Viewpoint Paper ■

Comparing Paper-based with Electronic Patient Records: Lessons Learned during a Study on Diagnosis and Procedure Codes

JÜRGEN STAUSBERG, PRIV-DOZ DR MED, DIETRICH KOCH, JOSEF INGENERF, DR RER NAT, MICHAEL BETZLER, PROF DR MED

Abstract Paper-based and electronic patient records generally are used in parallel to support different tasks. Many studies comparing their quality do not report sufficiently on the methods used. Few studies refer to the patient. Instead, most regard the paper record as the gold standard. Focusing on quality criteria, the current study compared the two records patient by patient, presuming that each might hold unique advantages. For surgical patients at a nonuniversity hospital, diagnosis and procedure codes from the hospital's electronic patient record (EPR set) were compared with the paper records (PPR set). Diagnosis coding from the paper-based patient record resulted in minor qualitative advantages. The EPR documentation showed potential advantages in both quality and quantity of procedure coding. As in many previous studies, the current study relied on a single individual to extract and transform contents from the paper record to compare PPR with EPR. The exploratory study, although limited, supports previous views of the complementary nature of paper and electronic records. The lessons learned from this study are that medical professionals should be cognizant of the possible discrepancies between paper and electronic information and look toward combining information from both records whenever appropriate. The inadequate methodology (transformations done by a single individual) used in the authors' study is typical of other studies in the field. The limited generalizability and restricted reproducibility of this commonly used approach emphasize the need to improve methods for comparing paper-based with electronic versions of a patient's chart.

■ **J Am Med Inform Assoc.** 2003;10:470–477. DOI 10.1197/jamia.M1290.

The authors report a "case"— their own imperfect study with its techniques and results—and then review the literature to illustrate how specific methodologic issues traditionally hinder the comparison of paper-based (PPRs) and electronic patient records (EPRs).

The electronic patient record has not yet fully replaced the paper-based one.[1,2] Rather, electronic documentation usually is used in addition to residual paper-based records. One might assume that the electronic data represent a subset of the patient data stored in the paper-based record. However, Mikkelsen and Aasly[3] found that "parallel use of electronic and paper-based patient records result[ed] in inconsistencies between the record systems" and "documentation [was] missing in both." The paper-based patient record is still the main source for information management in daily care delivery for several reasons. Utilization of the paper-based patient record, both as a reminder to health care providers to report events, such as the course of an illness, and as a tool for communication among clinicians, has already been documented in the literature.[4,5] The German legal system treats the paper-based patient record preferentially. Health insurance companies use the paper record to evaluate appropriateness of admission and length of stay. Conversely, electronic data storage is used for legislatively obliged standardized and structured documentation and reporting. This is true in Germany regarding communication between hospitals and health insurance companies; case grouping for hospital fees; data acquisition for national hospital statistics; and, in 2003, the introduction of diagnosis-related groups (DRGs), which particularly focus attention on grouping cases using the EPR.

Typically, a different level of information is present in each type of record. The paper-based record consists chiefly of unstructured or less-structured free text. The highly standardized "data abstract" component of the EPR provides structured elements and a controlled vocabulary. Furthermore, it consists of standard codes for classifications in main parts. To study both records' contents comparatively, researchers must transform the records into a common representation. One way to accomplish this is through retrospective coding of information from the paper-based record, as shown in Figure 1, and as used in our study. The focus of the authors' own investigation was to determine the

Affiliations of the authors: Institute for Medical Informatics, Biometry and Epidemiology, Medical Faculty, University of Duisburg-Essen, Germany (JS); Department of General, Trauma and Vascular Surgery, Alfried Krupp Hospital, Essen, Germany (DK, MB); Institute for Medical Informatics, University of Lübeck, Germany (JI).

The authors thank the *JAMIA* editorial staff, who advised them to focus on methodologic issues in study design and gave valuable support in further refining this paper. The authors also thank Frank Stellmacher, who provided them with the tool for the documentation of the PPR set. Preliminary results of this study were presented September 9, 2002, at the 47th Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology in Berlin, Germany.

Correspondence and reprints: Priv.-Doz. Dr. med. Jürgen Stausberg, Institute for Medical Informatics, Biometry and Epidemiology, Medical Faculty, University of Duisburg-Essen, Hufelandstr. 55, D-45122 Essen, Germany; e-mail: <stausberg@uni-essen.de>.

Received for publication: 11/11/02; accepted for publication: 04/19/03.

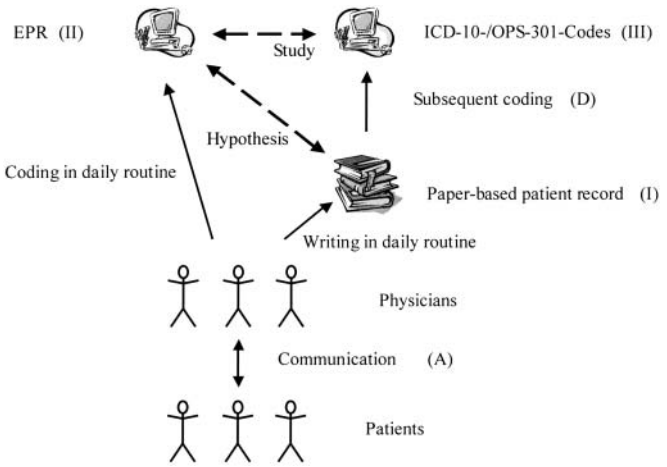

Figure 1. Overview of the different types of records (Roman numerals) and process steps (upper case) relevant for this study.

validity of EPR-based ICD-10-/OPS-301 codes as an equivalent to paper-based patient records. How to accomplish this is a crucial issue for the generalizability and applicability of the results for all studies, not just the one reported.

