journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Reliability of diagnoses coding with ICD-10

Jürgen Stausberg[a,*], *Nils Lehmann*[a], *Dirk Kaczmarek*[b], *Markus Stein*[c]

[a] *Institute for Medical Informatics, Biometry and Epidemiology, Medical Faculty,*
University of Duisburg-Essen, Hufelandstr. 55, D-45122 Essen, Germany
[b] *Sankt Marien-Hospital Buer gGmbH, Gelsenkirchen, Germany*
[c] *Clinical Centre Ludwigshafen, Germany*

ARTICLE INFO

Article history:
Received 30 January 2006
Received in revised form
1 September 2006
Accepted 20 November 2006

Keywords:
Coding
Data quality
Diagnoses
Diagnosis related groups
ICD-10
Reliability

ABSTRACT

Objective: Reliability of diagnoses coding is essential for the use of routine data in a national health care system. The present investigation compares reliability of diagnoses coding with ICD-10 between three groups of coding subjects.

Method: One hundred and eighteen students coded 15 diagnoses lists, 27 medical managers from hospitals 34 discharge letters, and 13 coding specialists 12 discharge letters. Agreement in principal diagnosis was assessed using Cohen's Kappa and the fraction of coincidences over the number of pairs, agreement for the full set of diagnoses with a previously developed measure p_{om}.

Results: Kappa values were fair (managers) or moderate (coders) for terminal codes with 0.27 and 0.42 (agreement 29.2% versus 46.8%), substantial for the chapter level with 0.71 and 0.72 (agreement 78.3% versus 80.8%). p_{om} was lower for the full set of diagnoses than for principal diagnoses, for example in case of managers with 0.21 versus 0.29 for terminal codes. Best results were achieved by students coding diagnoses lists. In summary, the results are remarkably lower than in earlier publications.

Conclusion: The refinement of the ICD-10 accompanied by innumerous coding rules has established a complex environment that leads to significant uncertainties even for experts. Use of coded data for quality management, health care financing, and health care policy requires a remarkable simplification of ICD-10 to receive a valid image of health care reality.

1. Introduction

The use of classified and coded medical entities for reimbursement, quality management, and health care policy has increased enormously in the last 30 years. The usefulness of these data relies basically on an identical coding of the same entity independent of the coding person and/or the time of coding. Thirty years ago the Institute of Medicine (IOM) analyzed the reliability of diagnoses coding from hospital discharge abstracts with the 8th Revision of the International Classification of Diseases (ICD) [1]. An independent re-coding of the principal diagnoses confirmed 65.2% of the original codes. Since then, various studies have raised issues such as whether hospitals use systematically wrong codes to increase reimbursement [2] or whether administrative data include the necessary elements for quality management [3]. Many studies have been published concerning the validity of coded data [4,5]. But it is still not clear whether diagnoses coding with ICD is more than a matter of chance.

Some established problems raise concerns about the present reliability of diagnoses coding with ICD:

[*] *Corresponding author*. Tel.: +49 201 723 4512; fax: +49 201 723 5933.
E-mail address: stausberg@ekmed.de (J. Stausberg).

1386-5056/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2006.11.005

- The ICD includes ambiguities and inconsistencies [6].
- Coding of abstracts and medical reports is influenced by different conclusions about existing diagnoses [7].
- Refinement of ICD for reimbursement and a high number of rules constitute a complex coding system, which is quite difficult to understand, even for coding experts.

Coding of medical entities with classifications is a hot topic in Germany. The codes are used for reimbursement and system design of the German Diagnosis Related Groups (G-DRGs), introduced on a mandatory basis to hospitals in 2004. Obligatory public quality reports from hospitals include performance statistics comprising codes. These reports were published first in 2005 for 2004. A system for risk compensation is in progress. Health insurance companies will establish morbidity scores derived from coded data.

We conducted an investigation on the reliability of diagnoses coding from discharge letters with the German modification of the ICD-10 for health care financing (ICD-10-GM) [8]. The ICD-10-GM is a successor of a pooling of an earlier German adaptation of WHO's 10th revision with the ICD-10 Australian Modifications (ICD-10-AM) Version 1. Due to the adoption of the Australian Refined DRGs (AR-DRGs) in 2003 compatibility with the ICD-10-AM was required. ICD-10-GM is revised each year according to requirements from the G-DRGs. For coding of procedures a national classification – abbreviated as OPS – is used based on WHO's International Classification of Procedures (ICPM), also adapted to the Australian DRGs. The ICD-10-GM 2004 included 12,983 terminal codes.

