. Panelists were asked to send a second round of estimated weights if they had updates to their original

estimates. The averages and standard deviations of the updated weights were presented at a teleconference meeting of the panel. If all panelists agreed that it was unnecessary to have another round of estimates, the weights would be used as the weights for the final calculation.

Determining the Candidate Code Subcategories

In order to produce a list of subcategories for inclusion in the study, we asked each panelist to submit a list, based on experience, of ICD-9-CM code subcategories that are responsible for a large degree of confusion when coding. Each panelist compiled a list independently. We combined all the submitted code subcategories into a final list with unique subcategories.

Determining the Final Rank Order

The criteria cost and volume can be determined objectively by querying a reimbursement database. We used the Medicare Provider Analysis and Review (MEDPAR) fiscal year (FY) 2007 inpatient acute care patient claims analysis data set to retrieve data on the cost and volume of the code subcategories. MEDPAR is a database that contains data from claims for services provided to beneficiaries admitted to Medicare-certified inpatient hospital and skilled nursing facilities. Although it does not encompass all the claims in the healthcare market, it is a sound surrogate given the large market share of Medicare in terms of both number of patients and number of encounters.⁴³

The MEDPAR fiscal year 2007 file consists of a total of 13,058,880 records. A report for each of the 13 individual code set categories selected for study was computed and yielded the following data ([View the complete report](#)):

1. the number of records in the analysis that contained one or more codes in the code set category coded as a secondary diagnosis or as a principal or secondary procedure;
2. the occurrence of the codes in the code set category coded as a secondary diagnosis or as a principal or secondary procedure;
3. the number of records where the codes in the code set category impacted the CMS version 26 (v26) DRG assignment; and
4. the estimated DRG payment impact for those records with codes in the code set category that impacted CMS v26 DRG assignment.

Volume

These four statistics were computed for cases assigned into a medical DRG, a surgical DRG, and an overall group (medical and surgical cases combined). Further, these overall totals for each code set category were also computed for each of the 25 major diagnosis categories (MDCs) plus the pre-MDC. A code category subset was not counted if it did not appear as a secondary diagnosis or if its impact as a secondary diagnosis failed to change the DRG assignment of the case. Malignant ascites did not appear as a secondary diagnosis because there were no cases in the MEDPAR FY 2007 data file matching this subcategory. This is because code 789.51 did not become effective until October 1, 2007, the beginning of FY 2008. The current study only included MEDPAR data from fiscal year 2007 (October 1, 2006, to September 30, 2007). Therefore, as the code for malignant ascites was not available until fiscal year 2008, no cases for malignant ascites (789.51) existed in the database. In the case of sepsis and bacteremia, a case was counted as “sepsis/bacteremia” if it contained either sepsis or bacteremia in the code set category as a secondary diagnosis.

Thus, a particular code set subcategory was ranked higher on the “volume” criterion if it had a greater number of cases with at least one appropriate code from its code set subcategory compared to the other coding subcategories.

Cost

The records in the MEDPAR data file were all grouped and assigned a CMS v26 DRG. In order to assess the “cost” impact of the codes in the code set category, first, the records in the analysis data set that were coded with one or more of the codes from the code set category as a secondary diagnosis or procedure were identified and then regrouped under CMS v26 DRG without passing any of the codes from the code set category for DRG assignment. The records that impacted CMS v26 DRG assignment were those records that, when regrouped without the codes from the code set category, resulted in a different CMS v26 DRG from the original CMS v26 DRG assignment. The number of cases impacting CMS v26 DRG assignment appears in the “cases affecting DRG assignment” column of the MEDPAR FY 2007 data file.

In order to quantify the actual dollar impact of a particular code set subcategory, it was necessary to calculate the relative weight of the cases in a CMS DRG. This was done as follows: cases grouped into a different CMS DRG when the codes from the code set category were not included in the CMS DRG assignment had a different relative weight associated with the case compared to the original CMS DRG assignment based on all the codes on the record. Next, the configured relative weight rate was multiplied by the base rate of \$6,320.21 according to the FY 2009 final rule CMS v26 DRG. The difference in the relative weight times the base rate generated the payment impact for the record when not allowing the codes from the code set category to impact CMS DRG assignment. The DRG payment impact is reflected by the number in the “overall impact \$” column of the MEDPAR FY 2007 data file.

