A Reliable Billing Method for Internal Medicine Resident Clinics: Financial Implications for an Academic Medical Center

Suraj Kapa, MD
Thomas J. Beckman, MD
Stephen S. Cha, MS
Joyce A. Meyer
Charlotte A. Robinet
Diane K. Bucher
Jeanne M. Hardy
Furman S. McDonald, MD, MPH

Abstract

Background The financial success of academic medical centers depends largely on appropriate billing for resident-patient encounters. Objectives of this study were to develop an instrument for billing in internal medicine resident clinics, to compare billing practices among junior versus senior residents, and to estimate financial losses from inappropriate resident billing.

Methods For this analysis, we randomly selected 100 patient visit notes from a resident outpatient practice. Three coding specialists used an instrument structured on Medicare billing standards to determine appropriate codes, and interrater reliability was assessed. Billing codes were converted to US dollars based on the national Medicare reimbursement list. Inappropriate billing, based on comparisons with coding specialists, was then determined for residents across years of training.

Results Interrater reliability of *Current Procedural Terminology* components was excellent, with κ ranging from 0.76 for examination to 0.94 for diagnosis. Of the encounters in the study, 55% were underbilled by an average of $45.26 per encounter, and 18% were overbilled by an average of $51.29 per encounter. The percentages of appropriately coded notes were 16.1% for postgraduate year (PGY) 1, 26.8% for PGY-2, and 39.3% for PGY-3 residents ($P < .05$). Underbilling was 74.2% for PGY-1, 48.8% for PGY-2, and 42.9% for PGY-3 residents ($P < .01$). There was significantly less overbilling among PGY-1 residents compared with PGY-2 and PGY-3 residents (9.7% versus 24.4% and 17.9%, respectively; $P < .05$).

Conclusions Our study reports a reliable method for assessing billing in internal medicine resident clinics. It exposed large financial losses, which were attributable to junior residents more than senior residents. The findings highlight the need for educational interventions to improve resident coding and billing.

Background

The financial success of academic medical centers depends on appropriate billing for encounters between resident physicians and patients. This is particularly true in resident physicians' primary care clinics, where resident-patient encounters generate the majority of revenue. Many academic medical centers require that attending physicians bill for resident-patient encounters. However, some institutions, including the Mayo Clinic, allow licensed residents to determine billing codes, with attending physicians providing only supervisory assistance. In all circumstances, residents must learn to determine accurate billing codes to practice independently after graduation from training.

Billing for patient encounters is determined by applying complex Medicare rules to determine *Current Procedural Terminology* (*CPT*) codes. The *CPT* requires documentation for several billing components.[1] Specifically, the final billing level is determined by the sum of 3 major components, which in turn are composed of 7 subcomponents. This system becomes more complex when applying additional rules regarding whether patients have been seen in the same clinic within 3 years, whether encounters reflect consultation or primary care visits, and whether visits constitute preventive care versus so-called

All authors are at the Mayo Clinic, Rochester, Minnesota. During the study, **Suraj Kapa, MD,** was a third-year internal medicine resident at the Mayo Clinic; he is now a cardiology fellow at University of Pennsylvania. **Thomas J. Beckman, MD,** and **Furman S. McDonald, MD, MPH,** are with the Department of Internal Medicine; **Stephen S. Cha, MS,** is with the Department of Biostatistics, and **Joyce A. Meyer, Charlotte A. Robinet, Diane K. Bucher,** and **Jeanne M. Hardy** are with the Department of Finance.

This study was supported by the Mayo Clinic Research Committee Small Grants Program administered by the Division of General Internal Medicine and by the Mayo Clinic Internal Medicine Residency Office of Educational Innovations as part of the ACGME Educational Innovations Project.

Corresponding author: Furman S. McDonald, MD, MPH, Department of Internal Medicine, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, 507.255.8719.

Received January 2, 2010; revision received April 11, 2010; accepted April 22, 2010.

DOI: 10.4300/JGME-D-10-00001.1

comprehensive general medical examinations. Adding to this complexity, the final billing level may reflect face-to-face encounter time.

Billing mistakes include overbilling and underbilling. Underbilling can have significant financial repercussions, and both underbilling and overbilling are considered Medicare fraud.[2] Billing errors may result in costly audits and legal consequences. Therefore, it is essential that attending physicians' documentation reflects the level of complexity suggested by the encounter and that residents receive adequate coding education to provide appropriate billing codes when entering into practice.