We aimed at calculating the reliability of diagnoses coding. Reliability measures the agreement of different persons coding the same case (inter-rater reliability) or the agreement of one person at different times coding the same case (intra-rater reliability). Reliability is different from validity. Validity measures the agreement with a gold standard. On the one hand it is possible to have high reliability but weak validity, if all raters agree in their wrong decisions. On the other hand, low reliability can be explained two-fold. It can be the consequence of insufficient education and training, and of inadequate standardization of the coders and the coding scenario. But it can also indicate weaknesses in the classification used for coding mentioned above. In the latter case, low reliability indicates poor quality of a coding system and should lead to a major revision!

The investigation was split into three studies: medical students coding diagnoses lists from discharge letters, physicians working in medical management in hospitals coding from discharge letters, and specialists in medical documentation also coding from discharge letters. Results from the first study with medical students were published previously [9]. Objectives of our study were to learn about the ICD-10, to find arguments for the discussion who should code and to get information on the quality of data coded in routine care.

2. Materials and methods

Discharge letters were used as basis for coding. The letters originate from a department of internal medicine of a medium sized municipal hospital and had been written by one physician in the early 1990s. They cover a full range of medical problems with special emphasis on nephrology. Personal data had been deleted including any datestamps concerning seldom events, rare diseases, or pathognomonic information. The length of the letters ranged from 1 to 4 pages (cf. Fig. 1). Participants were asked to code the diagnoses independent of each other, at the time and the location of their own choice. No rules were established concerning assisting tools like software or handbooks. The students were pointed to a WWW-version of the ICD-10 offered at http://www.dimdi.de/, the other participants were reminded to follow the German Coding Directives [10]. Students' results were collected on paper and subsequently entered into a database; results from the other participants were stored in simple Excel-files with columns for type of diagnosis and code that were distributed by the study center. The codes were transferred to a database from Excel using the Microsoft® Windows clipboard.

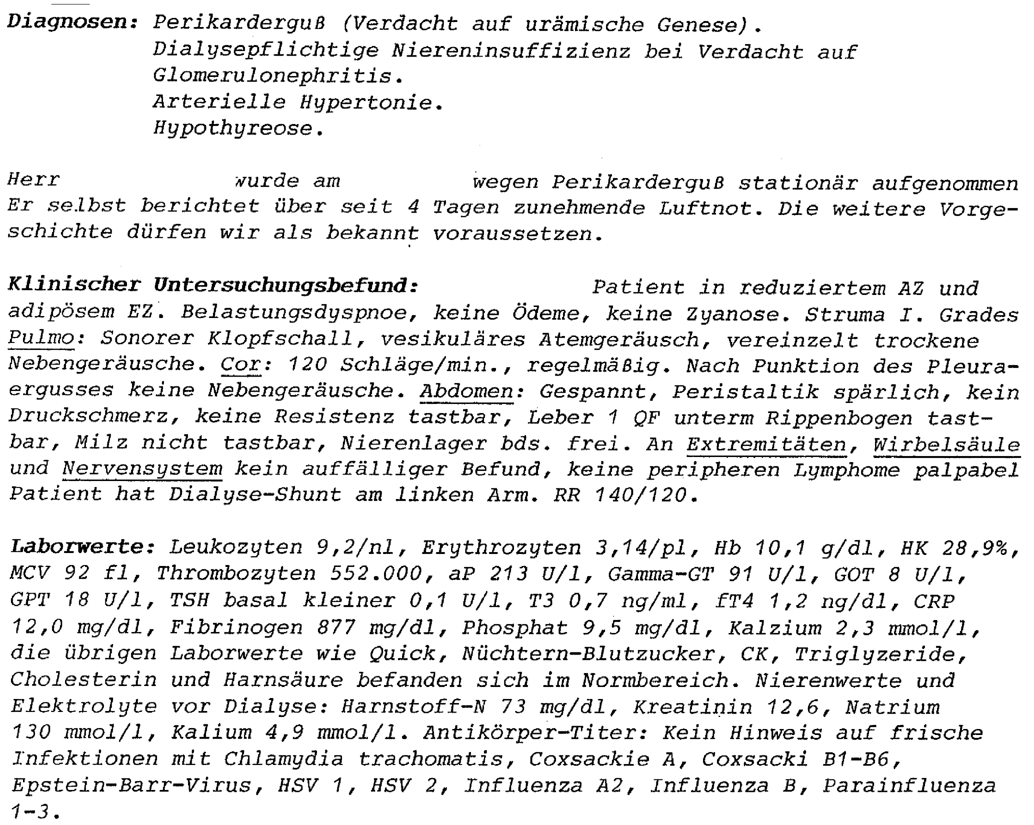
Table 1 – Interpretation of Kappa-values

Kappa	Grade of reliability
Landis and Koch [11]	
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect
Fleiss et al. [12]	
<0.40	Poor
0.40–0.75	Fair to good
>0.75	Excellent

Two special conditions in measuring observer agreement in our study should be mentioned. On the one hand, the number of possible categories (12,983 possible codes on the terminal level) in relation to the number of categories actually used is extremely high. On the other hand, we want to calculate agreement in one (the principal) diagnosis as well as in sets of diagnoses, not necessarily of equal size.