Records that did not have any codes from the code set category (i.e., malignant ascites) would not have any CMS DRG payment impact since the CMS DRG would not have been impacted by the code set category. Further, records that did have codes from the code set category but for which the CMS DRG assignment did not change when codes from the code set category were not passed for CMS DRG assignment would also not have any CMS DRG payment impact. For example, this was true of cases where the congestive heart failure (CHF) secondary diagnosis code 4280 was excluded. However, cases where 4280 was excluded yet the case included at least one other associated code (either for systolic or diastolic heart failure) or cases that included 4280 as well at least one other associated code resulted in a change to the CMS v26 DRG assignment and hence affected payment.⁴⁴

Thus, a particular code set subcategory was ranked higher on the “cost” criterion if it had a greater overall payment impact compared to the other coding subcategories.

Confusion Rate

The ranking of the confusion rate was determined subjectively by the panelists. Because there are no reliable data sources devoted to tracking the confusion rate of coders when they code medical cases, we asked panelists to rank the candidate code subcategories according to how confusing they consider them based on their past experiences. Confusion in medical coding could be from two sources: clinician documentation and coding guidelines. Poor documentation could be a challenge to coders because of incomplete or inaccurate information. On the other hand, if a coding guideline for a particular code subcategory is vague, it can cause large variations in how the code is applied when being coded. The panelists provided two ranks of confusion, one based on the confusion caused by poor documentation and the other based on the confusion caused by a vague guideline. The overall score for ranking each code subcategory is the average of these two ranks. For example, if Diabetes Mellitus is ranked first in confusion caused by poor documentation and second in confusion caused by poor coding guidelines, then it will receive an average score of 1.5 as the basis for ranking all the subcategories.

We again used the Delphi method to determine the final rank of the confusion rates of the individual subcategories. After each panelist sent us a ranked list of code subcategories based on level of confusion, we calculated the average and standard deviations of the rank orders and of each code subcategory in order to provide feedback to the panelists. Panelists were then asked to update their rank ordering after reviewing the statistics. The results of the updated averages and standard deviations were discussed at a

teleconference. If no one opposed or objected to the rank order from the aggregated data, it would be used as the panelists' final rank of the confusion rates.

Using the formula shown in Figure 1, we calculated the final weighted total score from the three criteria: cost, volume, and confusion. For instance, the score for the subcategory Excisional Debridement is 5.175 (calculated as 12×32.8 percent + 4×18.9 percent + 1×48.3 percent). The final rank of the code subcategories was determined based on their calculated scores from lowest to highest.

Results

Weights of Individual Criteria

The panelists agreed that the weights could be finalized after two rounds of estimation (Table 1). The standard deviations decreased between the two rounds. This indicates that the participants were reaching a consensus on the weights to apply for each individual criterion. Confusion was assigned almost half of the total weight (48.3 percent) among the three criteria.

Confusion Rate Ranking

We received a list of 13 unique ICD-9-CM code categories from the panelists. After two rounds of the Delphi process, the panelists agreed that there was no need to have another round of ranking. The final ranking is the average of the two rank orders—documentation and guideline (Table 2). Excisional Debridement was ranked as the ICD-9-CM code subcategory with the most confusion associated with it. Note that the standard deviation of the rank order value is zero after round 2, which showed that all the panelists agreed that the Excisional Debridement subcategory causes the most confusion during the coding process.

Final Weighted Total Rank

Table 3 shows the value and ranks of volume and cost of individual cases and the final rank of confusion rate of the code subcategories. The final weighted total scores were calculated using the formula in Figure 1. A lower score means that the subcategory has a higher level of importance based on the three criteria. Based on this scoring rubric, the top three ICD-9-CM code subcategories are Complications, Congestive Heart Failure, and Excisional Debridement. During the final discussion, panelists pointed out that the Complications subcategory is impractical to address because there is no designated guideline for it and the code could be dispersed throughout the entire code set. It was also pointed out that the guideline for the Congestive Heart Failure subcategory is already being investigated at CDC in order to improve coding performance. The panelists all agreed, then, that Excisional Debridement should be chosen as the candidate for further study, especially because it is number one in the confusion rate ranking. (See Table 3.)