Prior studies[3–5] indicated high variability for billing codes based on documentation review among coding specialists. If coding specialists are unreliable when assigning codes, then even greater unreliability may occur among less experienced busy resident physicians. Therefore, we developed and tested the reliability of a tool for determining billing codes within a large academic medical center. Based on these findings, we determined the frequency of billing errors (underbilling and overbilling) among internal medicine residents, the financial implications of these errors, and whether the frequency of billing errors differs among junior versus senior internal medicine residents.

Methods

Development and Pilot Testing of the Coding Instrument

An instrument was developed to determine the final billing code based on the standard Medicare rule set for billing levels. The instrument assessed the 3 major billing components (history, physical examination, and medical decision making) and the corresponding 7 subcomponents (history of present illness; review of systems; medical and surgical history; and physical examination diagnosis, data, and risk). As a pilot test, 3 coding specialists used the instrument to score 10 notes. The specialists then discussed their assigned codes to understand their discrepancies. Based on this process, it was determined that the instrument (a single sheet of paper) was sufficient and required no additional changes.

Checklists for the first 2 major billing components (history of present illness and physical examination) are used to determine the appropriate coding level. The coding sheet also includes rules for deciding appropriate billing levels for each of the subcomponents. Consequently, the instrument would be usable even by novice coders with limited medical knowledge and experience.

Resident Note Selection

We randomly selected 100 resident notes (31 postgraduate year [PGY] 1, 41 PGY-2, and 28 PGY-3) from all patient visits to the internal medicine resident outpatient continuity clinic at the Mayo Clinic. Notes were selected from the middle of the academic year to ensure that all residents had prior exposure to both continuity clinic and the billing process. This study was approved by the Mayo Clinic Institutional Review Board.

Coding Instrument Assessment

The coding instrument consisted of the following scales. Under *history* and *physical examination*, 3 subcomponents were scored from 1 to 4 based on whether they were "problem focused," "extended problem focused," "detailed," or "comprehensive." Under *medical decision making*, risk was scored from 1 to 4, ranging across the options of "minimal," "low," "moderate," or "high." Also under *medical decision-making*, *diagnosis* and *data* scores were obtained based on a summary score, which in turn was based on responses to all other elements of the coding tool. The "nature of the presenting problem" falls under the "risk" portion of medical decision making. This ranges from a self-limited or minor problem to multiple acute or chronic illnesses demonstrating severe progression or posing a threat to life.

Three independent raters (2 coding specialists [raters 1 and 2] and an internal medicine resident [rater 3]) used the instrument to score resident documents of 100 patient encounters. Raters were blinded to the resident provider, patients, and previously assigned billing codes. Furthermore, no rater had access to the billing assessment of any other rater. Raters utilized the coding instrument to score each billing component, to determine whether a preventive medicine modifier was indicated, and to determine the overall billing code.

Description of *CPT* Codes

Outpatient *CPT* codes are listed as 5 digit numbers, and define patient encounters as consult (physician requested) or primary care (nonphysician requested) visits. Primary care visits are further subdivided into new patients or established patients. Billing level is determined by either computing face-to-face time or by meeting the individual components. To enhance the feasibility of scoring, 5-digit number codes were converted to alphanumeric codes (99201–99205 to N1–N5, 99211–99215 to E1–E5, and 99241–99245 to P1–P5). Overall, billing corresponded to an E code (established), an N code (new), or a P code (consult). Finally, preventive medicine modifiers were assigned, when appropriate, to patients undergoing comprehensive general medical examinations.

Statistical Analysis

Data from the 3 raters were organized according to the 7 billing subcomponents, indication for a preventive medicine modifier (ie, whether the examination met criteria for a comprehensive general medical evaluation), and final billing code. The interrater reliability of coding instrument scores was determined by calculating simple κ for billing subcomponents, preventive medicine modifiers, and overall billing codes. κ coefficients were interpreted according to

TABLE 1 **Medicare Billing Component Assessments by Coding Specialists**[a]