The Kappa measure serves as indicator of agreement, when n cases are given one diagnosis each by two raters. It can be written as $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is the proportion of observed accordances and p_e is the expected rate of accordances, calculated from table marginals. As the number of answering categories increases, the table inflates and typically becomes sparse, such that $p_e \ll p_o$ (<1). Then $\kappa \approx p_o$. We used Cohen's Kappa for the measurement of reliability of the principal diagnosis with the graduation proposed by Landis and Koch [11] (cf. Table 1).

Given multiple codes for n cases classified by two raters, we calculate for each case $P = n_o/(n_1 + n_o + n_2)$. Here n_o is the number of accordances, n_1 the number of diagnoses of rater 1 not agreeing with any diagnosis of rater 2, and n_2 is the number of diagnoses of rater 2 not matching a diagnosis of rater 1 (cf. Fig. 2). As measure of agreement in this case we compute p_{om} as the mean of P over the cases. If there is only one diagnosis per case, P is either 0 or 1, and the mean value of P is



Diagnosen: Perikarderguß (Verdacht auf urämische Genese).
Dialysepflichtige Niereninsuffizienz bei Verdacht auf Glomerulonephritis.
Arterielle Hypertonie.
Hypothyreose.

Herr wurde am wegen Perikarderguß stationär aufgenommen Er selbst berichtet über seit 4 Tagen zunehmende Luftnot. Die weitere Vorgeschichte dürfen wir als bekannt voraussetzen.

Klinischer Untersuchungsbefund: Patient in reduziertem AZ und adipösem EZ. Belastungsdyspnoe, keine Ödeme, keine Zyanose. Struma I. Grades Pulmo: Sonorer Klopfschall, vesikuläres Atemgeräusch, vereinzelt trockene Nebengeräusche. Cor: 120 Schläge/min., regelmäßig. Nach Punktion des Pleuraergusses keine Nebengeräusche. Abdomen: Gespannt, Peristaltik spärlich, kein Druckschmerz, keine Resistenz tastbar, Leber 1 QF unterm Rippenbogen tastbar, Milz nicht tastbar, Nierenlager bds. frei. An Extremitäten, Wirbelsäule und Nervensystem kein auffälliger Befund, keine peripheren Lymphome palpabel Patient hat Dialyse-Shunt am linken Arm. RR 140/120.

Laborwerte: Leukozyten 9,2/nl, Erythrozyten 3,14/pl, Hb 10,1 g/dl, HK 28,9%, MCV 92 fl, Thrombozyten 552.000, aP 213 U/l, Gamma-GT 91 U/l, GOT 8 U/l, GPT 18 U/l, TSH basal kleiner 0,1 U/l, T3 0,7 ng/ml, fT4 1,2 ng/dl, CRP 12,0 mg/dl, Fibrinogen 877 mg/dl, Phosphat 9,5 mg/dl, Kalzium 2,3 mmol/l, die übrigen Laborwerte wie Quick, Nüchtern-Blutzucker, CK, Triglyzeride, Cholesterin und Harnsäure befanden sich im Normbereich. Nierenwerte und Elektrolyte vor Dialyse: Harnstoff-N 73 mg/dl, Kreatinin 12,6, Natrium 130 mmol/l, Kalium 4,9 mmol/l. Antikörper-Titer: Kein Hinweis auf frische Infektionen mit Chlamydia trachomatis, Coxsackie A, Coxsacki B1-B6, Epstein-Barr-Virus, HSV 1, HSV 2, Influenza A2, Influenza B, Parainfluenza 1-3.

Fig. 1 – First page of a discharge letter used for coding (in German). It is opened by a list of diagnoses in free text, followed by the reason for admission, the physical examination and lab values.

just the observed agreement, $p_{om} = p_o$. Thus, p_{om} is a possible generalization of p_o to the case of multiple codes. With more than two observers, we calculate P for each case for each distinct pairing of observers, and average over pairs. The value of p_{om}, a number between 0 (no agreement) and 1 (perfect agreement), is determined by averaging again, now over cases. A comparison of the different groups of discharge letters covered by each study is based on the estimates calculated for each study independently.