Inconsistencies between a patient's electronic and paper-based medical record can lead to significant problems for the health care staff in daily practice. Comparative studies are therefore necessary. Personnel cannot base their decisions on one record type alone if the two differ. For example, a physician working the night shift may deal with an established patient who is unknown to the physician. The physician must check the EPR in addition to the paper-based record to review all past complications and comorbidities. Medical assistants responsible for scheduling procedures will not rely on electronic information present in their scheduling tool exclusively; they must examine the paper record to be aware of relevant procedures recorded exclusively in it. This study sought to quantify discrepancies between the electronic abstract and the paper-based patient record. In addition, it critically discusses methodologic issues in the design of comparative studies, using the current work as a starting point. The authors focus on diagnosis and procedure codes, because they are generally available in an EPR.

Report of Current Study as an Example: Material and Methods

Sample: PPR Set and EPR Set

Alfried Krupp Hospital in Essen provides 540 beds and is divided into 12 departments. The sample consists of patients who were discharged from its Department of General, Trauma, and Vascular Surgery (117 beds) in September 2001. To compare paper and electronic records, the authors employed an experienced surgeon to code diagnoses and procedures in the paper-based records without knowledge of the medical data in the electronic abstracts. He followed the general coding rules,[6] which health care regulatory bodies (comprising hospital carriers and health insurance companies) published to establish a standard for coding with respect to billing with DRGs. The surgeon coded using DIACOS, a coding tool manufactured by ID GmbH (Berlin, Germany). The data were stored using Microsoft Access 2000. This work was done in December 2001, two to four months after discharge of the patients. The data derived from the paper-based patient record are denoted as the PPR set. Each individual hospital stay of a specific patient is denoted as a "case." Thus, a patient could have more than one case in this study, if he or she was discharged two or more times in September 2001.

The authors independently used the EPR to collect demographic data and related abstracted information about each patient's hospitalization. A separate hospital unit, responsible for electronic data processing, developed the EPR. The EPR is connected with the central administrative system IS-H from SAP, the laboratory management system and the picture archiving and communication system (PACS). The EPR includes information regarding operations, diagnoses, laboratory results, and reports from the radiology department, among other things. EPR diagnoses are stored as codes from a special edition of the International Classification of Diseases for inpatient care (abbreviated as ICD-10-SGB-V 2.0), and EPR procedures are stored as codes from a German adaptation of the ICPM called "Operationenschlüssel nach § 301 SGB V" version 2.0 (abbreviated as OPS-301 2.0). Normally, physicians enter diagnosis and operative procedure codes using DIACOS. When a patient is discharged, the responsible physician is confronted with the set of known diagnoses. The physician then is prompted to mark one code as the principal diagnosis. In addition, the physician is able to delete irrelevant diagnoses. The authors considered only those EPR diagnoses accepted at the time of discharge. The set of data from EPR is denoted as the EPR set. The Department of General, Trauma, and Vascular Surgery has an additional form of data control for EPR coding. An experienced surgeon checks the EPR codes for diagnoses and operative procedures entered for each case, whenever a patient is discharged. The authors use the term *electronic patient record* to indicate all kinds of electronic documentation, independent of the degree of structuring and the amount of information.

Calculation of Diagnosis-related Groups

Both sets of data, the EPR set and the PPR set, were grouped into the Australian Refined DRGs (AR-DRGs) Version 4.1[7] using DrGroup, a software program produced by Visasys Pty. Ltd., Canberra, Australia. The AR-DRGs are the basis for the German DRGs (G-DRGs), which had not been published when this study was conducted. The authors found ICD-10-SGB-V 2.0 to be comparable to the ICD-10-Australian Modifications (ICD-10-AM) first edition. As a result, the former was used for diagnoses. The OPS-301 procedure codes were translated into the Australian procedure classification, the Australian Minimum Benefits Schedule-Extended (MBS-Extended), using a translation table developed for the German self-government by the Essen Institute for Medical Informatics, Biometry, and Epidemiology. DrGroup also calculates a score for complications and comorbidities (CC) called patient clinical complexity level (PCCL). Possible values of the PCCL are 0 (no CC), 1 (minor CC), 2 (moderate CC), 3 (severe CC), and 4 (catastrophic CC). This score takes into account only additional diagnoses. The weight of a specific additional diagnosis is determined using a predefined range and depends on the principal diagnosis, other additional diagnoses, and the adjacent DRG. For efficient grading of the AR-DRGs we used the "Combined Cost Weights" of public hospitals from 1998/1999 from the National Hospital

Cost Data Collection.[8] The case mix index (CMI) is defined as the sum of cost weights divided by the number of cases, i.e., the mean case weight.