Component	3 Agree and 0 Disagree (n = 80)	2 Agree and 1 Disagree (n = 20)	All Notes (N = 100)
History			
Present illness	3.6 ± 0.7	3.4 ± 0.9	3.6 ± 0.8
Review of systems	3.2 ± 0.5	3.1 ± 0.6	3.2 ± 0.5
Medical and surgical	3.3 ± 0.7	3.3 ± 0.6	3.3 ± 0.7
Physical examination	3.0 ± 1.0	2.7 ± 0.9	3.0 ± 0.7
Medical decision making			
Diagnosis	3.5 ± 0.8	3.1 ± 1.0	3.4 ± 0.9
Data	1.5 ± 0.8	1.3 ± 0.7	1.5 ± 0.8
Risk	2.8 ± 0.4	2.5 ± 0.6	2.8 ± 0.5
Overall			
Charge, $ US	152.52 ± 48.07	156.88 ± 69.63	153.22 ± 52.00

Abbreviation: *CPT, Current Procedural Terminology.*

[a] The mean scores for each of 7 components comprising the overall *CPT* code based on the reviews of all 3 independent raters. See "Methods" section for how the numerical scores were reached. Also included is the average cost per encounter based on the assessment of 3 raters. Note that the only significant difference between groups was the risk component of medical decision making, for which there was a significantly lower risk attributed to encounters involving disagreement versus encounters involving complete agreement ($P = .008$).

method by Landis and Koch,[6] where less than 0.40 suggests poor, 0.4 to 0.75 represents fair to good, and greater than 0.75 indicates excellent reliability. In addition, billing codes assigned by the residents were abstracted from the electronic billing record for the same encounters that were coded by the specialists. All billing codes were converted to US dollars based on the national Medicare reimbursement list.[7]

For all 100 notes, we compared the codes assigned by the 3 coding specialists with the codes assigned by the residents. The gold standard was defined as all notes for which at least 2 of 3 raters agreed on the final billing code. These gold standard codes were then compared with those entered by the residents at the time of the encounter. The cost translation for these overbilled and underbilled notes was determined by summing the total differences in cost for all notes to determine an average cost of underbilling or overbilling per note.

Group comparisons for overbilling or underbilling across years of training (PGY-1, PGY-2, and PGY-3) were determined using Kruskal-Wallis one-way analysis of variance. Fisher exact test was used to determine if there were significant differences in the numbers of providers who had assigned the appropriate billing code.

Results

Note Selection

Breakdown of the 100 randomly selected notes (by number) were as follows: 31 PGY-1, 41 PGY-2, and 28 PGY-3. Combined, these notes included 5 consults, 73 established patients, and 22 patients who presented to establish primary care. PGY-1 residents do not have consult-based rotations, so all consult notes were documented by PGY-2 (3 consults) and PGY-3 residents (2 consults). Postgraduate year 1 encounters comprised 25 established and 6 new patients. Postgraduate year 2 encounters comprised 27 established, 11 new, and 3 consult patients. Postgraduate year 3 encounters comprised 21 established, 5 new, and 2 consult patients.

Instrument Score Reliability

Coding specialists universally agreed on whether the encounter met criteria for an N code, an E code, or a P code and whether a preventive medicine modifier should be applied to every encounter (which in this study accounted for 15 of 100 encounters). For 80 of 100 notes, all 3 specialists agreed on the final code. For the remaining 20, there was agreement between 2 raters and disagreement by 1. Rater 1 was in dissension on 7 of these notes, rater 2 on 5, and rater 3 on 8.

TABLE 1 summarizes the average billing level of each component for all 100 notes, including the 80 notes for which all 3 raters agreed and the 20 notes for which 1 of 3 raters disagreed. Also included is the encounter charge based on final *CPT* codes and, when indicated, the assignment of a preventive medicine modifier. The only significant difference between the groups was the risk component of medical decision making, for which there was a significantly lower risk attributed to encounters involving disagreement versus encounters involving complete

TABLE 2 **Interrater Reliability for 3 Expert Coding Specialists**[a]

Component	κ Statistic	95% Confidence Interval
History		
Present illness	0.86	0.78–0.89
Review of systems	0.94	0.92–0.96
Medical and surgical	0.93	0.89–0.94
Physical examination	0.76	0.66–0.83
Medical decision making		
Diagnosis	0.93	0.90–0.95
Data	0.94	0.91–0.95
Risk	0.91	0.87–0.94
Overall		
CPT code	1.00	1.00–1.00
Preventive medicine modifier	1.00	1.00–1.00
Charge, $ US	0.98	0.97–0.98

Abbreviation: *CPT, Current Procedural Terminology.*

[a] Interrater reliabilities for all 3 expert coders for each of 7 *CPT* code components, the decision on preventive medicine modifier inclusion, and the final *CPT* code and encounter cost across all 100 notes.

agreement ($P = .008$). Otherwise, no significant between-group differences were observed for the other components or charges.