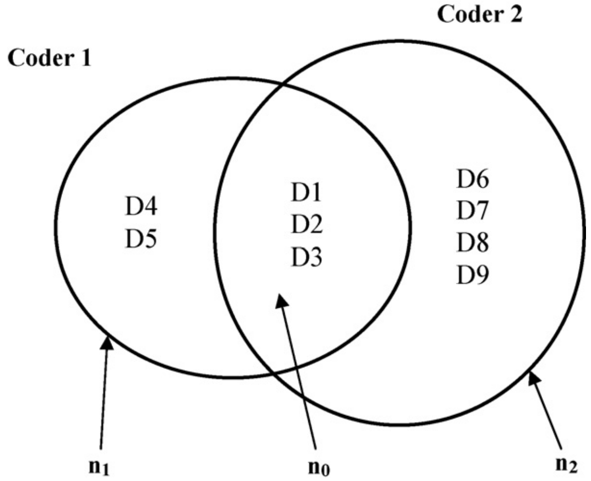


Fig. 2 – Calculation of p_{om}. First, we calculate P as $P = n_o/(n_1 + n_o + n_2)$ for each pair of coders. Secondly, we calculate the mean P for every discharge letter. Averaging the latter over discharge letters yields p_{om}.

The data were stored in a Microsoft® Office Access 2003-database, analyses were done using Access 2003, Microsoft® Office Excel 2003 and the Statistical Analysis System® (SAS) version 8.2. In the following we will refer to a diagnoses set of one participant for one discharge letter as "form". "Diagnosis" and "code" is used synonymously, being aware that a single diagnosis could lead to more than one code on the one hand (e.g. meningococcal meningitis: G01* and A.39.0+), and a code of ICD-10 could cover disease-independent information like procedures on the other hand (e.g. single delivery by caesarean section: O82).

Some plausibility checks concerning the diagnoses codes were performed before further analysis. First, duplicates of codes were deleted, thus any code could appear only once in a form. For example, if a code appears as principal diagnosis and as secondary diagnosis on the same form, the secondary diagnosis code was deleted. Clear typing errors were corrected, e.g. a blank between characters of the code. All codes were then checked against an official list of valid codes. Invalid codes were handled as follows: strings with a head identical to an official terminal code were truncated to the official code, strings identical with the first three/four digits of a four/five digit terminal code were supplemented with one character for "unspecified" if possible, all other strings were deleted from the forms.

2.1. *Participants*

2.1.1. *Medical students*

One hundred and twenty-nine students undertook courses in Epidemiology, Medical Biometry and Medical Informatics in winter 2003/2004. As part of this course they received 1 h training in diagnoses coding with a former German version of ICD-10-GM abbreviated ICD-10-SGB-V 2.0. As home work, every student got a diagnoses list from one out of 15 discharge letters randomly. One hundred and eighteen students filled out a respective form with codes from the ICD-10-SGB-V 2.0 without typing of diagnoses as principal or secondary. Six to 11 forms were available for each diagnoses list.

2.1.2. *Physicians in medical management*

A call for participation was made in a mailing list of the German Association of Medical Informatics, Biometry and Epidemiology 2004-10-29, addressing physicians responsible for organization of coding, for communication with case managers of health insurance companies, and for process reorganization in hospitals. Thirty-four eligible persons stated their interest. Thirty-four discharge letters were randomly distributed among the 34 persons, such that each person received five different discharge letters and each letter was passed to five different persons. Results were obtained from 27 physicians with five forms each leading to 135 out of 170 possible forms. At least two results were available for each discharge letter. The last form was received 2004-12-08. Coding was done using the ICD-10-GM 2004.

2.1.3. *Coding specialists*

The Hospital of Ludwigshafen offered its participation with the Department of Medical Documentation. While in many hospitals in Germany coding is performed by physicians, special trained personnel codes diagnoses and procedures for reimbursement in Ludwigshafen. Twelve discharge letters were randomly selected from the above-mentioned 34 and coded by 13 coding specialists, also with ICD-10-GM 2004. The results were received completely 2004-12-13 leading to 156 forms.

2.1.4. *Overlapping of the three studies*

Six discharge letters had been coded in all three studies, 12 by physicians as well as by coding specialists. Results of these subgroups defined by overlap are presented to exclude a bias due to different sets of discharge letters. Regularly we refer to the results of the full studies to achieve more precise estimates.