Discussion

The purpose of this study is to systematically determine problematic ICD-9-CM subcategories for future coding performance study. It is the first step in a series of planned studies to develop and assess ICD-9-CM coding guidelines with the intent to improve coding performance of the selected subcategory. We reported on a methodological framework that could be employed to identify problematic medical code set subcategories from the perspective of cost, volume, and degree of confusion to the coders. The methodological framework employed both objective and subjective measures in order to identify the problematic code subcategory for the next phase of the study, which will investigate the causes of coding confusion and their potential solutions. Subsequent studies that concentrate on the coding guidelines for the specific subcategory—Excisional Debridement—to improve the coding quality will be reported separately.

We employed both objective and subjective measures in the methodological framework in order to rank order the candidate coding subcategories. The objective measures are cost and volume of cases assigned a code in certain code subcategories. The subject measure is the degree of confusion when a medical coder determines codes for cases in these subcategories. The MEDPAR data file, which provides

information on cost as well as volume of services, was the source for the study's objective measures. The experts' experiences and perception were used to determine the subjective ranking of the coding confusion rate. The Delphi method was employed twice in the study to facilitate determining the weight of the three measures and the rank of the confusion rate of the subcategories.

Although we used national-level data (MEDPAR) to determine the final code subcategory, the same methodological framework can be adopted for identifying problematic subcategories of medical code sets at different levels. Given the size of any code set (ICD, CPT, etc.), it is often impractical to study the entire code set without sizeable resources. It is also cost ineffective to study certain subcategories that cause confusion for coders but apply to only a small number of cases. For example, the methodological framework can be used by hospitals when they invest resources to study certain subcategories of the ICD-9-CM code set in order to improve their coding quality to reduce claim denials or returns from the payers. They can use their local database of claim denials or returns as the source for objective measures on cost and volume of these claims. Then the hospital can ask a group of coders using the Delphi technique to identify and rank the code subcategories that are most problematic during the coding process. The same framework then could be employed to integrate these measures into one rank order.

Excisional Debridement was ranked as the ICD-9-CM code subcategory with the most confusion associated with it. It is significant to note that Excisional Debridement was the only code set category for which the standard deviation of the rank order value was zero after round 2. This meant that all the panelists agreed that among all the ICD-9-CM codes, the Excisional Debridement subcategory caused the most confusion during the coding process. As a result, though it ranked low in volume, it was selected as the code subcategory warranting further study because it ranked the highest in confusion rate (which represented almost half of the total weight among the three criteria) and ranked relatively high in cost.

One source of the confusion surrounding this category is certainly that medical documentation is often not detailed or accurate enough to support coding as an Excisional Debridement case. The panel experts ranked Excisional Debridement high in confusion caused by documentation, which has been discussed in previous *Coding Clinic* updates. For example, the first-quarter 2008 *Coding Clinic*⁴⁵ validates that if a physician's documentation in a progress note states that excisional debridement was carried out, the coder is correct to use 86.22, Excisional debridement of wound, infection, or burn. However, there are still dozens of exceptions depending on the site, the tissue, and the methodology. The coder needs to read the entire operative report, which can contradict the title of a procedure. For example, if the physician titles the report "Excisional Debridement" but it clearly describes an incision and drainage, the coder must query for further clarification in order to correctly assign a procedure code. O'Malley et al. (2005) identified error sources along the "paper trail" of patient records including variance in the records as one of the major error sources.⁴⁶ Therefore, coding professionals should obtain accurate and complete documentation to support the assignment of the Excisional Debridement code subcategory. A key element in preventing such coding inaccuracy involves educating providers on the components needed in the documentation to support the code and communicating with them about any deficiency in the medical records.

It is noteworthy that Excisional Debridement was also one of the coding topics that received attention through the CMS Recovery Audit Contractor (RAC) program. CMS RAC demonstration projects in New York, California, and Florida found that Excisional Debridement coding was the leading source of ICD-9-CM procedure coding errors.⁴⁷ Specifically, they found cases of coding of Excisional Debridement (86.22) that did not meet the definition of Excisional Debridement. Even though there have been frequent updates and clarifications in the AHA *Coding Clinic*, Excisional Debridement still remains one of the most problematic subcategories for coders at all levels.⁴⁸

Moreover, most of the overpayment amounts collected by the RACs (about 85 percent) were from inpatient hospital providers, and almost half of such overpayments were the result of incorrect coding. Thus, it is important for hospitals to proactively identify potential problem code set subcategories before they are identified to be out of compliance by a RAC audit. One way for facility staff to avoid such an audit may be to perform their own internal studies using methodological frameworks that identify problematic code subcategories in a systematic manner.