TABLE 2 summarizes agreement across 3 raters for the billing codes (components and overall), preventive medicine modifier, and cost. Notably, κ exceeds 0.70 for all elements, indicating excellent interrater reliability for the billing assessment method used in this study.

Assessment of Resident Coding

Gold standard codes determined by 3 independent raters were compared with the resident codes. Among encounters for which only 1 rater disagreed (n = 20), the code the other 2 raters agreed on was used as the gold standard. Charges were determined on the basis of these final codes and whether or not a preventive medicine modifier was applied to the encounter.

FIGURE 1 summarizes the percentage of encounters falling into each billing level based on the gold standard and resident codes. There were 2 encounters for which only the preventive medicine modifier was assigned without any additional encounter code. Both met criteria for an E3, or 99213, per the gold standard codes.

Of 100 encounters, only 27% were appropriately coded by the residents. Of 73 inappropriately coded encounters, 12 were attributed to an incorrect prefix (ie, E, N, or P) (FIGURE 2). Also shown is the preventive medicine modifier, which should have been applied to 15 encounters (8 PGY-1, 3 PGY-2, and 4 PGY-3) but was only assigned to 6 (3 PGY-1, 1 PGY-2, and 2 PGY-3). The preventive medicine modifier was never inappropriately added to an encounter by residents. Most residents correctly coded E codes and P codes (E codes were not appropriately coded by 2 residents, and P codes were not appropriately coded by 1 resident). The most frequent inappropriate coding occurred with N codes, for which only 13 of 22 encounters (4 of 6 PGY-1, 6 of 11 PGY-2, and 3 of 5 PGY-3) were coded appropriately.

FIGURE 3 summarizes coding outcomes by residency year, showing that 55% of notes in the study resulted in underbilling by an average of $45.26 per encounter and that 18% resulted in overbilling by an average of $51.29 per encounter. All 9 encounters for which a preventive medicine modifier was not coded fell within the underbilled group. Of these 9 encounters, 8 were otherwise correctly coded, and 1 also listed an N code inappropriately as an E code. The 1 inappropriately coded consult (P code) also fell within the underbilled group. In total, 16.4% of underbilled notes could be accounted for by lack of a preventive medicine modifier and 16.4% by an incorrect prefix code. Of the overbilled encounters, 3 could be accounted for by inappropriate prefix codes.

Underbilling significantly decreased over the PGY-1 through PGY-3 years ($P = .005$). Conversely, appropriate coding ($P = .037$) and overbilling ($P = .049$) increased over the PGY-1 through PGY-3 years. The average charge per encounter by which underbilling and overbilling occurred did not significantly change between residency years.

Estimated Financial Losses From Billing Errors

Total lost charges across all 100 encounters were estimated at $1786.73. On average, 100 to 110 residents (equally distributed among PGY years) have their continuity clinic in a given week, with PGY-1 residents seeing on average 3 to 5 patients per clinic session and PGY-2 and -3 residents seeing on average 5 to 7 patients. On the basis of our analysis, a PGY-1 will tend to underbill each encounter by $35.76, a PGY-2 by $9.37, and a PGY-3 by $11.28. Over the academic year, each PGY-1 resident will have 33 continuity clinics and see 4 patients per clinic session. Each PGY-2 and PGY-3 resident will have an average of 42 and 34 continuity clinics, respectively, and see an average of 6 patients per clinic. Consequently, we estimate that in a residency class of 48 residents, approximately $8660.94 in underbilling occurs weekly in the resident-based outpatient continuity clinic. Assuming equal patterns of underbilling and overbilling throughout the 52-week academic year, this would constitute an annual total of $450,368.64 in lost charges, of which 50% is attributable to inappropriate coding at the PGY-1 level.

