Appropriateness of ICD-coded Diagnostic Inpatient Hospital Discharge Data for Medical Practice Assessment*

A Systematic Review

H. Prins1; A. Hasman2

1School of Health Care, Windesheim University of Applied Sciences, Zwolle, The Netherlands;
2Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Keywords
ICD, diagnostic data, discharge data, sensitivity, positive predictive value

Summary
Objectives: We performed a systematic review to investigate the quality of diagnostic hospital discharge data (DHDD) in order to gain insight in the usefulness of these data for medical practice assessment. We investigated the methods used to evaluate data quality, factors that determine data quality and its consequences for medical practice assessment.

Methods: We selected studies in which both completeness (or sensitivity; SENS) and correctness (or positive predictive value; PPV) were measured. We used the random-effects model to calculate mean SENS and PPV and to explore the effect of a number of covariates.

Results: The 101 included studies were very heterogeneous. We distinguished six typical study designs. We found a mean SENS of 0.67 (95%CI: 0.62–0.73) and PPV of 0.76 (95%CI: 0.73–0.79); SENS was significantly lower for comorbidity and complication studies than for some single disease studies. PPV was significantly higher for Scandinavian countries than for other countries. Recoding compared to re-abstracting of the medical record as a gold standard gave a significantly lower PPV. Diagnostic data were considered appropriate by the authors of the studies for quality of care purposes when both SENS and PPV were at least 0.85. Only 13% of the studies fulfilled this criterion.

Conclusions: Variability in quality of care between settings can easily be overshadowed by variability in data quality. However, the use of DHDD by physicians to evaluate their own medical practice may be useful. But only if physicians are willing to critically interpret the meaning of the information for their medical practice assessment.

1. Introduction

If physicians want to assess retrospectively their medical care for certain patient groups in order to improve their medical practice, information about the diagnoses of the patients is of utmost importance. In most Western countries, diagnoses for each admission in a hospital are coded and electronically recorded after discharge of the patient for statistical, policy or reimbursement reasons [1]. Nowadays, this diagnostic information is coded using ICD-9-CM [2] or ICD-10 [3]. The use of these diagnostic hospital discharge data (DHDD) for medical practice assessment is appealing because of its availability, ease of access, standardization of documentation, complete coverage of all admissions and low additional cost [4]. However, high data quality is not self-evident. Since the 1970s, many studies have been conducted to evaluate the quality of these data for various purposes. One of the first major studies was performed in 1974 by the Institute of Medicine [5]. This study examined the agreement between the principal diagnoses of the hospital discharge data and the corresponding diagnostic information within the medical records. It was concluded that: “Diagnosis-specific discrepancies are of sufficient magnitude to preclude use of such (diagnostic hospital discharge, authors) data for detailed research and evaluation.” It also stated that: “These findings may be particu...
lar timely, since increasingly important decisions about the content of medical care and levels of reimbursement may be based on such information.” But is this situation still the case, or did the quality of data improve in later years? The objective of our review is to get insight in the appropriateness of diagnostic hospital discharge data for medical practice assessment.

Performance indicators are a promising tool to help physicians in a hospital to assess the quality of their specialist care for certain patient groups [6]. In the hospital information system (HIS) the patients with the clinical conditions of interest are identified and within this group of patients events such as comorbidities or complications are then determined. Often the clinical conditions of interest are diagnoses, comorbidities and complications being special forms. In order to get reliable performance indicators, the hospital diagnostic discharge data should be of high quality [7–11], especially in terms of completeness and correctness. To qualify the data in terms of completeness and correctness, a gold standard (GS) is needed to compare the data with, resulting in criterion validity” [12]. Using a GS, a diagnostic code can be true positive (TP), false positive (FP), false negative (FN) or true negative (TN), and this outcome can typically be placed in a 2 × 2 table. Completeness is then equivalent to sensitivity (SENS: TP/(TP + FN)) which is here [13]:

- The proportion of patients’ hospitalizations with a certain diagnosis according to the GS for which the corresponding diagnostic code is present in the hospital discharge record.

Correctness is equivalent to the positive predictive value (PPV: (TP)/(TP+FP)) which is [13]:

- The proportion of patients’ hospitalizations with a certain diagnostic code in the hospital discharge record for which the corresponding diagnosis is present according to the GS.

Both completeness and correctness are required to describe the quality of clinical data, since there is typically a tradeoff between them [14]. Generous allocation of a diagnostic code can lead to high completeness at the cost of correctness. Careful allocation of a diagnostic code can lead to high correctness at the cost of completeness. When identifying cases for medical practice assessment, incomplete data leads to undetected cases and incorrect data to unwanted cases. With regards to complications, incomplete and incorrect data lead to underestimation and bias of adverse outcomes respectively [15, 16]. With regards to comorbidities, incomplete and incorrect data lead to underestimation and bias of risks of adverse outcomes respectively. In the literature, no criteria can be found about minimum values for completeness or correctness of diagnostic data in order to be useful for medical practice assessment.

This systematic review investigates the quality of diagnostic inpatient hospital discharge data as reported in scientific journals. We selected only studies in which – at least – both completeness and correctness were measured. Our research question was: are ICD-coded diagnostic inpatient hospital discharge data appropriate for medical practice assessment? We investigated:

1. the characteristics of the settings where the evaluation took place;
2. the gold standards and designs used to assess completeness and correctness;
3. the values for completeness and correctness and relation with kind of diagnosis, setting, and evaluation method;
4. the determinants of data quality that were reported by the researchers;
5. evidence about the consequences of the data quality for medical practice assessment;
6. the appropriateness of the data for the objectives of the studies, especially for quality of care.