Limitations of the Current Study

There are several limitations of the study that need attention when drawing conclusions from the findings.

1. We used a relatively small number of experts in determining the subcategory of the ICD-9-CM code set that we need to study. As for all qualitative studies, there is a limitation on how to generalize from the results of the study. The study relies heavily on the experiences and knowledge of the experts. The experts we identified in the study are all well known in the field and have experience in educating coders and in the area of coding quality improvement. The Delphi method is used to minimize the bias in the final prediction.
2. MEDPAR is a data set used by CMS for Medicare patients only. As Medicare serves the elderly population, it may not represent the entire picture of the volume of healthcare utilization in the United States. However, given the percentage share of Medicare cases within the entire healthcare system, it may represent an important part of the picture. Needleman (2003) reported on a study using MEDPAR data as a surrogate to estimate the quality measures (i.e., adverse outcomes) for all patients in quality-of-care research using administrative data sets. They found that analyses of quality of care for medical patients using Medicare-only and all-patient data are likely to have similar findings.
3. The evaluation of the effect of a particular code set subcategory on cost was determined by utilizing its estimated DRG payment. Because this part of the study looked only at the effect of the study's code set subcategories on Medicare DRG payments, rather than at the effect on all types of provider payments, it may not have reflected the true costs associated with providing care for the identified coding subcategories. However, given that Medicare payments represent 20 percent of all national health expenditures, the estimated DRG payment is probably a valid and robust construct to use in assessing a coding subcategory's impact on costs in the present study.⁴⁹

Recommendations for Future Studies

The importance of certain code set subcategories will vary with changes in healthcare policies, reimbursement models, coding systems, and RAC regulations. However, the methods of this study can be applied to future studies of specific code sets in medical coding, regardless of the types or versions of the code sets investigated. We suggest, then, that future studies designed to look at problematic code set categories should consider using the framework proposed in this study to systematically review and choose the specific subcategory for study.

Acknowledgment

This study was part of the FORE-directed research project titled "Improving the Code Assignment Accuracy of Historically Problematic Codes," sponsored by the AHIMA Foundation of Research and Education (FORE) and 3M Health Information Services. Susan Fenton, PhD, RHIA, provided support and guidance on the study.

Xiaoming Zeng, MD, PhD, is an associate professor and chair of the Department of Health Services and Information Management in the College of Allied Health Sciences at East Carolina University in Greenville, NC.

Paul D. Bell, PhD, RHIA, CTR, is a professor in the Department of Health Services and Information Management, College of Allied Health Sciences at East Carolina University in Greenville, NC.