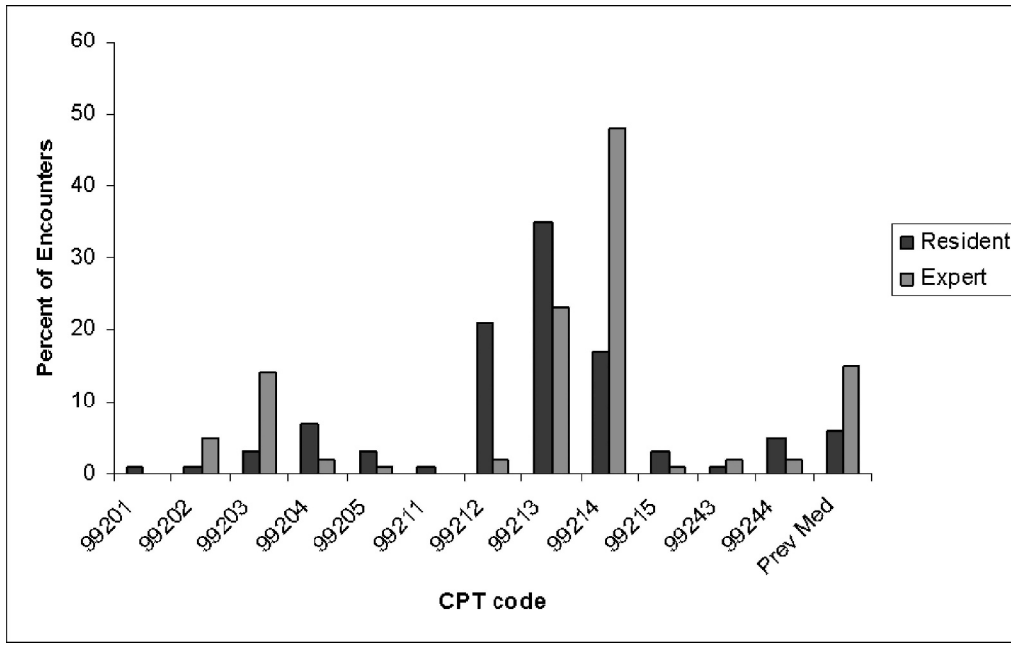

FIGURE 1 **Distribution of Notes as Coded by the Expert Observers and Residents for Each *Current Procedural Terminology* Code**

Note: Two encounters were coded by residents using the preventive medicine modifier only; 15% of encounters met criteria for preventive medicine, but only 6% were actually coded for by residents.

Discussion

We report a reliable method for assigning billing codes within a resident physician clinic at a large academic medical center. Our method assisted in estimating potentially large financial losses, which seemed more attributable to junior residents than senior residents. These findings have important financial and educational implications for internal medicine residency programs.

Our findings add to previous studies that mainly involved family medicine practices. Horner et al[8] found that billing errors are common at the resident level and at the faculty level. Another study[5] of 84 family practices in Ohio reported that billing errors were most common for new patient visits and that underbilling and overbilling occurred at similar frequencies. Findings from these studies suggest that the problems associated with current coding and billing extend beyond practitioners to include systems and the need to interpret complex rules. While previous investigations have demonstrated that billing errors are common, we believe that our study is the first to reveal coding disparities between junior and senior internal medicine residents and the financial implications of underbilling at an academic medical center.

Previous studies also showed that physicians are less accurate than expert auditors when assigning billing levels for patient visits[4] and that physicians are particularly less accurate than experts when determining codes for new patients.[3] However, research has shown that even professional coding specialists tend to underbill and have poor interrater agreement, ranging from 50% to 71%.[4] We describe a coding method that allows excellent interrater agreement among a small team consisting of coding experts and an internal medicine resident and that, based on its simplicity, may provide a framework for training resident and faculty physicians regarding accurate coding.

Indeed, internal medicine residents report a substantial need for improved training on outpatient Medicare billing during residency.[9] In addition, a study[10] of pediatric residents revealed low agreement for outpatient visit coding, leading the authors to suggest that inadequate training could cause large financial losses. In response to the need for improved training, authors have described problem-based learning curricula.[11] Furthermore, investigators reported that the use of lectures combined with specialized history and physical examination forms might improve billing accuracy, thereby increasing billable income.[12] However, none of these studies reported the reliability of these curricular interventions, nor did they estimate an actual savings.