2. Methods

2.1 Selection Procedure

We selected the papers in three steps. First, we used PubMed to perform a sensitive search. We limited our search to studies written in the English language. We developed the search strategy iteratively, learning relevant subject headings and text words that seemed potentially relevant. In line with the advice of the Cochrane Collaboration [17], we also selected some key articles that met the inclusion criteria for the review to note common text words and their variants (such as synonyms, abbreviations and spelling variants) as well as subject headings the database indexers assigned to the articles. However, there are not many subject headings with a meaning related to our topic, so we used many text words and tried many queries in order to find a balance between the number of articles needed to read and a sensitive search. Our query was composed of four sets of terms, combined with “AND”: 1) terms about quality, 2) terms about coded diagnostic hospital discharge data, 3) title words about quality or coded diagnostic hospital discharge data and 4) relevant MeSH headings. We used a previously defined set of eligible studies to optimize the query, which was executed on March 21st, 2006 (Supplementary Online File (Appendix A) for the complete query).

Second, the two researchers manually selected, by applying in- and exclusion criteria (see below) to the title and abstract, a set of possibly eligible studies. Possibly eligible were those studies for which the abstracts did not make it clear that they could be excluded. About half of the abstracts was independently evaluated by both researchers. In case of disagreement, both researchers had to come to consensus and the interpretation of the criteria was discussed to increase interrater reliability. The rest of the abstracts was equally divided among the two researchers.

Thirdly, the two researchers independently screened the resulting set of possibly eligible studies based on full-text papers, again using the list of in- and exclusion criteria. In case of disagreement, both researchers had to come to consensus.

2.2 In- and Exclusion Criteria

Included were studies that fulfilled all of the following criteria; the study:

- evaluated ICD coded diagnostic hospital discharge / administrative / claims data (when DRG’s were evaluated, the study could be included if also the quality of the underlying ICD codes was evaluated);
- reported the evaluation of inpatient hospital discharge codes separately in
used a GS as a reference for data quality;
- measured both completeness (SENS)
  and correctness (PPV); r
- reported absolute numbers for TP, FP
  and FN or gave sufficient information
  to compute these numbers a.

Excluded were studies that fulfilled one
or more of the following criteria:
- reviews;
- only about, or mingled with out-patient
data b;
- only about, or mingled with procedural
data c;
- only about, or mingled with physician
claims data d;
- only about, or mingled with E-codes
(external cause of injury) e, f;
- with an ‘experimental’ design instead of
evaluation of routinely collected data;
- with PPV and SENS not sharing the
same set of TPs.

2.3 Data Extraction

Both researchers independently extracted
the necessary data from the selected full-
text papers and the results were discussed
in order to come to consensus. We used a
data extraction form (Supplementary
online File (Appendix B)). Data were
extracted about:
- setting of study (country, type of hospi-
tal, number of hospitals, who coded,
etc);
- routine use of data;
- objectives of data evaluation;
- kind of diagnostic data;
- data evaluation methods;
- numbers of TP, FP, FN and – if measured – TN;
- data quality in terms of SENS, PPV and
  – if possible – SPEC;
- authors’ conclusions about determinants
data quality and usability of data.

For each study we constructed one $2 \times 2$
table. If the authors themselves reported
overall measures of TP, FP, FN (and – if pos-
sible – TN) and the corresponding SENS
and PPV (and – if possible – SPEC) even when
several disease subcategories were distin-
guished, we extracted the overall numbers. If
no overall measures were reported, we
pooled TPs, FP, FN (and – if possible – TNs) of
disease subcategories into one $2 \times 2$
table, computed the SENS, PPV (and – if
possible – SPEC) and labeled the pooled data
with an appropriate disease category. If
numbers were given over several periods, we
extracted only the numbers of the latest peri-
od. In case of intervention studies, we only
used the post intervention data.

2.4 Data Analysis

We did a meta-analysis to estimate the
overall SENS and PPV averaged over the
studies. Since the studies were very hetero-
genous (because of different hospitals and
coding settings, populations, diseases and
evaluation methods), we used the random-
effects model of Dersimonian and Laird
[18] instead of a fixed-effects model to cal-
culate the pooled SENS and PPV. The ran-
dom-effects model explicitly takes be-
tween-study variance into account and es-
timates this variance. We used no trans-
formations. To determine the weighted
average of SENS (or PPV), the estimate of
SENS (or PPV) of each individual study was
multiplied by a weight, being the in-
verse of the sum of the variance of SENS
(or PPV) of the individual study and the es-
timate of the between-study variance (Tau-
squared). The mean SENS (or PPV) was
then computed as the sum of the products
(SENS or PPV multiplied by weight) di-
vided by the sum of the weights. The be-
tween-study variance was also included
when calculating the 95% confidence inter-
val (95% CI). We used the $I^2$-statistic
to measure the heterogeneity of the studies.
The $I^2$-statistic represents the percentage
of variability in SENS and PPV estimates due
to heterogeneity rather than chance.

The random-effects model was also
used to determine the effect of a number of
covariates on both SENS and PPV. We lim-
ited our subgroup analysis to predefined
covariates related to kind of diagnostic
data, period of original coding, setting and
methodological issues. We hypothesized
that disease category, coding year, country
and type of GS would influence the ob-
erved data quality. Within the studies that
used the medical record as GS we also ana-
lyzed the influence of blinding and re-
coding. In a preliminary study in 2003,
we tested that these covariates could be
of influence were explicitly tested [19].