Notes

1. Iezzoni, L. I. "Assessing Quality Using Administrative Data." *Annals of Internal Medicine* 127, no. 8, pt. 2 (1997): 666.
2. Ginde, A. A., C.-L. Tsai, P. G. Blanc, and C. A. Camargo. "Positive Predictive Value of ICD-9-CM Codes to Detect Acute Exacerbation of COPD in the Emergency Department." *The Joint Commission Journal on Quality and Patient Safety* 34, no. 11 (2008): 678–80.
3. Cherkin, D. C., R. A. Deyo, E. Volinn, and J. D. Loeser. "Use of the International Classification of Diseases (ICD-9-CM) to Identify Hospitalizations for Mechanical Low Back Problems in Administrative Databases." *Spine* 17, no. 7 (1992): 817.
4. Tsui, F.-C. "Value of ICD-9-Coded Chief Complaints for Detection of Epidemics." *Journal of the American Medical Informatics Association* 9, no. 90061 (2002): 41S–47.
5. Iezzoni, L. I. "Using Administrative Diagnostic Data to Assess the Quality of Hospital Care: Pitfalls and Potential of ICD-9-CM." *International Journal of Technology Assessment in Health Care* 6, no. 2 (1990): 272–81.
6. Hirschtick, R. E. "A Piece of My Mind: Copy-and-paste." *Journal of the American Medical Association* 295, no. 20 (2006): 2335–36.
7. Dimick, C. "Documentation Bad Habits: Shortcuts in Electronic Records Pose Risk." *Journal of the American Health Information Management Association* 79, no. 6 (2008): 40–43; quiz 45–46.
8. O'Malley, K. J., K. F. Cook, M. D. Price, et al. "Measuring Diagnoses: ICD Code Accuracy." *Health Services Research Journal* 40, no. 5, pt. 2 (2005): 1620–39.
9. Surján, G. "Questions on Validity of International Classification of Diseases–coded Diagnoses." *International Journal of Medical Informatics* 54, no. 2 (1999): 77–95.
10. Iezzoni, L. I. "Assessing Quality Using Administrative Data."
11. Ibid.
12. Iezzoni, L. I. "Using Administrative Diagnostic Data to Assess the Quality of Hospital Care: Pitfalls and Potential of ICD-9-CM."
13. Centers for Disease Control and Prevention, National Center for Health Statistics. *International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)*. Available at <http://www.cdc.gov/nchs/icd/icd9cm.htm>.
14. Hsia, D. C., W. M. Krushat, A. B. Fagan, J. A. Tebbutt, and R. P. Kusserow. "Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-Payment System." *New England Journal of Medicine* 318, no. 6 (1988): 352–55.
15. Hsia, D. C. "Accuracy of Medicare Reimbursement for Cardiac Arrest." *Journal of the American Medical Association* 264, no. 1 (1990): 59–62.
16. Lloyd, S. S., and J. P. Rissing. "Physician and Coding Errors in Patient Records." *Journal of the American Medical Association* 254, no. 10 (1985): 1330–36.
17. Campos-Outcalt, D. E. "Accuracy of ICD-9-CM Codes in Identifying Reportable Communicable Diseases." *Quality Assurance and Utilization Review* 5, no. 3 (1990): 86–89.
18. O'Malley, K. J., K. F. Cook, M. D. Price, et al. "Measuring Diagnoses: ICD Code Accuracy."
19. Gibson, N., and S. A. Brideman. "A Novel Method for the Assessment of the Accuracy of Diagnostic Codes in General Surgery." *Annals of the Royal College of Surgeons of England* 80, no. 4 (1998): 293–96.
20. Hagen, E. M., T. Rekan, N. E. Gilhus, and M. Gronning. "Diagnostic Coding Accuracy for Traumatic Spinal Cord Injuries." *Spinal Cord* 47, no. 5 (2009): 367–71.