Our method draws on abundant validity evidence.[13,14] Specifically, content validity is supported by the use of coding elements based on Medicare guidelines, review by a panel of coding experts, and clarification through pilot testing. Internal structure validity was supported by high interrater reliability across all coding elements. Consequences validity, although incomplete, was supported by applying the

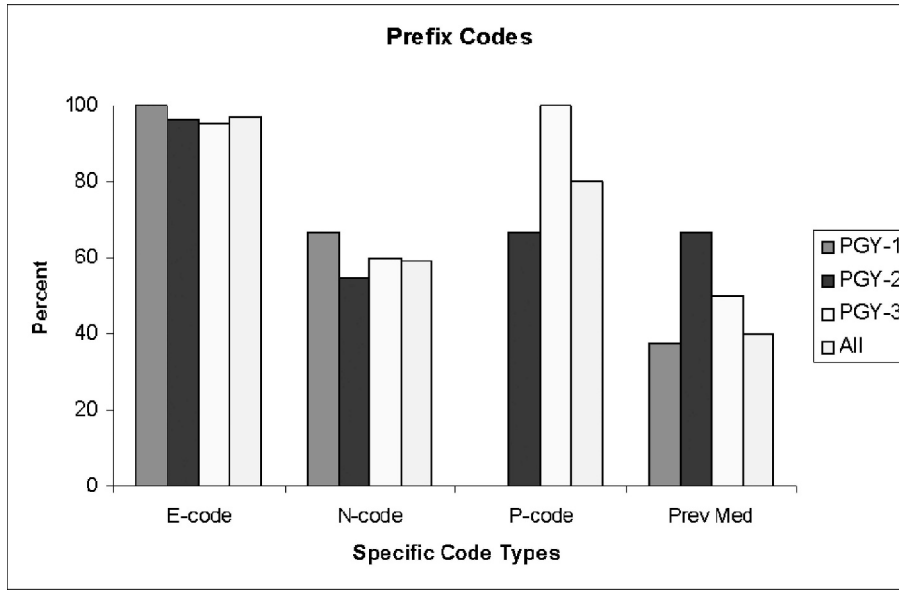

FIGURE 2 **Percentage of Resident-Coded Notes Appropriately Coded With E Code, P Code, or N Code Based on the Gold Standard**

Note: E code indicates established patient visit; N code, new patient visit; and P code, consultation visit. Shown is the percentage of encounters by resident year that were coded with the appropriate prefix code and for which a preventive medicine modifier was included when appropriate.

gold standard coding method to expose financial losses; if this method ultimately leads to improved coding accuracy and savings, then validity will also be attained.

Our study has several limitations. First, although interrater reliability for individual coding elements was excellent, all 3 raters agreed on the final code 80% of the time, with 2 of 3 agreeing the other 20%. However, this level of reliability still exceeds that reported in previous studies.[3–5] Second, the reliability of our method depended on thoroughly developing and applying the coding scale, which

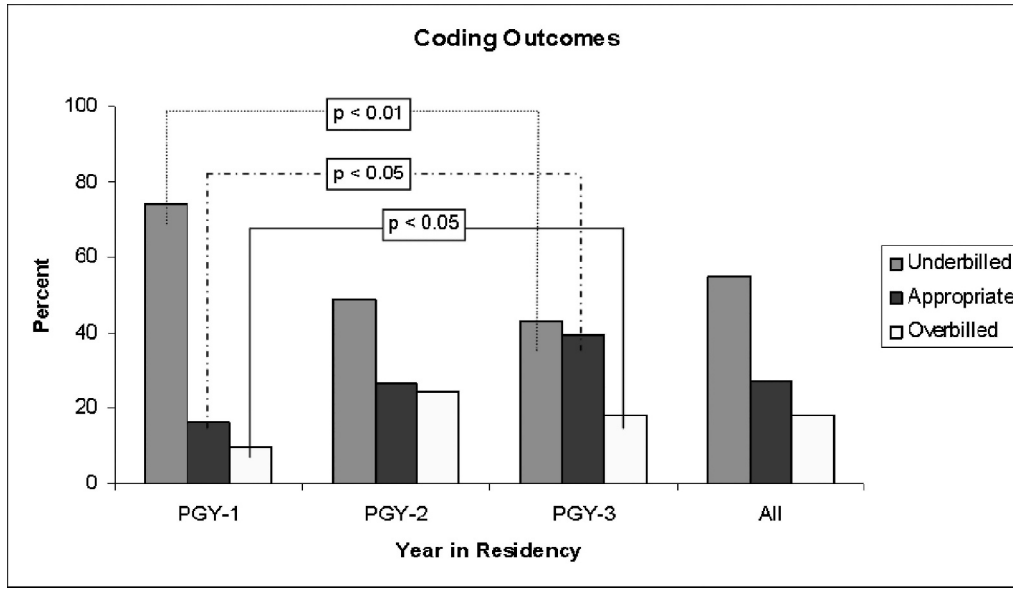

FIGURE 3 **Distribution of Underbilling, Overbilling, and Appropriate Billing by Residency Year**