We calculated mean SENS and PPV
with their 95% CI using Excel® based on
the random-effects formulas of Borenstein
et al. [20]. We used the t-test to compare
subgroups of studies. We first computed
(with the random-effects formulas) the
mean SENS (or PPV) and standard errors
of two subgroups independently. Then we
used the two independent means and
standard errors for the t-test. Since we did
multiple tests, we chose a conservative sig-
ificance level of 0.01 multiplied by $1/k$
where $k$ is the number of hypothesis tests
within a group (Bonferroni correction).

Many authors of the included papers
discuss determinants of diagnostic data
quality. We mention only those determin-
ants that were significant after a statistical
analysis like Pearson Chi-square tests or
multivariate regression analyses.

Based on the conclusions of the authors
in each individual study, we qualified the
authors’ opinion about the appropriateness
of the DHDD for their specified (second-
ary) purposes as ‘yes’, ‘no’ or ‘doubt’. For
threshold setting we constructed a boxplot
with SENS and PPV for the different
qualifications of appropriateness.

3. Results

3.1 The search

Our PubMed search yielded 4040 results,
out of which 283 were labeled, based on
title and abstract, as “possibly eligible”.
Agreement between the two researchers
over the first 2081 abstracts was 0.917
and Kappa was 0.651. After a manual review
of 276 full-text papers (seven papers could

a This makes it possible to carry out a meta-analysis
b If a study also includes, and explicitly reports about inpatient diagnostic codes, we did not exclude the study.
c ICD-codes from Chapter ‘Injury and Poisoning’
(ICS-9-CM) do not belong to this exclusion-crite-

\[ \text{PPV} = \frac{TP}{TP + FP} \]

\[ \text{SENS} = \frac{TP}{TP + FN} \]

\[ \text{SPEC} = \frac{TN}{TN + FP} \]

\[ \text{TN} = \text{True Negative} \]

\[ \text{TP} = \text{True Positive} \]

\[ \text{FP} = \text{False Positive} \]

\[ \text{FN} = \text{False Negative} \]

\[ \text{SPEC} = \text{Specificity} \]

\[ \text{PPV} = \text{Positive Predictive Value} \]

\[ \text{SENS} = \text{Sensitivity} \]

\[ \text{PPV} = \text{Prevalence} \]

\[ \text{SPEC} = \text{Prevalence} \]

\[ \text{TP} = \text{True Positive} \]

\[ \text{FP} = \text{False Positive} \]

\[ \text{FN} = \text{False Negative} \]

\[ \text{TN} = \text{True Negative} \]

\[ \text{PPV} = \text{Positive Predictive Value} \]

\[ \text{SENS} = \text{Sensitivity} \]

\[ \text{PPV} = \text{Prevalence} \]

\[ \text{SPEC} = \text{Prevalence} \]
not be obtained), 101 studies were included based on our eligibility criteria. Agreement between the two researchers about the full-text based paper selection was 0.826 and Kappa was 0.645. The reasons for exclusion are given in ▶ Table 1.

A table with extracted data of all studies can be found in the ▶ Supplementary Online Files (Appendix C).

The number of studies included per publication year increased gradually from one or no study in the eighties to 15 studies in 2005.

3.2 Characteristics of the (Coding) Settings Where the Evaluation Took Place

Almost half of the studies came from the USA; the rest of the studies came from Anglo-Saxon, Scandinavian and some other European countries.

In 37 studies, the evaluation was performed in a single hospital setting. In 64 studies the evaluation setting consisted of more than one hospital: 18 studies with 2–10 hospitals, 22 studies with 11–100 hospitals, 7 studies with > 100 hospitals and 17 with an unknown number of hospitals. In 39 studies the evaluation took place only in general hospitals, in 16 studies only in university hospitals, in 32 studies in a combination of general and university hospitals and in 32 studies the type of hospital was not specified.

▶ Table 2 shows, by country, the reasons for coding and who is coding. Reimbursement as reason for coding is mainly limited to studies done in the United States and other Anglo Saxon countries, and the doctor as coder is relatively often seen in Scandinavian studies.

In 8 studies, the use of ICD-8 was evaluated. In respectively 34, 49 and 10 studies ICD-9, ICD-9-CM and ICD-10 were evaluated.

3.3 Gold Standards and Designs Used to Assess Completeness and Correctness

First we distinguished studies that evaluated diagnostic data quality of 1) all kinds of diseases together, 2) a single disease or 3) all or some diseases in a subgroup of patients who a) had a specific disease or b) underwent a specific procedure. For subgroups of patients with a specific disease, the interest is on the quality of morbidity or complication data. For subgroups of patients who underwent a specific procedure comorbidities, complications or principal diagnoses were studied.

Secondly we distinguished studies according to the GS that was used: 1) the medical record, 2) a disease specific registry or 3) prospectively collected diagnostic data.

When the medical record was used as GS, the re-abstracted or recoded medical record was compared with the diagnostic hospital discharge codes. Electronically available test results were considered to be part of the medical record. In re-abstraction relevant information from the medical record is retrieved, but not recoded. Reabstracting took place blinded or not blinded for the original codes although blinding was not always (clearly) specified.