Statistics

A variety of parameters from the published literature were used to compare the two samples. A comprehensive discussion of quality criteria for a DRG system is presented by Roeder et al.[9] Quantitative parameters are reported with absolute and relative frequencies, and distribution characteristics are reported using mean, median, and range. To assess the interrater variability for categorical data, the authors used an extended version of Cohen's kappa[10] for multiple categories (weighted kappa). Results were calculated using Microsoft Access 2000, Microsoft Excel 2000, and SPSS for Windows Release 10. The confidence limits of weighted kappa were determined using PROC FREQ of SAS System for Windows 8.02. Repeated diagnosis codes within a case were deleted before further analysis. Redundant procedure codes were accepted because procedures could realistically occur several times within the same case. English terms for ICD-10-codes were taken from a list provided by the National Center for Health Statistics <ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/publications/ICD10/>.

Results

Sample Characteristics

Each data set had 254 cases. The paper-based patient record could not be retrieved from the archive in nine cases. Due to different times of data acquisition, one case was only present in the EPR set. Another case was mentioned twice in the PPR set. Removing these irregularities left a common set of 244 cases. Because one patient had two hospital stays in September 2001, the 244 case set comprises 243 patients. The common set of 244 cases used in all calculations included 142 (58.2%) women and 102 (41.8%) men. The mean age in years at the date of admission was 61.6 and the median was 64.0 (range, 15 to 98 years). Mean length of stay (calculated as 1 day if the patient leaves the day of admission, otherwise counting every day except the day of discharge) was 11.2 days with a median of 7 days (range, 1 to 82 days).

Reliability

Two circumstances affect calculation of reliability. First, each case typically has more than one diagnosis or procedure code assigned to it. In addition, the number of codes a case contains may vary between the data sets. For these reasons, the authors used the weighted kappa statistic to provide a good estimate of intercoder reliability for the principal diagnosis. Two steps are necessary to reach an agreement in the definition of the principal diagnosis: the identification of the disease and the acceptance of that diagnosis as the principal diagnosis. This two-step process causes a trend toward underestimating reliability in comparison with studies that take into account only the first step. Table 1 (available as an online data supplement at www.jamia.org) shows the weighted kappa with respect to different levels of the ICD-10-SGB-V 2.0 structure. The assessment is taken from Landis and Koch.[11]

Quantity of Documentation

The PPR set includes 909 diagnoses with a mean of 3.7 diagnoses per case (median, 3); 384 different codes were used. About 20% (n = 55) of the cases had only one diagnosis; the maximum was a case with 16 diagnoses, which had only 11 diagnoses in the EPR set. More than half of the diseases (represented by a code from the ICD-10-SGB-V 2.0) were seen only once during the study. The PPR set included 765 procedures with a mean of 3.1 procedures per case (median, 2); 7.8 % of the cases (n = 19) had no procedure; and the maximum was a case with 37 procedures involving a patient in prolonged intensive care. Physicians performed 236 different procedures. Over half of the procedures (represented by a code from the OPS-301 2.0) were performed only once a month.

The EPR set includes 959 diagnoses with a mean of 3.9 diagnoses per case (median, 3); 436 different codes were used with a median of 1 case per disease. Like the PPR set, roughly 20% (n = 51) of the EPR cases had only one diagnosis; the maximum was 19 diagnoses, which had 10 diagnoses in the PPR set. The EPR set included 940 procedures with a mean of 3.9 procedures per case (median, 2); 12.3% of the cases had no procedure (n = 30); and the maximum was the previously noted intensive care case with 48 procedures. Care providers completed 272 different procedures. As in the PPR set, more than half of the procedures were performed only once a month.

Precision of Documentation

Avoidance of imprecise classes, such as "other" and "unspecified," is a recommended coding procedure not always followed in practice. One author (JS) reviewed each "imprecise" code to determine whether it represented an imprecise class (Table 2; available as an online data supplement at www.jamia.org).

Appropriateness of Documentation

The relative frequency of surgical procedures (indicated by a leading 5 in the code of the OPS-301) was nearly identical with 38.0% (n = 291) in the PPR set and 38.4% (n = 361) in the EPR set. The most frequent operative procedures (with more than 10 occurrences) are shown in Table 3 (available as an online data supplement at www.jamia.org). The three most frequent procedures are the same in both sets, but they are ranked in a different order. The most frequently used section of ICD for principal diagnoses in both data sets was digestive system diseases, followed by injuries, poisoning, and occupational diseases; and cardiovascular diseases. A knowledgeable surgeon (DK) classified 217 PPR-set principal diagnoses and 203 EPR-set diagnoses as surgical. In turn, 27 PPR-set diagnoses and 41 EPR-set diagnoses were not surgical in nature. In both sets, K40.90 "Unilateral or unspecified inguinal hernia, without obstruction or gangrene" was the only principal diagnosis with more than 10 occurrences: 16 (6.6%) in the PPR set and 15 (6.1%) in the EPR set. The authors analyzed diabetes mellitus and hypertension as "tracer diagnoses" independently of their use as principal or additional diagnosis to gain an understanding of the diagnosis of common diseases. "Essential (primary) hypertension" was used in 22.5% (n = 55) of the cases in the PPR set and 18.4% (n = 45) in the EPR set. Diabetes mellitus (coded as E10 to E14) appeared in 13.1% (n = 32) of the PPR-set cases and in 9.8% (n = 24) of EPR-set cases.

The mean PCCL was 1.30 in the PPR set and 1.39 in the EPR set; nearly one third of the cases were at level 3 or 4, indicating severe or catastrophic comorbitities/complications (Fig. 2).