3. Results

Table 2 gives an overview of the study groups. One hundred and eighteen student forms from 15 discharge letters include 516 codes with a mean of 4.4 codes per form. The most frequent code was I10 "essential (primary) hypertension" (38 forms). One hundred and eighteen different codes were used. One hundred and thirty-five manager forms include 751 codes with a mean of 5.6 codes per form. The most frequent code was E66.0 "Obesity due to excess calories" (23 forms). Three hundred and twelve different codes were used. One hundred and fifty-six coder forms include 1137 codes with a mean of 7.3 codes per form. The most frequent code was I10.90 "essential (primary) hypertension, unspecified, without statement of a hypertensive crisis" (51 forms). Two hundred and forty-seven different codes were used. Fig. 3 shows the number of forms categorized for the number of codes used per form.

Table 2 – Characteristics of the study groups

	Manager	Coder	Student
Participants (number)	27	13	118
Discharge letters (number)	34	12	15
Forms (number)	135	156	118
ICD-10-codes per form (mean)	5.6	7.3	4.4
ICD-10-codes per form (S.D.)	3.0	3.9	1.9
Pairs of forms for comparison (number)	212	936	416

3.1. *Principal diagnosis*

Each form in the studies on physicians and on coders includes a single principal diagnosis. Principal diagnosis was defined as "the diagnosis established after study to be chiefly responsible for occasioning the patient's episode of care in hospital". Table 3 presents Cohen's Kappa with 95% confidence limits and the fraction of coincidences over pairs. Most frequent principal diagnosis is N18.89 "other chronic renal failure, not end-stage, stage undefined" (23 forms) in the coders' study and N18.0 "end-stage renal disease" (seven forms) in the managers' study. According to Landis and Koch [11] the results could be interpreted as fair/moderate (manager/coder) for terminal codes and substantial for chapters. The calculation of p_{om} shows similar results as Cohen's Kappa (cf. Table 3).

3.2. *Full set of diagnoses*

We calculated p_{om} for different levels of ICD-10 within each group (cf. Table 4). Duplicates on the higher levels of ICD-

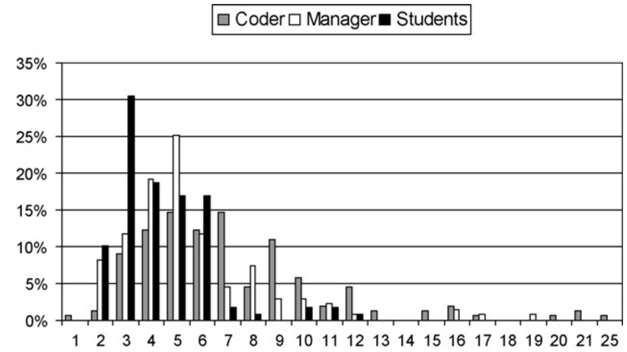


Fig. 3 – Relative number of forms within each study categorized by the number of codes per form (coders left grey column, managers middle white column, students right black column).

Table 3 – Reliability of coding of the principal diagnosis according to different levels of the ICD-10

ICD-level	Manager		Coder	
	N	%	N	%
Number of identical pairs				
Terminal code	62	29.2	438	46.8
Three-digits	128	60.4	641	68.5
Group	145	68.4	707	75.5
Chapter	166	78.3	756	80.8

ICD-level	Manager		Coder	
	Mean	95% confidence limits	Mean	95% confidence limits
Kappa				
Terminal code	0.27	0.22–0.32	0.42	0.39–0.46
Three-digits	0.56	0.50–0.62	0.63	0.59–0.66
Group	0.64	0.58–0.70	0.71	0.68–0.74
Chapter	0.71	0.65–0.77	0.72	0.68–0.75

ICD-level	Manager		Coder	
	p_{om}	95% confidence limits	p_{om}	95% confidence limits
p_{om}				
Terminal code	0.29	0.18–0.40	0.47	0.31–0.63
Three-digits	0.61	0.48–0.73	0.68	0.53–0.84
Group	0.67	0.55–0.79	0.76	0.58–0.93
Chapter	0.78	0.67–0.90	0.81	0.67–0.95

Table 4 – Reliability of coding of the full set of diagnoses according to different levels of the ICD-10

ICD-level	Manager		Coder		Student	
	p_{om}	95% confidence limits	p_{om}	95% confidence limits	p_{om}	95% confidence limits
Terminal code	0.21	0.17–0.24	0.28	0.22–0.34	0.46	0.39–0.53
Three-digits	0.40	0.35–0.46	0.39	0.32–0.47	0.58	0.52–0.64
Group	0.50	0.45–0.55	0.46	0.38–0.54	0.73	0.67–0.79
Chapter	0.64	0.59–0.70	0.60	0.54–0.65	0.87	0.80–0.94

Table 5 – Reliability of coding of the full set of diagnoses according to different levels of the ICD-10 in six discharge letters present in all groups