After re-abstraction the re-abstractor de-

### Table 1

<table>
<thead>
<tr>
<th>Source</th>
<th>Title + Abstract</th>
<th>Full-text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of studies</td>
<td>4040</td>
<td>283</td>
</tr>
</tbody>
</table>

**Reason for exclusion**

<table>
<thead>
<tr>
<th>Reason for exclusion</th>
<th>—</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-text paper could not be obtained</td>
<td>2828</td>
<td>9</td>
</tr>
<tr>
<td>Not about ICD-coded hospital discharge data</td>
<td>745</td>
<td>36</td>
</tr>
<tr>
<td>ICD codes are used in study, but not evaluated</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Review study</td>
<td>71</td>
<td>37</td>
</tr>
<tr>
<td>Inpatient hospital diagnostic data not (separately) evaluated, but about (or mingled with) outpatient, physician, procedure or ‘cause of injury’ data</td>
<td>77</td>
<td>30</td>
</tr>
<tr>
<td>No (correct) criterion validity of routinely collected data</td>
<td>30</td>
<td>62</td>
</tr>
</tbody>
</table>

**Number of (possibly) eligible studies**

<table>
<thead>
<tr>
<th>Country</th>
<th>Policy</th>
<th>Management</th>
<th>Reimbursement</th>
<th>Quality</th>
<th>Research</th>
<th>N.S**</th>
<th>Doctor</th>
<th>Trained coder</th>
<th>N.S**</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States (n = 45)</td>
<td>24</td>
<td>5</td>
<td>29</td>
<td>15</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>32</td>
<td>12</td>
</tr>
<tr>
<td>Other Anglo Saxon countries (n = 31)</td>
<td>20</td>
<td>2</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>21</td>
<td>8</td>
</tr>
<tr>
<td>Scandinavian countries (n = 15)</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Other European countries (n = 10)</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>All countries (n = 101)</td>
<td>62</td>
<td>7</td>
<td>45</td>
<td>29</td>
<td>19</td>
<td>15</td>
<td>12</td>
<td>56</td>
<td>33</td>
</tr>
</tbody>
</table>

* Reasons for coding categories not mutually exclusive ** N.S.: not specified
cides whether the re-abstracted data confirm the original code. In case the re-abstracter is not blinded for the original code the results of re-abstraction may be biased. In recoding, the medical record is re-abstracted and again coded, and the newly obtained codes are compared with the hospital discharge codes. Recoding is always done blind for the original codes.

In Table 3 the number of studies as a function of type of GS and range of diseases is given. In 35 of the 61 studies where the medical record was used as GS, re-abstraction took place of which 14 blinded. Recoding took place in the other 26 studies; in 12 studies a medical record coder was the re-coder, in 9 studies a doctor (in 5 studies not specified). Agreement between two or more re-abstractors or recorders was reported in 21 studies. In 23 studies, clinical criteria (instead of, or in addition to ICD criteria) were applied to re-abstract or re-code. Clinical criteria to diagnose a disease are formulated by a professional organization of specialists.

In 20 of the 31 studies where a disease specific registry was used as GS, a patient ID in combination with a time frame was used for record linkage in order to match the hospital discharge data with the registry data. In 4 studies more or less anonymous patient data with a time frame was used for record linkage, e.g. via birth date, sex and Zip code. In 7 studies the matching procedure was not specified. A time frame was used to limit matching to those disease registry events having a date stamp close to the period of hospitalization (e.g. within 28 days before or after the admission date [21]).

In the 11 studies that used prospectively collected data as GS, already available diagnostic data prospectively collected for clinical research or other purposes during hospitalization of the patients, were ‘gratefully’ used as GS.

There is a relationship between the type of GS and the range of diseases studied. In studies where all kinds of diseases were combined, only the medical record was used as GS. Most of the studies evaluating diseases in clinical subgroups also use the medical record as GS. Studies that evaluated single diseases used the medical record, a disease specific registry or prospectively collected data.

When matching the DHDD with the GS data, studies differ in the number of digits of the ICD codes that have to match exactly for agreement; studies compared the GS data at the 5-, 4-, or 3-digit level of codes, or with a group of codes in respectively 5, 13, 31 and 52 studies.

We inferred the following typical designs to evaluate diagnostic data quality in terms of completeness and correctness:

- **Design 1 – all kinds of diseases, GS is medical record**
  A (random/stratified/every nth case/consecutive) sample of hospitalizations is drawn from a routine discharge database. Then the coded diagnoses are compared with those in the corresponding recoded medical record. In case of re-abstraction it is determined whether the information confirms the code. All data are represented by one 2x2 table, resulting in one SENS and PPV per study. Some large studies also report SEN and PPV for a number of single diagnoses with relatively high prevalence – 18 studies [5, 13, 16, 22–36].

- **Design 2a – single disease, GS is medical record**
  First a GS is determined/constructed for a specific diagnosis (e.g. all true myocardial infarctions) based on a comprehensive medical record review with abstracting or recoding. Then all cases with the hospital discharge code(s) of interest are compared with the GS – 14 studies [37–50].

- **Design 2b – single disease, GS is disease specific registry**
  The GS is based on a disease specific registry. All cases from the hospital discharge database with the relevant code(s) are compared with the GS, sometimes combined with a medical record review for extra or additional (in case of non-matching) verification – 25 studies [21, 51–74].

- **Design 2c – single disease, GS is prospectively collected diagnostic data**
  The GS is constructed prospectively and independently of coding practice, e.g. cases included in a clinical study of a specific disease. Then all cases with the relevant code(s) from the hospital discharge database are compared with the GS – 8 studies [75–82].

- **Design 3a – all or some kinds of diseases in patients having a specific disease**
  (A sample of) all admissions in which the patient had a specific disease is obtained and for these selected cases a verification takes place of the coded comorbidities or complications using a GS, most of the times the medical record – 12 studies [83–94].

- **Design 3b – all or some kinds of diseases in patients who underwent a specific procedure**
  (A sample of) all admissions in which the patient underwent a specific procedure is obtained and for these selected cases a verification takes place of the coded comorbidities or complication data using a GS, most of the times the medical record – 24 studies [95–118].

The collection of studies contained three intervention studies: doctors that code [36], a discharge letter-linked diagnosis registration [13] and a database for clinical use as origin of discharge data [35].