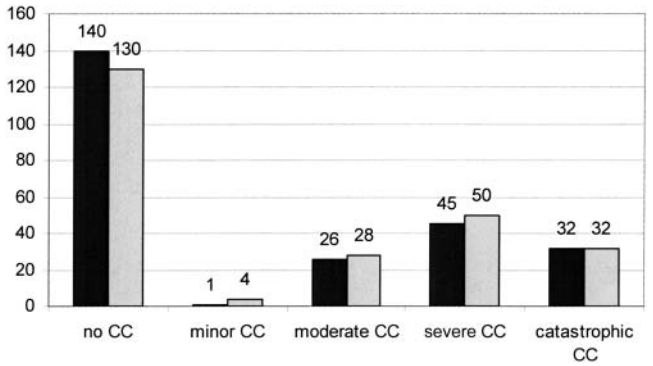

Figure 2. PCCL-distribution of the PPR set (*black, left column*) and the EPR set (*gray, right column*). The numbers indicate the cases at that level.

Impact on DRG Coding

The 244 cases led to 96 different DRGs in the PPR set and 102 different DRGs in the EPR set. Nearly half of the DRGs occurred only once: 48 in the PPR-set and 59 in the EPR set. Five cases led to erroneous DRGs in both sets. A comparison of the two sets is shown in Table 4 (available as an online data supplement at www.jamia.org). The PPR-set CMI was 2.09 while the EPR-set had a CMI of 2.06. The median value for both sets was 1.26. An increase of only 1.5% could be reached by documentation with the paper-based patient record.

Discussion

Current Study

The current study indicated that diagnoses coded from the paper-based patient record may have minor qualitative advantages. In the PPR set the use of imprecise diagnostic classes was reduced, and the proportion of surgical principal diagnoses and the frequency of tracer diagnoses were increased. The higher total number of diagnoses in the EPR set may account for the improved PCCL (1.39 vs. 1.30) in that set. The EPR documentation showed potential advantages in both quality and quantity of procedure coding: a lower number of imprecise codes and a higher number of codes pertaining to operations (corresponding to the higher number of operative DRGs). But the broad-based definition of DRGs makes these differences immaterial. From an economic efficiency point of view, it appears that no additional reimbursement would be achieved by coding diagnoses and procedures from the full paper record.

Implications for Studies Comparing Paper and Electronic Chart Abstracts

The Patient as the Gold Standard for Comparisons

A gold standard representing the truth regarding the patient should be the reference point for the comparison of medical record types. This would allow the comparison of sensitivity and specificity on levels II and III in Figure 1. It is straightforward to achieve a high level of accuracy regarding the concrete steps taken to care for a patient diagnostically and therapeutically, e.g., procedures. Diagnoses, however, represent the results of complex processes in medical decision making. The authors note a pioneering study that used the patient as the gold standard carried out by Pringle et al.[12] This group validated the entities stored in the EPR through review of video-recorded patient encounters (consultations). They found that the EPR was incomplete. However, they also showed that the number of diagnoses in the EPR was double that recorded in the PPR. This finding supports the authors' perception of an EPR as including more than a simple subset of the written chart contents. To test their approach on data quality, Logan et al.[13] used video-taped patient encounters as a gold standard as well.

The Paper-based Patient Record as the Gold Standard for Comparisons

Most studies comparing EPR with PPR coding consider the paper-based patient record as the gold standard. Comparison of electronically available data with the paper record often is called *validation*.[14] Hassey et al.[15] used different references for different parts of the record in their study on validity of electronic patient records in a general practice. For example, they checked the completeness of prescribed information through comparisons with pharmacy data. Barrie and Marsh[16] reported a completeness of 62% and an accuracy of 96% in an orthopedic database comparing stored key words with "ideal key words" gained from clinical notes. A study on the availability and accuracy of data for medical practice assessment in pediatrics was carried out by Prins et al.[17] Their information system provides nine of 14 criteria regarded as clinically relevant for medical practice assessment. Accuracy was defined as the degree to which information from the paper record was present in the EPR (which seems to be a combination of completeness and accuracy from Barrie and Marsh[16]). The accuracy was between 0.65 for diagnosis codes and 1.0 for test results (and some other criteria).

Hogan and Wagner[18] express a similar view to Barrie and Marsh[16] in their meta-analysis of data quality in EPRs. They defined completeness in terms of the number of observations recorded and correctness as the proportion of correctly recorded observations. This corresponds to recall (completeness) and precision (correctness) in the assessment of retrieval methods. Hogan and Wagner[18] called for further studies to improve the knowledge of data accuracy in EPRs. Logan et al.[13] extended this definition (see Appendix; available as an online data supplement at www.jamia.org) by distinguishing among items that were present and correct, present and incorrect, and absent in a trial. In comparison with the definition of recall, this approach obtains higher measures of completeness.