ICD-level	Manager		Coder		Student	
	p_{om}	95% confidence limits	p_{om}	95% confidence limits	p_{om}	95% confidence limits
Terminal code	0.24	0.09–0.39	0.26	0.14–0.37	0.49	0.37–0.61
Three-digits	0.39	0.22–0.55	0.34	0.22–0.47	0.58	0.47–0.68
Group	0.53	0.38–0.69	0.42	0.29–0.54	0.68	0.59–0.77
Chapter	0.69	0.57–0.82	0.56	0.49–0.63	0.83	0.66–0.99

10 were deleted within each form. Thus, the numbers of codes were reduced from 751/1137/561 to 94/370/151 (manager/coder/student) on the chapter level.

3.3. *Overlapping discharge letters*

Managers' coding is nearly unchanged recalculating p_{om} for the 12 discharge letters used by the coders: terminal code 0.16/0.24/0.33 (lower 95% confidence limit/p_{om}/upper 95% confidence limit), three-digits 0.29/0.40/0.51, group 0.42/0.51/0.60, and chapter 0.57/0.65/0.72. A comparison of p_{om} with six discharge letters present in all studies doesn't alter the ranking between the groups (cf. Table 5). Confidence intervals reveal a large overlap between managers and coders as well as a small overlap between managers and students.

4. Discussion

The coding of diagnoses with ICD-10-GM is of great importance for hospitals in Germany today. Their revenue depends mainly on the coding of diagnoses and procedures that build the definition for DRGs. Appropriateness of care is systematically monitored by a timely communication with health insurance companies using the same codes. In questionable cases an assessment of the correct coding, the appropriateness of admissions and the appropriateness of medical decisions is carried out analyzing the paper record through a service engaged by the health insurance companies. Typically this ends with a discussion on the correct codes [13]. The national program for quality control of inpatients uses a filter with these codes as well. Thus, the inclusion of cases for provider comparisons depends also on the right coding.

Our investigation reveals weak results concerning the reliability of diagnoses coding with ICD-10. Closing of hospitals, patient empowerment, and level of quality will depend on data with only fair to moderate reliability.

A Swedish study characterized the reliability of diagnoses coding as poor [14]. Six general practitioners (GPs) coded 152 medical problems with a subset of 972 codes from a Swedish version of ICD-10. Each GP coded the medical problems in three subsets with different coding tools. For terminal codes the best Kappa was 0.58 (59% agreement) using a book. The value increases up to 0.82 (84%) aggregating the codes to the level of chapters. These results, better as compared to our estimates for agreement in principal diagnoses, could be explained with the smaller coding space of the Swedish study on the one hand. On the other hand the determination of the principal diagnoses requires two decisions, coding the diagnosis and characterizing it as principal.

The IOM analyzed 3301 abstracts from patients discharged in 1974 available for re-abstracting [1]. Selection criteria include a list of 14 so-called target diagnoses and seven so-called satellite diagnoses. The study covered 50 hospitals, from which 48 used four different abstract services. The IOM published in its report a percentage of 65.2% principal diagnoses without discrepancy between the original hospital discharge abstract and the independent re-coding. This figure varies with the principal diagnoses (between 30.2% and 100.0% agreement). As in our study the agreement improves to 74.0% on the level of three-digits. Relationships between other variables concerning personnel & training, abstracting process, and hospital characteristics were judged as difficult to interpret, because of unstable and not meaningful results. In a subgroup, the agreement of IOM's coding with a consultant was examined. The percentage of agreement was clearly higher with 86.1% for terminal codes and 88.1% for three-digits.

As example of a regional re-coding study Dixon et al. analyzed a random sample of 354 and 348 available case notes from two hospitals [15]. Re-coding the principal diagnosis with ICD-9, an external coder agreed with the local coding in 43%/60% (hospital A/hospital B) the four-digits level and 55%/72% the three-digits level. The results concerning the four-digits level are better compared with our study (29.2% manager, 46.8% coders) due to the further use of a five-digit specialization in Germany. At the level of three-digits the results are quite similar (60.4% manager, 68.5% coder). Analyzing the disagreements at the level of three-digits, even a third coder neither confirmed the local nor the external coder in 53%/35%.

Coding the principal diagnosis from a discharge letter, five medical managers will create about three different results. The coders receive significantly better results for terminal codes. The gap between managers and coders nearly disappeared on the chapter level. Managers disagree especially in the refinement of a three-digit code. One can argue that the physicians' medical knowledge may cause a speculative coding, whereas the coders concentrate on the given information. Furthermore, the higher reliability of the coders in terminal codes could be an effect of an internal standardization in Ludwigshafen. The latter explanation is supported by the fact that the managers use each code 2.4-times, the coders 6.4-times in the full set of diagnoses.

