Barriers and challenges of using medical coding systems
Surján, G.

Citation for published version (APA):
Surján, G. (2011). Barriers and challenges of using medical coding systems

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter II. QUESTIONS ON VALIDITY OF INTERNATIONAL CLASSIFICATION OF DISEASES-CODED DIAGNOSES

Abstract
International Classification of Diseases (ICD) codes are used for indexing medical diagnoses for various purposes and in various contexts. According to the literature and our personal experience, the validity of the coded information is unsatisfactory in general, however the 'correctness' is purpose and environment dependent. For detecting potential error sources, this paper gives a general framework of the coding process. The key elements of this framework are: (1) the formulation of the established diagnoses in medical language; (2) the induction from diagnoses to diseases; (3) indexing the diseases to ICD categories; (4) labelling of the coded entries (e.g. principal disease, complications, etc.). Each step is a potential source of errors. The most typical types of error are: (1) overlooking of diagnoses; (2) incorrect or skipped induction; (3) indexing errors; (4) violation of ICD rules and external regulations. The main reasons of the errors are the physician's errors in the primary documentation, the insufficient knowledge of the encoders (different steps of the coding process require different kinds of knowledge), the internal inconsistency of the ICD, and some psychological factors. Computer systems can facilitate the coding process, but attention has to be paid to the entire coding process, not only to the indexing phase.

1. Introduction
As is well known, different versions of the International Classification of Diseases: ICD (ICD-9, ICD-10, and ICD-9-CM) are widely used in many countries for several purposes, such as reimbursement, epidemiology, for grouping in case-mix systems and resource management. Large databases became available containing billions of medical cases indexed to ICD. The use of these data repositories is limited however by the questionable validity of the coded data. In this paper our aim first is to investigate-based on the available literature-how serious this problem is. Then I present an overview of the entire information process resulting in the coded data, and define a general framework for detecting potential error sources. By better understanding the nature of these errors we can hope to find ways to increase the accuracy of coded data.

2. Key features of International Classification of Diseases
Although ICD is a widely known coding system it seems to be reasonable to summarise some of its features which are important from the point of view of this study. ICD has different versions, which are updates (ICD-8, ICD-9, ICD-10) or modifications (ICD-9-CM). Whenever in this paper we are referring to ICD without specifying the version, we mean ICD-10, which is the most recent version,
published by WHO [1]. The key features presented in this section are more or less the same in the other versions.

The only original aim of ICD is to serve as a classification for standard, internationally comparable mortality and morbidity statistics. The ICD-10 titled International Statistical Classification of Diseases and Related Health Problems was published in 1995 in three volumes.

The first volume contains a tabular list of the categories in a three level hierarchy. On the first level ICD-10 contains 21 sections (e.g. ‘Infectious and transmittable diseases’). On the second level each section contains a list of the so-called 3-character categories, sets of diseases identified by a three character code (e.g. A36 = diphtheria). Most of the 3 digit categories are subdivided into 4 character categories, which represent the third level of the hierarchy (e.g. A36.6 = diphtheria of larynx).

ICD is not a nomenclature but a classification of the diseases. Consequently the four digit categories usually do not represent single diseases but sets of them, consisting of one or more diseases. The tabular list contains the code value, e.g. ‘I15.0’, and the description of the class, which in many cases is nothing more than a name (‘renovascular hypertension’ for I15.0). Explanations, inclusions and exclusions optionally might appear at the four character category, at the three character category header or at the beginning of the section to which the given item belongs.

ICD tends to be comprehensive. It practically never occurs, that a disease can not be assigned an ICD code. This is achieved by the ‘other’ and ‘not otherwise specified’ categories (e.g. H73.8 = ‘other diseases of eardrum’ for a specified disease of the eardrum which is not listed explicitly, and H73.9 = ‘disease of eardrum, not otherwise specified’ for a case where no more is known about the patient than that he or she had some eardrum disease). So the comprehensiveness is achieved by very generic categories. In case the disease can not be identified at all section 18 should be used, which contains categories of symptoms and signs. Conditions that are not diseases but might be reasons for providing some healthcare service are listed in section 21 (e.g. pregnancy and delivery, social problems, etc.). These two parts are not very detailed.