Issues of Concordance and Reliability

The authors believe that the paper-based patient record should not be taken as the gold standard over the electronic record when circumstances create two different and supplemental records. The degree of concordance or reliability could then be a first level of analysis. Despite the authors' report suggesting that, at the coarse level of DRG abstraction, the two methods were roughly equal with respect to the institution studied, the authors observed more serious differences at the more detailed level. Nilsson et al.[19] analyzed the reliability of diagnosis coding in Sweden using that country's primary health care version of the ICD, which included 972 codes. Six general practitioners (GPs) coded 152 problems from 89 encounters using three different methods: book lookup, computerized book lookup, and a computer tool that provided a compositional approach to finding diagnostic terms. The best kappa reached was 0.58 on the

code level and 0.82 on the ICD organ system (chapter) level, both attained using the book. Morris et al.[20] evaluated the reliability of procedure coding in intensive care with the Current Procedure Terminology (CPT). Seven people recorded the CPT Evaluation & Management level of services codes for 100 charts. A computer tool also measured the level. Kappa, a measure of interrater reliability, was calculated pair wise and against a consensus. The mean interrater kappa was 0.38, and the mean rater-versus-consensus kappa was 0.53. The best kappa for pairwise agreement was 0.57 between a medical expert and a person for standard billing.

There are substantial differences between coding in practice and coding in the experimental study design of Nilsson et al.[19] and Morris et al.[20]

- The number of possible categories for diagnoses and procedures is quite higher, with 12,401 codes in the ICD-10-SGB-V 2.0 and 23,160 codes in the OPS-301 2.0.
- In addition, the authors noted two steps underlying coding of the principal diagnosis, identifying the disease and characterizing it as the principal diagnosis.
- There are several diagnoses and procedures per case, not single codes. Furthermore, the classifications themselves assume the use of multiple codes per clinical concept. A good example of this is the dagger–asterisk system of the ICD. A measurement of reliability should ideally be able to cope with the comparison of sets.
- Internal inconsistencies of the classifications themselves cause coding errors and coding weaknesses.[21]
- The weighted kappa coefficient was designed for ordinal values. It could be used for nominal values,[22] but then it cannot weigh the different distances between values. For example, the disagreement between K21.0, "Gastroesophageal reflux disease with esophagitis," and I21.0, "Acute transmural myocardial infarction of anterior wall," is bigger than the difference between I21.0, "Acute transmural myocardial infarction of anterior wall," and I21.2, "Acute transmural myocardial infarction of other sites." Table 1 (available as an online data supplement at www.jamia.org) includes the kappa coefficient for the levels of the ICD-10-SGB-V 2.0, which provides a good example of the different distances.
- In practice, it is difficult to control Step A in Figure 1, which could create noise in the data. Studies on medical documentation are affected by less-standardized clinical diagnostics.

Previous studies on the relative reliability of paper-based and electronic coding are substantially confounded by the methodologic problems listed above. The authors' study and previous studies indicate that a more complex study design, sufficient to demonstrate reliability, must be developed as a basis for further analysis of differences between the paper-based and electronic patient records.

Proposed Quality Criteria for Comparing Electronic and Paper Chart Abstracts

Rationale

The authors believe it is insufficient to assess only retrieval capacity when comparing different types of records. The authors propose use of "quality criteria for documentation" to compare paper-based and electronic patient records. These criteria should focus on content such as precision and appropriateness. Then, the respective figures for each record type could be validated against external "gold standard" ones.

Quantity of Items of Interest Documented

In evaluating paper and electronic records, it is important to have a reasonable expectation for the amount of relevant data that each should contain. For example, for DRG systems, the quantity of diagnoses codes is a well-established quality criterion. However, for hospitals using the ICD, little is known about the "expected" rate for documenting comorbidity. Hohnloser et al.[23] analyzed discharge summaries of intensive care unit patients and reported means between 3.2 and 3.64 for free-text diagnoses. With the introduction of an EPR, 83% of the free-text diagnoses were coded (mean, 2.75 per discharge summary). Iezzoni et al.[24] found a mean number of diagnoses per case of 5.5 (median, 5 codes) in a study based on computerized hospital discharge data from California, which allowed up to 25 diagnoses per discharge. They showed hospital-based differences with a range between 2.5 and 11.7 diagnoses per case. Kerby et al.[25] presented data from two data sets of 18 family medicine clinics: an administrative data set and a clinical data set. One can calculate a mean number of diagnoses of 4.4 (201,871 diagnoses from 45,617 patients) in the administrative and 4.6 (122,449 diagnoses from 26,511 patients) in the clinical data set per patient. In comparison with the literature, the authors' study found no diagnosis overcoding among the PPR and EPR sets. Three to four diagnoses codes per case could be expected from existing evidence. Studies quantifying the number of procedures performed in hospitals are rare: Ingenerf et al.[26] reported 3.19 (EPR) versus 3.72 (paper record) procedures per case from the Surgical Department at the University of Lübeck in Germany. Both are more than the authors found in the current study. Stausberg et al.[27] found a median of 3 surgical procedures per operation diagnosis (represented by an ICD-9 code) in 3.5 years worth of general surgery documentation in a university hospital. The current study detected a 1-case-per-diagnosis-code median in the EPR set and the PPR set. It is unclear whether the observation of many codes with a low number of cases is an artifact of classifications or a valid representation of a highly specialized medical practice.

Dexter and Macario[28] reported problems in operation room scheduling based on historical data. They analyzed raw data from the U.S. National Survey of Ambulatory Surgery, which included 228,332 visits from 1994 to 1996 and found that 36% of all cases in the United States had a procedure or a combination of procedures that occurred fewer times than the number of surgical specialists performing ambulatory surgery. Twenty percent of the cases had a procedure or a combination that occurred 1,000 times or less.