In general one can conclude that the studies were very heterogeneous and that

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Number of studies as a function of type of GS and range of diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GS – Main</td>
</tr>
<tr>
<td></td>
<td>Medical Record</td>
</tr>
<tr>
<td>Range of</td>
<td>All kinds of diseases</td>
</tr>
<tr>
<td>diseases</td>
<td>Single disease</td>
</tr>
<tr>
<td></td>
<td>Diseases in clinical subgroup</td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
there is no standard procedure of how to determine diagnostic data quality. However, each study can be classified in one of the six typical designs that can be distinguished based on the range of diseases evaluated and the type of GS used.

### 3.4 Values for Completeness and Correctness and Relation with Kind of Diagnoses, Years, Settings and Evaluation Methods

Figure 1 shows the SENS and PPV of each study for the disease categories that we distinguished. Figure 2 shows the mean SENS and PPV with 95% CI of all studies together and of subgroups. The subgroup analyses for country and GS were limited to the studies that evaluated a single disease (n = 47) since only within this set of diseases the four groups of countries and three types of GS’s were reasonably distributed over the studies. Studies that evaluated all kinds of diseases (n = 18) or diseases in a clinical subgroup (n = 36) were almost always done in the USA or other Anglo Saxon countries and not in Scandinavian or other European countries. For the effect of blinding and recoding for GS construction we limited the set to studies that used the medical record as GS and we further excluded all ‘Single disease’ studies since these studies were overrepresented in the non-blinded re-abstracting studies.

The I²-statistic for all studies together as well as for subgroups of studies was almost always > 98% which means that the studies (even within subgroups based on common characteristics) were highly heterogeneous. Complications and comorbidities had a statistically significantly lower mean SENS than AMI (and other single diseases) had. Complication and comorbidity studies were strongly correlated with clinical subgroup studies and secondary diagnoses. So not surprisingly, studies that evaluated diagnoses in clinical subgroups (21 of these 36 studies evaluated comorbidities or complications) had a statistically significantly lower mean SENS than studies that evaluated a single disease. Studies that only evaluated secondary diagnoses (22 of these 23 studies evaluated comorbidities or complications) had a statistically significantly lower mean SENS (0.50; 95% CI 0.41–0.59) than studies that evaluated only principal diagnoses (0.73; 95% CI 0.70 to 0.77, n = 14) or a combination of principal and secondary diagnoses (0.72; 95% CI 0.65–0.79, n = 64) (results not presented in Figure 2).

We did not find an improvement in later years compared to earlier years of coding ('years' as a dichotomous factor: before and from ’94 - the median of coding years), while the characteristics of the studies (disease categories, countries, GS used etc.) in earlier and later years were quite comparable. Related to the year of coding, also the version of ICD had no influence on SENS or PPV (results not shown here). With regards to countries, PPV was statistically significantly higher for Scandinavian studies.
There were no differences between types of GS. Of studies that used the medical record as GS, studies with recoding had statistically significantly lower mean PPV than studies with blinded re-abstraction. Application of clinical criteria instead of ICD-criteria for re-abstracting or recoding had no effect on SENS or PPV (results not shown in Figure 2).

3.5 Determinants of Data Quality Reported by the Researchers

Several authors reported determinants of data quality obtained via a statistical analysis like Pearson Chi-square tests or multivariate regression analyses. We categorized these determinants in diagnostic type dependent, disease dependent, disease manifestation dependent, patient dependent and hospital dependent determinants. Since these determinants of the quality of diagnostic data were obtained in separate studies they may not be generalizable.

With regards to the diagnosis dependent determinants it was found that data quality is higher for:
- principal than for secondary diagnoses [52, 109].

With regards to the disease dependent determinants it was found that data quality is higher for:
- more severe than for less severe diseases [106, 114];
- symptomatic than for asymptomatic diseases [105];
- common than for rare disorders [26].

With regards to the disease manifestation dependent determinants it was found that data quality is higher for:
- patients in a later stage of a disease [67];
- patients having longer present [50], more frequently occurring [50], more

![Figure 2](Forrest plot of SENS and PPV with mean and 95% CI (random-effects model) for all studies together and for subgroups based on covariates. * Statistically significant; ** Gold standard construction based on medical record)
Methods Inf Med 1/2013 © Schattauer 2013

H. Prins, A. Hasman: ICD-coded Diagnostic Data for Medical Practice Assessment

severe [83], or more [80] manifestations of the disease;
• patients having no history of the disease [53] (explanation: true cases of AMI with history were often falsely coded as CHF);
• patients having a disease confirmed with an important test than for patients where such a test had not been performed [80].

With regards to patient dependent determinants it was found that data quality is higher for:
• infants than for older children [57];
• neonatal than for maternal patients [117];
• patients discharged alive than for patients who died during hospitalization [83] (possible explanation: when patients died the need to code completely was less felt);
• patients who also have comorbidities [57, 67];
• patients not having other, more severe diseases [45];
• patients undergoing surgery for the disease [67, 119];
• patients having a clear risk-factor for the disease (e.g. smoking for AMI)[53].

Some findings regarding patient dependent determinants are conflicting. Data quality is found to be higher for younger patients [42, 72, 80] on one hand and for patients over 63 [39] or 64 years old [42, 106] on the other hand. The latter may be due to the influence of the Medicare reimbursement system for patients over 64 years old where DHDD are necessary for payment and the on average higher disease severity in patients over 64 years old. Some studies found that data quality is higher for male patients [53, 65, 71], other studies for female patients [39]. Some studies found that data quality is higher for white patients [71], other studies for non-(Hispanic)white patients [39]. This can possibly be explained by the fact that acute myocardial infarction (AMI) is better diagnosed in white male patients than in female patients [120] and non-(Hispanic)white patients [121], resulting in higher data quality for white males in case of AMI. As a consequence, the ICD-code for heart failure would be used more in female and non-(Hispanic)white patients, resulting in higher SENS (and thus data quality) for this code. Both shorter length of stay (LOS) and longer LOS were associated with better data quality. Shorter LOS is on average correlated with less comorbidity which is often underreported. This can explain why a shorter LOS results in higher completeness. On the other hand, a longer LOS can also lead to higher completeness since in general more test results are available before discharge of the patient [24].