Neither of the official versions of ICD contain exact definitions for the meaning of categories, so the ‘semantic content’ of a category can only be ‘estimated’. This ‘estimation’ is based:

1. mostly on the name of the category, e.g. J67.9 hypersensitivity pneumonitis due to unspecified organic dust
2. on the officially listed inclusions and exclusions, like J67.9 hypersensitivity pneumonitis due to unspecified organic dust allergic alveolitis (extrinsic) NOS hypersensitivity pneumonitis NOS or K14.0 glossitis (inflammation of the tongue). Excludes: atrophic glossitis (K14.4)
3. on the place of the entry in the structure of the system. The description behind the code does not have to be complete:
D02.2 = bronchus and lung has the same description as D14.2 = bronchus and lung
but D02.2 belongs to D02 = carcinoma in situ of the respiratory system and middle ear and D14.2 belongs to D14 = benign neoplasm of the respiratory system and middle ear.

In ICD-10 there is another section in addition to the 21 sections, which contains concepts for the histological categorisation of neoplasms. These so-called M codes are identical to the corresponding part of the SNOMED morphology axis [2].

The second volume of ICD-10 contains rules and guidelines for encoders. The most important rules from our point of view are: (1) the rule of the highest possible specificity; (2) the rules describing the use of ‘daggers’ and ‘asterisks’; and (3) the rules which are pertinent to the selection of the ‘principal disease’.

The rule of highest possible specificity says that whenever there are more choices for a given disease, the most specific one has to be chosen. The ‘dagger’ (-†-) and the ‘asterisk’ (*), attached to the code values in the tabular list, are used when a disease is indexed by two different codes. In principle the ‘dagger’ codes represent the etiological aspects, whereas the asterisk codes refer to the anatomic localisation (e.g. in ‘renal tuberculosis’ the list states the following codes: A18.1† + N29.1* (A18.1† = ‘urogenital tuberculosis’ and N29.1* = ‘infectious disease of kidney and ureter classified elsewhere’). There are other cases, in which more codes have to be used for a single disease. E.g. in case of injuries and intoxications the external cause (the toxic agent or the circumstance of the injury) also has to be assigned a code (these concepts are listed in section 20 of ICD-10).

The main rule for the selection of the ‘principal disease’ states, that the principal disease is the condition that is the main reason for the medical service. Only one principal disease is allowed. In case two or more such independent diseases fulfil the criteria, that one should be chosen which is predominantly responsible for the costs of treatment. In case of ‘dagger’-‘asterisk’ pairs the appropriate ‘dagger’ category has to be chosen as principal disease for morbidity statistics.

Additional local or national regulations are possible mainly concerning the selection of principal diseases and for further labelling diseases, like ‘co-morbidity’, ‘complication’, ‘underlying disease’, etc.

The third volume contains an alphabetical list of the expressions that appear in the tabular list. This is just a mere index, without any explanation. So it is not advisable to use the alphabetic list alone. Once an item is found here, the tabular list should be checked to make sure that the selection was correct.

Some critical remarks have to be made here.

• The 3-level hierarchy of ICD is not represented clearly in the code values. While the 3-character categories are meaningful, the first and second characters are meaningless.

• A consistent categorical structure should be divided according to a unique semantic axis, where the ‘sibling’ categories have to be mutually exclusive [3]. The first level categorisation of ICD violates this rule. There are, for example, sections for infectious diseases, neoplasms, respiratory tract diseases and urogenital diseases. It is clear that both in the respiratory and the urogenital tract neoplasms and infectious diseases may occur as well, so these categorisations are not exclusive.

• The categorisation is confusing throughout the whole structure. For example the ‘dagger’ categories unnecessarily contain some reference to the
localisation (as in the above given example A18.1†, urogenital tuberculosis), while ‘asterisk’ codes contain some reference to the aetiology (N29.1* is only for infectious diseases of kidney).