The slightly higher number of diagnoses in the EPR set corresponds to a higher PCCL mean of 1.39 in comparison with 1.30. The clinical profiles in Australia[29] show a mean PCCL of 0.62 with and 0.63 without error-DRGs. It seems appropriate for a university-associated surgical department to have a more severe casemix as it could be calculated on a national level for Australia.

Proportion of Imprecise Codes Used for Entities of Interest

In the PPR set, the authors were able to reduce the total number of imprecise diagnosis codes. The more frequent

use of imprecise codes in the EPR set might reflect time constraints for coding in a daily routine. One has to note that the coding tool DIACOS induces the use of "unspecified" codes, because these codes are offered at the top of a result list for an input string. However, the authors also detected problems caused by the structure of the classifications. The characterization of imprecise classes by elements of the codes is inconsistent, especially for general surgery. In two frequently occurring cases in general surgery, appendicitis and inguinal hernia, codes have to be used for uncomplicated cases that indicate "unspecified." By taking into account this false-positive result, the authors reduced the frequency of "unspecified" codes for diagnoses to 7.3% in the PPR set and 11.1% in the EPR set. The OPS-301 showed its more sophisticated structure in comparison with the ICD. All codes indicating an imprecise class could be confirmed.

Calibration of Entities Coded with Respect to Known Population Parameters (e.g., Prevalence of Tracer Diagnoses in Charts and in the General Population)

Pringle et al.[12] assessed the completeness of electronic patient records in four British GP practices. They used prevalence of diabetes mellitus and glaucoma as "tracer" diagnoses and reported identifying additional cases (not present in the EPR) from paper records. They found 27 "new" cases of diabetes mellitus in addition to the 785 cases previously documented in the EPR (+3.4%) and 17 cases of glaucoma in addition to the 205 (+8.3%) in the EPR. The reported prevalence of diabetes mellitus in the EPR was 2.1% of the practice population (versus 2.7% in national data, based on consultation rate per 100 person-years at risk). The reported prevalence of hypertension was 5.8% (versus 10.3% from national data, based on consultation rates per 100 person-years at risk).

In the authors' current study, the authors' paper-based review identified 17% more cases of diabetes mellitus, and 33% more cases of hypertension than were documented previously in the EPR, indicating, like the findings of Pringle et al.,[12] incompleteness in the EPR set. There is adequate agreement between our findings and the available German data about the national prevalence of diabetes mellitus. Thefeld[30] reported a prevalence rate of 4.7% in a representative sample of 18- to 79-year-olds, In the 60- to 69-year age range, he found a prevalence rate of 12.9% in comparison with 12.5% in the PPR set, including counting seven patients with diabetes mellitus solely as an additional diagnosis in relation to 56 patients in this age category. In contrast to diabetes mellitus, the relative frequency of hypertension is below national data. Using Thefeld's representative sample, Thamm[31] reported a prevalence rate of about 30% for men and 26% for women. The authors found hypertension in 19.6% (20 of 102 cases) of men and 24.6% (35 of 142 cases) of women using the PPR set. Different economical impacts of diabetes mellitus and hypertension in the AR-DRGs may have caused this disparity. For the same reason, hypertension without therapy may have been omitted in coding records as well.

Economic and Quality-of-care Effects of Differences in Coding Efficiency

Recent findings in Germany suggested a potential difference of 14% (or more) for economic return based on optimization (or lack thereof) in the documentation of care performed by clinicians routinely.[26,32,33] The authors' current study results are remarkably different, suggesting a benefit of optimal coding of only 1.5% (Table 5; available as an online data supplement at www.jamia.org). This difference might be explained by the information and workflow management of the Alfried Krupp Hospital, which is the only nonuniversity hospital among the studies. The higher absolute CMI of the other studies could be explained by the different casemixes of university clinics and a nonuniversity hospital.

With the hospitals' self-developed and self-maintained electronic patient record system, it was possible to react quickly to the announcement of the upcoming German DRG system. Organizational structures for data monitoring had been implemented before our study took place. Furthermore, the focus of our study was not an investigation about potentials for economic optimization. Thus, the coding from the PPR was based on the same concept as that in routine use: to capture valid and formal consistent information. Notably, there is no evidence in the current study for the DRG creep predicted by Simborg[34] and documented to occur in the United States after introduction of DRGs.[35]

The use of routinely collected data for quality measurement and quality improvement in health care is controversially discussed.[36,37] On one hand, the administrative nature of data that had been coded with ICD-9-CM typically is seen as a handicap for quality measurement and benchmarking.[38] On the other hand, the impact of automatic alerts,[39] automatic sentinel event detection,[40] and feedback of rate-based quality indicators could be demonstrated.[41]

Study Methodology in Comparing Two Patient Record Formats

All studies comparing paper and electronic records have to carry out a transformation of at least one of the two records, because, by definition, they are in different formats. Because such studies require a transformation of one or both of PPR and EPR, the manner in which the transformation is done can critically influence study results. For example, if the decision is made to compare a transformation of the PPR with the EPR, process step D in Figure 1 has to be managed. Whereas most previously published studies do not mention anything about the methodology used,[3,14,15,17,23,42] other studies, including the current authors' study, used a single individual to make the transformation.[12,16] A single individual, who has not been calibrated as to performance compared with coders in the general community, or to "best of practice" expert coders, cannot stand as a representative for all of them (especially if that individual is also an author of the study, as occurred in the current study). Studies using single coders to transform one or the other record format then are critically limited by the unknown specificity and sensitivity of that individual (Level III in Fig. 1) in comparison with the best transformation of the data. Future studies should improve this process, for example, by using expert consensus of calibrated individuals, rather than a single arbitrary individual, to transform the records—a method used by Logan et al.[13]

The authors made the case above for using the patient (i.e., all that is known and documented about the patient, independent of source) as the gold standard for comparisons, as opposed to the PPR or the EPR. However, to determine the best coding for a given chart, it is similarly beneficial to have multiple, calibrated individuals (with reported interrater

reliability measures) determine the gold standard, not just a single individual (or a blind, unsupervised merger of the two medical record formats).