With regards to hospital dependent determinants it was found that data quality is higher for:
• public than for private hospitals [31, 43, 92];
• teaching/university than for non-teaching/university hospitals [96, 109];
• regional than for local hospitals [65];
• urban than for rural hospitals [92];
• big than for small hospitals [48];
• hospitals having a higher volume of procedures [110].

Several multicenter studies reported differences in data quality between hospitals or regions for the same diseases. Some researchers reported differences between specialties [42], others between hospitals or regions [33, 65, 67, 70, 71, 80, 91]. Often it is not clear where these differences come from, but authors speculate about reasons that have to do with the local coding practices and settings: differences in diagnostic practices, differences in case-mix, clarity and structure of documentation of diagnoses in the medical record, training of medical record coders, access to patient data, time per case to code and coding instructions.

Some researchers explicitly reported factors that did not influence data quality: year of coding [103], insurance status [62], code position [41], age [103], gender [42, 103] and ethnicity [62].

Of the three studies that evaluated interventions to improve data quality, Yeoh et al. [36] found that the participation of doctors in coding leads to higher accuracy (SENS and PPV). Prins et al. [13], however, could not prove their hypothesis that the participation of doctors in coding by means of a discharge letter-linked diagnosis registration would improve data quality. Van Walraven et al. [35] found that coding by doctors with the help of a clinical database (instead of the standard chart review by medical record coders) significantly improved completeness and correctness of secondary diagnoses.

3.6 Impact of Data Quality on Medical Practice Assessment

Losina et al. [110] reported a substantial overestimation of the true effect of avascular necrosis (AVN) on the risk of perioperative dislocation when using AVN diagnoses derived from hospital discharge data compared to this risk when using diagnoses derived from medical record data. They also found that hospital discharge diagnoses led to an 80% overestimation of the association between low functional status three years after total hip replacement and rheumatoid arthritis. This finding was the result of selection bias due to the more sensitive coding of severe cases.

Rinaldi et al. [82] observed a large difference in case fatalities in FP and FN stroke cases (respectively, one month after onset of the disease, 32.7%–36.8% vs. 6.9%–21.1%). According to them, this may signify that patients could be misclassified on the basis of clinical severity, by coding different high mortality disorders as stroke cases, and not coding minor strokes as such. As a consequence the case fatality rate of ischemic stroke could be overestimated.

Sapsford et al. [48] on the other hand, found that hospital coding misses a substantial proportion (22.5%) of AMI cases, but without any apparent systematic bias, and thus provides a suitably representative and robust basis for audit.

Romano et al. [113] found that half of the difference in risk-adjusted complication rates between low and high outlier hospitals was attributable to reporting variation.

3.7 Appropriateness of the Data for the Objectives of the Studies

The purposes of the studies could be categorized into ‘Quality of Care’ (n = 33), ‘Research’ (n = 29), ‘Surveillance’ (n = 34)
It appeared that the criteria for the appropriateness for quality of care purposes were stricter than for other purposes (▶ Figure 3). If we take the first quartile as a minimum, SENS and PPV should both be at least 85% for quality of care purposes. In only 13 [21, 36, 42, 50, 51, 55, 61, 68, 72, 91, 93, 98, 110] of all 101 included studies the data quality met this criterion. Five of these studies were performed in Scandinavian countries and eight studies were about single diseases.

4. Discussion

4.1 Heterogeneity of Studies

We carried out a systematic review to get insight in the appropriateness of diagnostic hospital discharge data for medical practice assessment. The retrieved studies were very heterogeneous. In Scandinavian countries for example, the diagnoses are often encoded by physicians, whereas in other countries this is often done by a medical record coder. Some studies evaluated the data quality of a broad range of diseases, whereas other studies evaluated data quality of a single disease or diseases within a clinically defined subgroup of patients. The medical record, disease specific registries and prospectively collected data were used as GS. We could distinguish six typical designs based on range of diseases studied and GS used.

4.2 Findings

There was a great variability in observed SENS and PPV between the individual studies. Our meta-analysis using the random-effects model resulted in an average SENS of 0.67 (95% CI: 0.62–0.73) and PPV of 0.76 (95% CI: 0.73–0.79). For some single disease studies SENS was significantly higher and for comorbidity and complication studies SENS was significantly lower. When the medical record was used to construct a GS, recoding of diagnoses gave a significantly lower PPV than blind re-abstraction of diagnoses. PPV was significantly higher for Scandinavian countries compared to non-Scandinavian countries. Determinants of data quality, obtained via statistical analysis, could be
categorized in diagnostic type dependent, disease dependent, disease manifestation dependent, patient dependent and hospital dependent determinants. Severity of a disease and severity of the manifestations of a disease lead to a higher chance of the disease to be coded. Studies that evaluated the effect of data quality on quality of care estimates showed that the use of diagnostic hospital discharge data can easily lead to an overestimation of sentinel outcomes. The cases selected for medical practice assessment based on these data represented the more severe cases. From studies with quality of care purposes, we inferred from the authors’ opinion about appropriateness that both SENS and PPV of the diagnostic data should be at least 85%. Only 13% of all included studies fulfilled this criterion.