Note: There were significant decreases in underbilling ($P < .01$) and significant increases in appropriate billing ($P < .05$) and overbilling ($P < .05$) from postgraduate year (PGY) 1 to PGY-3.

suggests that replication of our findings might require the same degree of rigor. Nonetheless, such iterative processes are essential when developing reliable and valid scales.[15] Third, coding consistency was not compared with what could have been achieved had no form been used, although the consistency achieved using this tool was greater than what has been previously reported. Fourth, this is a single-institution study, and nuances of the practice at our institution may limit generalization.

Billing by residents in our study generally improved throughout training. These findings reflect research showing that billing accuracy may naturally improve as physicians mature professionally and learn more about appropriate billing.[16] Nonetheless, our study also demonstrated overbilling by almost 20%, which may result in loss of reimbursement for patient encounters and have serious repercussions for providers. Consequently, our findings have implications for trainees at all levels, including senior medical residents who are soon to graduate and enter practice.

In summary, we describe a Medicare-based billing method that is feasible and reliable when used by 3 coders. Our method exposed substantial potential financial losses from underbilling in an outpatient internal medicine resident practice and found that most financial losses are attributable to billing errors by junior residents. These findings suggest the need to target billing curricula at junior residents, perhaps by incorporating a simple instrument like that used in this study. Future research should determine whether our instrument is an effective instructional tool that leads to favorable financial outcomes.

References

1 Iezzoni LI. The demand for documentation for Medicare payment. *N Engl J Med*. 1999;341:365–367.
2 *Medicare Program Integrity Manual*. Baltimore, MD: Centers for Medicare and Medicaid Services, Dept of Health and Human Services. December 12, 2001.
3 King MS, Sharp L, Lipsky MS. Accuracy of *CPT* evaluation and management coding by family physicians. *J Am Board Fam Pract*. 2001;14:184–192.
4 King MS, Lipsky MS, Sharp L. Expert agreement in *Current Procedural Terminology* evaluation and management coding. *Arch Intern Med*. 2002;162:316–320.
5 Kikano GE, Goodwin MA, Stange KC. Evaluation and management services: a comparison of medical record documentation with actual billing in community family practice. *Arch Fam Med*. 2000;9:68–71.
6 Landis JR, Koch GG. The measure of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
7 Centers for Medicare and Medicaid Services. Overview: physician fee schedule look-up. 2008. http://www.cms.hhs.gov/PFSlookup/. Accessed May 18, 2010.
8 Horner RD, Paris JA, Purvis JR, Lawler FH. Accuracy of patient encounter and billing information in ambulatory care. *J Fam Pract*. 1991;33:593–598.
9 Adiga K, Buss M, Beasley BW. Perceived, actual and desired knowledge regarding Medicare billing and reimbursement: a national needs assessment survey of internal medicine residents. *J Gen Intern Med*. 2006;21:466–470.
10 Ng M, Lawless ST. What if pediatric residents could bill for their outpatient services? *Pediatrics*. 2001;108:827–834.
11 As-Sanie S, Zolnoun D, Wechter ME, Lamvu G, Tu F, Steege J. Teaching residents coding and documentation: effectiveness of a problem-oriented approach. *Am J Obstet Gynecol*. 2005;193:1790–1793.
12 Sprtl SJ, Zlabek JA. Does the use of standardized history and physical forms improve billable income and resident physician awareness of billing codes? *South Med J*. 2005;98:524–527.
13 Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med*. 2005;20:1159–1164.
14 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119:166e167–166e116.
15 Beckman TJ. Determining the validity and reliability of clinical assessment scores. In: Henderson MC, ed. *A Textbook for Internal Medicine Education Programs*. 9th ed. Washington, DC: Association of Program Directors in Internal Medicine and Association of Specialty Professors; 2007:139–146.
16 Lasker RD, Marquis MS. The intensity of physicians' work in patient visits: implications for the coding of patient evaluation and management services. *N Engl J Med*. 1999;341:337–341.