The authors present the first study that compares full diagnoses sets. With this set we could detect no significant and meaningful differences between managers and coders in terms of internal agreement. If two persons process independently the same discharge letter with 10 diagnoses they will agree in two or three codes. A comparison of p_{om} for the full set of diagnoses with p_{om} for the principal diagnoses shows inferior results on all levels. Here it is relevant that the definition of a secondary diagnosis according to the German Coding Directives [10] is handicapped by the judgment, whether that diagnoses has led to additional work during hospital stay.

An explicit list of diagnoses could simplify coding in comparison to full discharge letters. The more a coder knows about a clinical case, the more coding alternatives he or she has to consider. That's why medical students achieve more homogenous results on all ICD-levels in comparison to outstanding experts of hospitals (managers) or to well-trained coding specialists of one hospital (coders). Furthermore, students might have used a common sense coding, not influenced by difficult coding rules or clinical speculations about the course of a disease.

Our study might be biased by some methodical restrictions. The discharge letters cover only internal medicine with special emphasis on nephrology. One can argue that coding will lead to a better reliability in medical fields with simpler diagnoses, such as ophthalmology. The terminology and structure used in these letters follow the concept of a single physician. Thus, the conclusions are restricted to the selected set of discharge letters. Coding of managers, coders, and students was not controlled. Persons were able to share their codes to increase agreement as well as to note nonsense on the forms. We have no hints for a bias in any of the two directions. Also p_{om} is until now used only for this investigation and not formally compared to Cohen's Kappa or other indicators of agreement. Nevertheless, the results are plausible and consistent with the available literature. Further reliability studies will profit from a semantic distance measure for ICD-10-codes beyond the hierarchy. Therefore, the support of such a measure is a methodological issue that should be considered in the further maintenance of ICD-10.

Starting with the patient's complaints, many factors could lead to discrepancies in the final diagnosis code, which are independent of the diagnosis classification: communication between the patient and the health professional, decisions on diagnostic and therapeutic procedures, wording in reports, and so on. Some of these factors are also present in our study using discharge letters as source material. So the results presented here do not represent an artificial optimum that is achievable with the ICD-10 in ideal circumstances, it rather represents a realistic measure of what could be expected in daily practice. But, one should have in mind that reliability in diagnosis coding from discharge letters depends not only on the classification used but also on the reliability of the diagnostic process itself.

5. Conclusions

We argue that the stated fair reliability is caused by the extensive refinement of the ICD-10 in Germany, accompanied by the introduction of complex and numerous coding rules. It is obvious to all coding experts that it is impossible to obtain reliable data on such a base. It is surprising, that re-coding studies as presented by Dixon et al. [15] did not recognize the role of the classification itself, even if they conclude a “low level of agreement between coders over main diagnosis and procedure codes”. The results are not a reflection of the coders’ weaknesses. Thus, further education and training, motivation, supplementation of coding rules, and higher standardization will not change the results significantly. Rather we have to be aware that coding in daily practice will be a good deal worse than identified with our study involving experts on a voluntary basis.

Classifications and coding rules have to be radically simplified. It is also erroneous to think that detailed classifications are a prerequisite of a valid image of the clinical situation. Previous studies have shown that detailed data are neither necessary for grouping in the DRG-system [16] nor a realistic performance measure of surgical work [17].

The IOM had mixed the question of reliability and validity in their report, because the discrepancies were judged as coders’ mistakes [1]. The results presented by Nilsson et al. demonstrate that discrepancies are due to ambiguities and other weaknesses of the classifications and not caused by wrong coding of the studies’ participants [14]. We should be aware that ICD-10 offers valid code alternatives in many situations. These alternatives lead to weak results in reliability, but will not influence validity. Notwithstanding the attempt to find one correct code for the principal diagnoses, IOM failed in 10.7% of all abstracts. One has to accept different valid codes for the same clinical situation in a significant number of cases.

Coding with detailed information does not increase reliability. We expect that the results will decline further if coding is done using full records. More information should improve validity but reliability will deteriorate due to an increase in coding alternatives, in necessary decisions about relevancy, etc.

ICD is used worldwide for mortality statistics and for health care reimbursement in many developed countries. Considering the present study together with the studies by the IOM and the Swedish group indicates that the problem of reliability is an international one, independent of the national ICD-version. We assume that this problem is also independent of the type of terminology systems. Thus, a replacement from ICD as classification with a thesaurus like the Medical Subject Headings or a nomenclature like SNOMED CT will not change the point unless studies become available yielding better results for reliability with these instruments. In summary, the reliability of the specific coding system used is a vital issue when it comes to discussing quality management, health care financing or health care policy.