### 4.3 Coding Performance and Data Quality

SENS and SPEC are an expression of coding performance and are independent of prevalence. They indicate how well coders are able to detect and code diagnoses (SENS) and how well coders are able not to code non-existing diagnoses (SPEC). As such, SENS and SPEC also are an expression of data quality. However, in most studies SPEC is not reported or only roughly estimated. Theoretically it can be argued that SPEC in most of the studies is very high, about 0.98 or 0.99. Studies in which SPEC was measured, showed this. Romano et al. [115] for example wrote: “We did not report specificity ... because this parameter was never below 97%, and nearly always exceeded 99%.” PPV is determined by the values of SENS, SPEC and prevalence. Since prevalence is a factor that cannot be influenced by the coding process, PPV is not an expression of coding performance. Nevertheless, PPV is a very valuable measure for data quality and determines the informational value or usability of the codes in a specific setting, namely the chance that the code represents a true diagnosis.

### 4.4 Do Differences in SENS and PPV Indicate Different Coding Performance?

Studies evaluating some specific diseases showed higher mean values of SENS than other studies. This was also true for ‘single disease’ and ‘principal diagnosis’ studies (these studies were highly correlated with specific diseases). We were curious whether the higher mean SENS indeed was due to a better coding performance or instead indicated a more generous (but not better) coding (a threshold effect). In case of better coding we – mathematically – expect also a higher PPV. The magnitude of this increase depends on the value of the prevalence [122]. In case of more generous coding we expect a higher SENS but lower SPEC (threshold effect), and thus a less – mathematically – expected increase or even a decrease of PPV. The lower the prevalence, the bigger the chance that PPV decreases. Since the often unknown prevalence of many diseases will be quite low, a more generous coding will usually lead to an increase of SENS and a decrease of PPV. The higher SENS for some subgroups in our analyses was accompanied by an unchanged or increased (but not statistically significant) PPV. Since differences in prevalence between subgroups may exist [122], we only cautiously conclude that coding performance is better in these subgroups. Differences between subgroups of studies are observational in nature and are prone to bias and confounding [123]. Other factors could be responsible for the differences.

We found that mean PPV is significantly higher for Scandinavian countries compared to the USA and other countries while mean SENS did not significantly differ. We also found that studies using a recoded medical record as GS had a significantly lower mean PPV, but no significantly different mean SENS than studies using a blindly re-abstracted medical record. If prevalence would explain the differences in PPV, then coding performance would not differ between the two groups of studies. Coding performance can be represented by the diagnostic odds ratio (DOR). This DOR can be expressed in terms of SENS and SPEC with the formula:

\[
\text{DOR} = \frac{\text{SENS} \times \text{SPEC}}{(1 - \text{SENS}) \times (1 - \text{SPEC})}
\]

We found that mean PPV is significantly higher for Scandinavian countries compared to the USA and other countries while mean SENS did not significantly differ. We also found that studies using a recoded medical record as GS had a significantly lower mean PPV, but no significantly different mean SENS than studies using a blindly re-abstracted medical record. If prevalence would explain the differences in PPV, then coding performance would not differ between the two groups of studies. Coding performance can be represented by the diagnostic odds ratio (DOR). This DOR can be expressed in terms of SENS and SPEC with the formula:

\[
\text{DOR} = \frac{\text{SENS} \times \text{SPEC}}{(1 - \text{SENS}) \times (1 - \text{SPEC})}
\]
possibly due to physicians’ involvement in the coding process.

Applying the same sensitivity analysis on recoding compared to blinded re-abstracting also produces an unlikely two times lower prevalence and DOR for recoding (Table 4). Recoding therefore leads to a lower observed coding performance than re-abstracting. Dixon et al. [24] and Prins et al. [13] showed that a substantial part of the disagreements between DHDD and the GS were due to the fact that DHDD represents conditions that were closely related to, but not covered by the codes representing the GS. Compared to a recoded GS, matching with a re-abstracted GS leaves more room for subjective assessment of the DHDD and may be influenced by knowledge of the GS, known as test review bias [126]. This makes it plausible to assume that, given the observed PPVs, re-abstracting compared to recoding leads to an overestimation of data correctness.

4.5 Unexpected Findings

For rare diseases we expect very low PPVs. This is the logical consequence of our expectation of a similarly high SPEC but a much lower PREV than for common diseases. However, several authors report PPVs for rare diseases that are comparable to, or only somewhat lower than those for common diseases (e.g. Bogliun et al. [52]: PPV of 0.55 for Guillain-Barré Syndrome; Beghi et al. [38]: PPV of 0.60 for ALS (Amyotropic Lateral Sclerosis); Chancellor et al. [54]: PPV of 0.70 for Motor Neuron Disease). Possibly, medical record coders may not consider a code for a rare disease unless there is a very clear indication in the medical record, while they may consider a code for a common disease even when there is a less clear indication. This will result in a higher SPEC for rare diseases compared to common diseases which can explain the higher than expected PPV for rare diseases.

4.6 Gold Standard

The observed data quality is partly the result of the GS quality. Romano et al. [114] showed that not only coding, but also recoding for GS construction is susceptible to interrater variability. They compared DHDD and independently recoded ICD-9-CM data with complications abstracted from the medical record by clinicians using detailed criteria. The recoded data captured 56% of all severe complications, whereas DHDD data captured 44%. According to Dixon et al. (24), expert recoders may have had access to information added to the notes after the local coding was done. This indicates that retrospective GS construction and original coding may not be based on the same information. One can also question whether the GS really includes all true cases. Sometimes the capture-recapture method is used, e.g. Nielsen 1996 [68], leading to a somewhat higher estimate of the number of true diagnoses and thus to a lower SENS. Some studies possibly measured data accuracy at the patient level including several hospitalizations, although this is not clearly specified. If so, it can lead to overestimation of SENS and PPV from the point of view of data quality at admission level.