Most studies are limited also by the use of classifications as a representation system, as Brennan and Stead pointed out.[43] The "honest, true-to-life depiction of the patient"[43] is available in neither the paper-based nor the electronic patient record. Furthermore, classifications, coding schemes, and coding guidelines are tailored to specific needs, for example, reimbursement.[44] In agreement with previous studies by Nilsson et al.[19] and Morris et al.,[20] the authors have shown that coding the same way does not lead to the same codes in a significant number of cases. Morris et al.[20] state, "perspective and motivation changes coding outcomes, especially when left with loose guidelines to govern behavior."

Conclusions

Objective measures are required to evaluate the quality of documentation in an EPR.[45] No standard set of assessment criteria has been approved or adopted, and a gold standard for comparisons currently is missing. It is extremely difficult pragmatically to take the patient per se as the gold standard for interpretative coding and abstraction entities such as diagnoses. Thus, most previously published studies compared different kinds of medical documentation (e.g., electronic and paper records) with each other. Previous studies used criteria tailored to reflect the performance of documentation ("How many diagnoses had been documented?"), or outcome parameters relevant outside the clinical process, for example, economical impact parameters.

The authors are aware of methodologic shortcomings in their current study. Nevertheless, current study results support the finding of Mikkelsen and Aasly[3] that parallel use of electronic and paper-based patient records can lead to inconsistencies in the medical documentation. Medical professionals should be aware of this situation and combine the information from both records whenever possible. The authors concede that it may be too expensive to strive for a total concordance between paper and electronic data sets, which often are used for dramatically different purposes in medical practice.[46] It is ultimately the goal to join all data into one ubiquitous electronic record. But this is only possible if care providers accept that valid data must be present to represent the truth in patient records for all intended uses of the record.

References ■

1. Dick RS, Steen B, Detmer DE (eds). The Computer-Based Patient Record. An Essential Technology for Health Care. Revised edition. Washington, DC: National Academy Press, 1997.
2. NHS Executive. Information for health. An information strategy for the modern NHS 1998–2005. A national strategy for local implementation. Great Britain: Wetherby; September 1998.
3. Mikkelsen G, Aasly J. Concordance of information in parallel electronic and paper based patient record. Int J Med Inform. 2001;63:123–31.
4. Nygren E, Johnson M, Henriksson P. Reading the medical record II. Design of a human-computer interface for basic reading of computerized medical records. Comput Methods Progr Biomed. 1992;39:13–25.
5. Tange HJ. The paper-based patient record: is it really so bad? In: Barahona P, Veloso M, Bryant J (eds). Proceedings of the Twelfth International Congress of the European Federation for Medical Informatics. Lisbon, 1994, pp 459-63.
6. Deutsche Krankenhausgesellschaft, Spitzenverbände der Krankenkassen, Verband der privaten Krankenversicherung. Deutsche Kodierrichtlinien. Allgemeine Kodierrichtlinien für die Verschlüsselung von Krankheiten und Prozeduren. Version 1.0, April 2001.
7. Australian Refined Diagnosis Related Groups, Version 4.1, Definitions Manual, Vol. 1–3. Canberra, 1998.
8. Commonwealth Department of Health and Aged Care. Report on the national hospital cost data collection. 1998–99 (Round 3). Draft for Comment. Canberra, 2000.
9. Roeder N, Irps S, Juhra C, et al. Erlöse sichern durch Kodierqualität. Messung und Interpretation von Kodierqualität. das Krankenhaus. 2002;94:117–27.
10. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure. 1960;20:37–46.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
12. Pringle M, Ward P, Chilvers C. Assessment of the completeness and accuracy of computer medical records in four practices committed to recording data on computer. Br J Gen Pract. 1995; 45:537–41.
13. Logan JR, Gormann PN, Middleton B. Measuring the quality of medical records: a method for comparing completeness and correctness of clinical encounter data. In: Bakken S (ed). A Medical Odyssey: Visions of the Future and Lessons from the Past. Proc AMIA 2001 Annu Symp.2001:408–12.
14. Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resources in the United Kingdom. BMJ. 1991;302:766–8.
15. Hassey A, Gerrett D, Wilson A. A survey of validity and utility of electronic patient records in a general practise. BMJ. 2001; 322:1401–5.
16. Barrie JL, Marsh DR. Quality of data in the Manchester orthopaedic database. BMJ. 1992;304:159–62.
17. Prins H, Kruisinga FH, Büller HA, Zwetsloot-Schonk JHM. Availability and accuracy of electronic patient data for medical practice assessment. In: Hasmann A, Blobel B, Dudeck J, Engelbrecht R, Gell G, Prokosch HU (eds). Medical Infobahn for Europe. Proceedings of MIE2000 and GMDS2000. Amsterdam: IOS, 2000, pp 484–8.
18. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc. 1997;4:342–55.
19. Nilsson G, Petersson H, Åhlfeld H, Strender L-E. Evaluation of three Swedish ICD-10 primary care versions: reliability and ease of use in diagnostic coding. Methods Inf Med. 2000;39:325–31.
20. Morris WC, Heinze DT, Warner HR, et al. Assessing the accuracy of an automated coding system in emergency medicine. Proc AMIA Annu Symp. 2000;595–9.
21. Surján G. Questions on validity of International Classification of Diseases-coded diagnoses. Int J Med Inform. 1999;54:77–95.
22. Krummenauer F. Extensions of Cohen's kappa coefficient for multi rater trials: an overview. Informatik, Biometrie und Epidemiologie in Medizin und Biologie. 1999;30:3–20 [in German].
23. Hohnloser JH, Puerner F, Soltanian H. Improving coded data entry by an electronic patient record system. Methods Inf Med. 1996;35:108–11.
24. Iezzoni LI, Foley SM, Hughes J, Fisher ES, Heeren T. Comorbitities, complications, and coding bias. Does the number of diagnosis codes matter in predicting in-hospital mortality? JAMA. 1992;267:2197–203.
25. Kerby J, Keshavjee K, Holbrook AM. Comparison of diagnostic codes in a clinical-research database and an administrative database. Proc AMIA Annu Symp. 2000;1045.
26. Ingenerf J, Stellmacher F, Stausberg J, et al. Analyse der rechnergestützten Kodierqualität durch Vergleich mit der