Summary points

What was already known in this field?

- ICD-10 has weaknesses in its structure from a terminological point of view.
- Two coder receive an agreement on the ICD-10-code of the principal diagnoses in about two third of the cases.
- Coded data are increasingly used in health care beyond accounting.

What this study has added to our knowledge?

- Reliability of diagnoses coding using the refined German version of the ICD-10 leads to inferiority results.
- Quality of coding depends scarcely on the profession of the coder.
- Including secondary diagnoses, the set of diagnoses present by patients will not be reliably coded with ICD-10.

Acknowledgements

We are very grateful to all the voluntary participants in our study, who received no additional fee for coding.

References

[1] Institute of Medicine, Reliability of hospital discharge abstracts, Report of a study, National Academy of Sciences, Washington, 1977.

[2] D.C. Hsia, W.M. Krushat, A.B. Fagan, J.A. Tebbutt, R.P. Kusserow, Accuracy of diagnostic coding for medicare patients under the prospective-payment system, N. Engl. J. Med. 318 (1988) 352–355.

[3] L.I. Iezzoni, Assessing quality using administrative data, Ann. Intern. Med. 127 (1997) 666–674.

[4] P.F. Brennan, W.W. Stead, Assessing data quality: from concordance, though correctness and completeness, to valid manipulatable representations, J. Am. Med. Inform. Assoc. 7 (2000) 106–107.

[5] W.R. Hogan, M.M. Wagner, Accuracy of data in computer-based patient records, J. Am. Med. Inform. Assoc. 4 (1997) 342–355.

[6] G. Surján, Questions on validity of international classification of diseases-coded diagnoses, Int. J. Med. Inform. 54 (1999) 77–95.

[7] C.P. Friedmann, G.G. Gatti, G.C. Murphy, T.M. Franz, P.L. Fine, P.S. Heckerling, T.M. Miller, Exploring the boundaries of plausibility: empirical study of a key problem in the design of computer-based clinical simulations, in: I. Kohane (Ed.), Proceedings of the AMIA Annual Symposium, Hanley & Belfus, Philadelphia, 2002, pp. 275–279.

[8] Deutsches Institut für Medizinische Dokumentation und Information (Hrsg.) ICD-10-GM Systematisches Verzeichnis. Version 2004, Internationale statistische Klassifikation der

Krankheiten und verwandter Gesundheitsprobleme, 10. Revision. German Modification, videel, Düsseldorf, 2003.

[9] J. Stausberg, N. Lehmann, Coding training for medical students: how good is diagnoses coding with ICD-10 by novices? GMS Med. Inform. Biom. Epidemiol. 1 (2005) (Doc04 (20050407), http://www.egms.de/en/journals/mibe/2005-1/mibe000004.shtml (in German)).

[10] Deutsche Krankenhausgesellschaft, Spitzenverbände der Krankenkassen, Verband der privaten Krankenversicherung, Institut für das Entgeltsystem im Krankenhaus. Deutsche Kodierrichtlinien. Version 2004, Deutsche Krankenhaus Verlagsgesellschaft, Düsseldorf, 2003.

[11] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.

[12] J.L. Fleiss, B. Levin, M.C. Cho Paik, Statistical Methods for Rates and Proportions, Wiley and Sons, Hoboken, NJ, 2003.

[13] B. Klaus, A. Ritter, H. Große Hülsewiesche, B. Beyrle, H.-U. Euler, H. Fender, M. Hübner, G. von Mittelstaedt, Study of the quality of coding of diagnoses and procedures under DRG conditions, Gesundheitswesen 67 (2005) 9–19 (in German).

[14] G. Nilsson, H. Petersson, H. Åhlfeld, L.-E. Strender, Evaluation of three Swedish ICD-10 primary care versions: reliability and ease of use in diagnostic coding, Method Inform. Med. 39 (2000) 325–331.

[15] J. Dixon, C. Sanderson, P. Elliott, P. Walls, J. Jones, M. Petticrew, Assessment of the reproducibility of clinical coding in routinely collected hospital activity data: a study in two hospitals, J. Public Health Med. 20 (1998) 63–69.

[16] J. Stausberg, Design of classifications for diagnoses and procedures in a DRG system, Gesundh. Qual. Manage. 7 (2002) 297–303 (in German).

[17] J. Stausberg, H. Lang, U. Obertacke, F. Rauhut, Classifications in routine use: lessons from ICD-9 and ICPM in surgical practice, J. Am. Med. Inform. Assoc. 8 (2001) 92–100.