21. Ginde, A. A., C.-L. Tsai, P. G. Blanc, and C. A. Camargo. "Positive Predictive Value of ICD-9-CM Codes to Detect Acute Exacerbation of COPD in the Emergency Department."
22. Hsia, D. C. "Accuracy of Medicare Reimbursement for Cardiac Arrest."
23. Quach, S., C. Blais, and H. Quan. "Administrative Data Have High Variation in Validity for Recording Heart Failure." *Canadian Journal of Cardiology* 26, no. 8 (2010): 306–12.
24. McMahon, L. F., and H. L. Smits. "Can Medicare Prospective Payment Survive the ICD-9-CM Disease Classification System?" *Annals of Internal Medicine* 104, no. 4 (1986): 562–66.
25. Kern, E. F. O., M. Maney, D. R. Miller, et al. "Failure of ICD-9-CM Codes to Identify Patients with Comorbid Chronic Kidney Disease in Diabetes." *Health Services Research* 41, no. 2 (2006): 564–80.
26. Fillit, H., D. S. Geldmacher, R. T. Welter, K. Maslow, and M. Fraser. "Optimizing Coding and Reimbursement to Improve Management of Alzheimer's Disease and Related Dementias." *Journal of the American Geriatrics Society* 50, no. 11 (2002): 1871–78.
27. Fisher, E. S., F. S. Whaley, W. M. Krushat, et al. "The Accuracy of Medicare's Hospital Claims Data: Progress Has Been Made, But Problems Remain." *American Journal of Public Health* 82, no. 2 (1992): 243.
28. Ginde, A. A., C.-L. Tsai, P. G. Blanc, and C. A. Camargo. "Positive Predictive Value of ICD-9-CM Codes to Detect Acute Exacerbation of COPD in the Emergency Department."
29. Hsia, D. C., W. M. Krushat, A. B. Fagan, J. A. Tebbutt, and R. P. Kusserow. "Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-Payment System."
30. Lloyd, S. S., and J. P. Rissing. "Physician and Coding Errors in Patient Records."
31. Hsia, D. C. "Accuracy of Medicare Reimbursement for Cardiac Arrest."
32. Quach, S., C. Blais, and H. Quan. "Administrative Data Have High Variation in Validity for Recording Heart Failure."
33. McMahon, L. F., and H. L. Smits. "Can Medicare Prospective Payment Survive the ICD-9-CM Disease Classification System?"
34. Fillit, H., D. S. Geldmacher, R. T. Welter, K. Maslow, and M. Fraser. "Optimizing Coding and Reimbursement to Improve Management of Alzheimer's Disease and Related Dementias."
35. Birman-Deych, E., A. D. Waterman, Y. Yan, et al. "Accuracy of ICD-9-CM Codes for Identifying Cardiovascular and Stroke Risk Factors." *Medical Care* 43, no. 5 (2005): 480.
36. Goldstein, L. B. "Accuracy of ICD-9-CM Coding for the Identification of Patients with Acute Ischemic Stroke: Effect of Modifier Codes." *Stroke* 29, no. 8 (1998): 1602.
37. Romano, P. S., L. L. Roos, and J. G. Jollis. "Adapting a Clinical Comorbidity Index for Use with ICD-9-CM Administrative Data: Differing Perspectives." *Journal of Clinical Epidemiology* 46, no. 10 (1993): 1075.
38. West, J., Y. Khan, D. M. Murray, and K. B. Stevenson. "Assessing Specific Secondary ICD-9-CM Codes as Potential Predictors of Surgical Site Infections." *American Journal of Infection Control* 38, no. 9 (2010): 701–5.
39. Benesch, C., D. M. Witter, A. L. Wilder, et al. "Inaccuracy of the International Classification of Diseases (ICD-9-CM) in Identifying the Diagnosis of Ischemic Cerebrovascular Disease." *Neurology* 49, no. 3 (1997): 660.
40. Waikar, S. S., R. Wald, G. M. Chertow, et al. "Validity of International Classification of Diseases, Ninth Revision, Clinical Modification Codes for Acute Renal Failure." *Journal of the American Society of Nephrology* 17, no. 6 (2006): 1688–94.
41. Brown, B. B. *Delphi Process*. RAND Corporation, 1968. Available at <http://www.rand.org/pubs/papers/P3925.html>. Accessed March 21, 2011.
42. Keeney, S., H. McKenna, and F. Hasson. *The Delphi Technique in Nursing and Health Research*. Chichester, West Sussex: Wiley-Blackwell, 2011.

43. Needleman, J., P. I. Buerhaus, S. Mattke, M. Stewart, and K. Zelevinsky. "Measuring Hospital Quality: Can Medicare Data Substitute for All-Payer Data?" *Health Services Research* 38, no. 6, pt. 1 (2003): 1487–1508.
44. Ingenix. "Challenges for Coding Heart Failure." *Advance for Health Information Professionals*. May 22, 2007. Available at <http://health-information.advanceweb.com/Article/Challenges-for-Coding-Heart-Failure.aspx>. Accessed April 29, 2011.
45. American Hospital Association. *Coding Clinic for ICD-9-CM*, first quarter 2008, p. 4.
46. O'Malley, K. J., K. F. Cook, M. D. Price, et al. "Measuring Diagnoses: ICD Code Accuracy."
47. Wilson, D. D. "Five RAC Coding Targets: Demonstration Program Identified Key Areas of Improper Payment." *Journal of the American Health Information Management Association* 80, no. 5 (2009): 64–66; quiz 67–68.
48. Sturgeon, J. "Debridement Revisited." *For the Record* 21, no. 17 (2009): 30.
49. Centers for Medicare and Medicaid Services. "NHE Fact Sheet." Available at https://www.cms.gov/NationalHealthExpendData/25_NHE_Fact_Sheet.asp. Accessed April 29, 2011.