konventionellen Krankenakte: Bewertung der klinischen und ökonomischen Qualität (DRG-System). Informatik, Biometrie und Epidemiologie in Medizin und Biologie. 2002;33:174.
27. Stausberg J, Lang H, Obertacke U, Rauhut F. Classifications in routine use: lessons from ICD-9 and ICPM in surgical practice. J Am Med Inform Assoc. 2001;8:92–100.
28. Dexter F, Macario A. What is the relative frequency of uncommon ambulatory surgery procedures performed in the United States with an anesthesia provider? Anesth Analg. 2000; 90:1343–7.
29. Commonwealth Department of Health and Aged Care. DRG clinical profiles for all AR-DRGs (v4.1), public and private acute hospitals combined, 1998–99. <http://www.health.gov.au/casemix/report/hospmo18.htm>. Accessed July 30, 2001.
30. Thefeld W. Prevalence of diabetes mellitus among adults in Germany. Gesundheitswesen. 1999;61:S85–S89 [in German].
31. Thamm M. Blood pressure in Germany—update review of state and trends. Gesundheitswesen. 1999;61:S90–S93 [in German].
32. Langrehr JM, Lohmann R, Birkner K, Neuhaus P. Evaluierung des 'Case Mix Index' und Fehleranalyse der Diagnosen- und Prozedurendokumentation einer chirurgischen Klinik. Ein einfaches Verfahrensmodell. Abstract CHIR-1214. 119. Kongreß der Deutschen Gesellschaft für Chirurgie, Berlin, 2002.
33. Mieth M, Wolkener F, Schmidt J, Glück E, Klar E, Kraus T. Prospective comparison of the effects of maximum and limited clinical documentation on the calculated hospital revenue in a surgical unit, based on the AR-DRG system. Chirurg. 2002; 73:492–9 [in German].
34. Simborg DW. DRG creep. A new hospital-acquired disease. N Engl J Med. 1981;304:1602–4.
35. Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. N Engl J Med. 1988;318:352–5.
36. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Policy and the future of adverse event detection using information technology. J Am Med Inform Assoc. 2003; 10:226–8.
37. Schwartz RM, Gagnon DE, Muri JH, Zhao QR, Kellogg R. Administrative data for quality improvement. Pediatrics. 1999; 103:291–301.
38. Iezzoni LI. Assessing quality using administrative data. Ann Intern Med. 1997;127:666–74.
39. Kuperman GJ, Gardner RM, Pryor TA. HELP: A Dynamic Hospital Information System. NewYork: Springer, 1991.
40. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse event using information technology. J Am Med Inform Assoc. 2003;10:115–28.
41. Stausberg J, Kolke O, Albrecht K. Using a Computer-based patient record for quality management in surgery. In: Cesnik B, McCray AT, Scherrer J-R (eds). MedInfo 98. Amsterdam: IOS, 1998. pp 80–4.
42. Aronsky D, Haug PJ. Assessing the quality of clinical data in a computer-based record for calculating the pneumonia severity index. J Am Med Inform Assoc. 2000;7:55–65.
43. Brennan PF, Stead WW. Assessing data quality: from concordance, though correctness and completeness, to valid manipulatable representations. J Am Med Inform Assoc. 2000;7:106–7.
44. Stausberg J. Design of classifications for diagnoses and procedures in a DRG system. Gesundheitsökonomie & Qualitätsmanagement. 2002;7:297–303 [in German].
45. Baptista J, Reis Abreu J, Correia C, Cordeiro A. Quality of computerized health data. In: Barahona P, Veloso M, Bryant J (eds). Proceedings of the Twelfth International Congress of the European Federation for Medical Informatics. Lisbon, 1994, p. 699.
46. Starmer CF. Hitting a moving target: toward a compliance-driven patient record. J Am Med Inform Assoc. 2002;9:659–60.