4.7 Improvement over the Years

We did not find an improvement in data quality over the years. However, Pajuunen et al. [70] saw an increase in data quality in Finland in the years 1998–2002 compared to 1988–1992. Leibson et al. [127] showed an increase of SENS over time (1970 to 1989) in the USA while PPV remained the same, due to the introduction of the prospective payment system in 1982. In our meta-analysis, data from only few studies date from 1981 or earlier, so we were not able to analyze the effect of the prospective payment system. We also did not find an improvement of data quality for successive versions of the ICD. Quan et al. [128] compared ICD-9-CM and ICD-10 coding of the same set of hospitalizations, but could also not demonstrate that ICD-10 did perform better than ICD-9-CM.

4.8 International Classification of Diseases

Some authors criticize the limitations of the ICD. Romano et al. [113] mentioned the vague definitions of ICD codes, Bogliun et al. [52] stated that Guillain-Barré syndrome is included in code 357.0 and does not have a separate diagnostic code; Tetsche et al. [72] warned that borderline tumors could not be excluded from the code for ovarian cancer, and McNaughton et al. [77] stated in general that in ICD there is no one single code for a disorder, the codes are not mutually exclusive and definitions are more based on pathological than on clinical information. Cimino [4] argues that some of the problems of ICD-coded data relate to the design aspects of the terminology, such as lack of detail limited by the restrictive nature of the numbering system, the strict hierarchical structure and changes of meaning of terms when the terminology is updated.

4.9 Limitations of the Study

The limitation to studies written in English led certainly to the exclusion of studies...
only published in languages like Danish, Spanish or Italian. However, we were not able to understand these languages and to use such studies for the review.

Compared to the fixed-effects analyses, the random-effects analyses yielded in a higher overall SENS (0.67 vs. 0.61) and a lower overall PPV (0.76 vs. 0.87). In our heterogeneous set of studies, the random-effects meta-analysis awarded relatively more weight to smaller studies than such studies received in the fixed-effect meta-analysis. This means that the smaller studies had higher SENS, but lower PPV then the larger studies. It remains unclear whether this is the result of publication bias, methodological differences in study design or true differences in data quality. Probably all three explanations are valid. We also think that settings where value is placed on data quality have a higher chance of critically evaluating their diagnostic data and publish the results. Many of the included studies had methodological flaws and in diagnostic studies methodological shortcomings have been shown to lead to overestimation of the accuracy of diagnostic tests [129]. Our study showed that compared to abstracting, recoding was associated with lower data quality. Bias may also have incurred by restricting the review to studies that could be meta-analyzed. However, we think that studies that could not be meta-analyzed were in general methodologically less sound, and thus even more prone to bias. This may also be true for studies that can be found in the ‘grey’ literature which we did not take into account. All in all, our overall picture based on the included publications may even be too optimistic.

The strength of our review is that we could analyze the effects of factors that could not be analyzed in individual studies, e.g. the type of GS used. So, we could also determine typical designs to evaluate diagnostic data quality.

4.10 Can Hospital Discharge Data Be Used for Medical Practice Assessment?

The observed data quality of diagnostic hospital discharge data is a function of the diagnostic process, diagnoses documenta-

tion, coding practice, characteristics of the disease, manifestation of the disease in patients, prevalence of the disease, and the way data quality is measured. Studies were highly heterogeneous with respect to these factors and showed highly variable data quality. The effect of several forms of bias remains unclear.

We conclude that quality of diagnostic hospital discharge data leaves much to be desired. In only 1% of the studies completeness and correctness are both at least 85%. Completeness of complications and comorbidities barely gets the 50%. Despite all the efforts to improve data quality, it did in general not lead to better data quality in the course of the years. The use of diagnostic hospital discharge data can easily lead to a biased idea of the quality of care. On the one hand cases selected for quality of care assessment possibly represent the more severe cases with a chance of overestimating sentinel outcomes, especially when comorbidities are also underreported. On the other hand, complication data are usually incomplete thus leading to an underestimation of sentinel outcomes. These shortcomings can differ between settings and thus complicates comparisons between hospitals or geographical areas. True variability in quality of care between settings can easily be overshadowed by the unknown variability in data quality which makes it very difficult to interpret the observed variability in the quality of care based on diagnostic hospital discharge data.

Despite the moderate data quality, we think that the use of these data by physicians to assess their own medical practice may be useful. Of all stakeholders, physicians have the best insight into the quality of their diagnostic data and its implications for the interpretation of performance indicators. Physicians are able to compare the information with their own experiences and can reason about what the information means for the quality of their care. However, this will only work when they are willing to critically reflect on their own medical practice.

Acknowledgment

We would like to thank Prof. Dr. A. H. Zwinderman for his support of the statistical analyses and Albert-Jan Schol for his preliminary data gathering and analyses.

References

16. Quan H, Parsons GA, Ghalia WA. Assessing accuracy of diagnosis-type indicators for flagging com-
31. Preen DB, Holman CD, Lawrence DM, Baynham NJ, Semmens JB. Hospital chart review provided more accurate comorbidity information than data from a general practitioner survey or an adminis-


120. Albarran JW, Clarke BA, Crawford J. 'It was not chest pain really, I can't explain it!' An exploratory study on the nature of symptoms experienced by women during their myocardial infarction. J Clin Nurs 2007; 16 (7): 1292–1301.