Figure 1

Formula for Calculating the Weighted Rank Order of ICD-9-CM Categories

$$\text{Weighted Total} = Wc \times \text{Cost} + Wv \times \text{Volume} + Wcf \times \text{Confusion}$$

Wc: weight for cost of case

Wv: weight for volume of case

Wcf: weight for confusion

Table 1**Mean and Two Standard Deviations of Weights of Criteria after Two Rounds of the Delphi Process**

	Cost (mean % ± 2 SD)	Volume (mean % ± 2 SD)	Confusion (mean % ± 2 SD)
Round 1	23.25 ± 12.6	32 ± 6.3	44.75 ± 9.5
Round 2	18.9 ± 11.0	32.8 ± 5.5	48.3 ± 6.9

Table 2

Mean and Two Standard Deviations of Rank Order of Confusion Rate after Two Rounds of the Delphi Process

Subcategories	Round 1		Round 2		Final Rank
	Confusion by Documentation Ranking (Mean \pm 2 SD)	Confusion by Guideline Ranking (Mean \pm 2 SD)	Confusion by Documentation Ranking (Mean \pm 2 SD)	Confusion by Guideline Ranking (Mean \pm 2 SD)	
Excisional debridement	2.8 \pm 5.0	2.8 \pm 7	1.0 \pm 0	1.0 \pm 0	1
Sepsis/bacteremia	3.8 \pm 2.0	4.8 \pm 3.0	3.4 \pm 1.0	3.6 \pm 7.2	2
Complication	3.0 \pm 2.8	8.0 \pm 7.2	2.2 \pm 0.8	5.2 \pm 4.4	3
Congestive heart failure	4.0 \pm 6.6	5.7 \pm 11.2	4.6 \pm 4.4	4.4 \pm 6.2	4
Epilepsy and seizures	10.0 \pm 8.2	9.3 \pm 12.2	9.2 \pm 3.0	7.4 \pm 4.6	5
COPD	9.7 \pm 10.4	10.2 \pm 6.1	8.6 \pm 5.0	9.2 \pm 8.2	6
Respiratory failure	11.2 \pm 6.6	9.2 \pm 5.8	10.8 \pm 3.8	8.2 \pm 0.8	7
Diabetes mellitus	9.3 \pm 9.4	11.5 \pm 11.6	7.4 \pm 2.2	13.2 \pm 3.6	8
SIRS with cellulitis	12.7 \pm 4.6	8.8 \pm 5.4	14.2 \pm 2.6	7.2 \pm 0.8	9
Renal insufficiency	10.5 \pm 9.6	12.0 \pm 4.6	10.0 \pm 7.8	14.8 \pm 2.2	10
Arterial embolization	15.5 \pm 5.4	11.0 \pm 12.1	18.0 \pm 0	9.2 \pm 8.2	11
Chronic kidney disease	11.8 \pm 7.6	12.8 \pm 5.4	13.4 \pm 1.8	16.2 \pm 0.8	12
Malignant ascites	13.3 \pm 7.8	13.7 \pm 7.6	14.2 \pm 8.0	18.0 \pm 0	13

Note: COPD: chronic obstructive pulmonary disease; SIRS: systemic inflammatory response syndrome.

Table 3

Final Weighted Rank Order of the ICD-9-CM Code Subcategories

Subcategory	Number of Cases	Volume Rank	Overall Financial Impact	Cost Rank	Confusion Rate Rank	Weighted Total	Weighted Total Rank
Complications	672,594	5	\$257,099,062	5	3	4.034	1
Congestive heart failure	2,320,629	2	\$100,475,422	8	4	4.1	2
Excisional debridement	100,646	12	\$576,461,896	4	1	5.175	3
Sepsis/bacteremia	512,632	8	\$29,235,465	10	2	5.48	4
Diabetes mellitus	3,271,828	1	\$33,252,579	9	8	5.893	5
COPD	632,026	6	\$179,653,514	6	6	6	6
Respiratory failure	538,248	7	\$948,974,508	3	7	6.244	7
Renal insufficiency versus failure	1,924,071	4	\$1,898,106,569	1	10	6.331	8
Chronic kidney disease	1,981,013	3	\$1,657,965,276	2	12	7.158	9
Epilepsy and seizures	478,038	9	\$8,916,045	12	5	7.635	10
SIRS with cellulitis	299,454	10	\$13,729,966	11	9	9.706	11
Arterial embolization	127,026	11	\$138,279,695	7	11	10.244	12
Malignant ascites	0	13	\$0	13	13	13	13

Note: COPD: chronic obstructive pulmonary disease; SIRS: systemic inflammatory response syndrome.