- Since the categorisation is based mostly on epidemiological considerations clinical entities do not always fit the ICD categories. Clinically similar entities are sometimes located in very different places. On the other side, conditions of clinically different nature sometimes have to be classified into the same item. Beyond the ‘dagger and asterisk’ problems there are other situations where a clinical entity can not be assigned to a single category, because neither of the possible codes completely describes the case (e.g. the clinical entity ‘chronic cholesteatomous otitis media’ should be described at least by two codes (H66.2 = ‘chronic suppurative otitis media’ and H71 = ‘cholesteatoma of middle ear’).

- The principle of one and only one principal disease is clinically not sensible [4]. The rules for selecting the principal disease are not necessarily intelligible for clinicians. In case of polymorbidity, the costs not necessarily can be assigned to one or the other disease.

- There is a lack of consensus about and stability in medical terminology and categorisation of diseases and related concepts

The successive versions of ICD are published in periods of about 10 years. This 10 year revision period was decided at the turn of the century! Nowadays the development in medicine is much more rapid while the preparation phase of a new update takes a long time, so certain parts are out of date at the time of publication. No wonder, that experts disagree with certain parts of the categorisation, and experts of different schools and traditions disagree with each other about how to map their concepts to an out of date categorical system. Such disagreements might cause discrepancies in the third character level or sometimes even in the level of sections.

3. The problem of International Classification of Diseases coding

Several papers have been published on the accuracy of data coded according to the different versions of ICD throughout the world. There is a large variability in encoding methods described in these papers and also in the methods used for accuracy measurements. The majority of them reports low accuracy results. In some papers ‘minor’ and ‘major’ errors are distinguished. The error is ‘minor’ if there is a difference only in the fourth character of the code [5,6]. Others differentiate ‘serious’ (the codes are semantically totally different) and ‘moderate’ errors (there is some semantic relation between the assigned and the correct codes [4]). Typically the error rate exceeds 10% on the third character level, and is above 25-30% on the fourth character level, even in case of computer-assisted encoding [6-9]. The reported error rates only rarely were below 10%, as in [10]. The accuracy is domain dependent. Low error rate was found in ophthalmology, while extremely high in case of cardiovascular diseases [5].

When the data are used for DRGs, the resulting error in grouping varies between about 8 and 20% [11,12]. In papers, directly focusing on the coding problem, the reported coding error rates are usually higher than in those studies which are dealing with the data quality of discharge reports or charts in general [7,13]. This
means that the accuracy measurement is subjective, the result significantly depends on the applied method and on the interest of the researcher. The problem may be viewed from different angles. One may be interested in the effect of coding errors on the fairness of the reimbursement. This was the typical viewpoint of studies on ICD coded data accuracy in the eighties [5,12,14-16], [17]. One can also reverse the question and study the effect of reimbursement on the accuracy of the coding [18]. Later on the focus turned to the usability of the same information for epidemiological and even clinical research [19-24]. This distinction in viewpoints is important, because the error rate that is still tolerable depends on the purpose for which the data are used. In general, nobody is satisfied with the data accuracy. But those, who are focusing on reimbursement only, usually are convinced that the necessary accuracy level can be achieved without radical changes, but just by refining the methods already in use or by increasing the bureaucratic rigor. A need for better methods—perhaps better coding schemes—is emphasised by those, who want to use the same data for clinical or research purposes [19,20,27]. Many of the papers pay very little attention to the problem of the gold standard. The accuracy measurements in all publications are based on a reviewing process. In some cases the reviewers are qualified experts (e.g. registered record administrators or accredited record technicians [5,14]), however even these papers do not mention whether the original encoders had the same qualifications. If there is no coding process without error, then the reviewing process is no exception. In case of disagreement between the original and the reviewer’s code assignment—which is the right one? In addition even in case of agreement a systematic error can not be excluded. The agreement percentage between the original and reviewed data shows the reproducibility rather than accuracy of the information. In other words, the discrepancy is a measure of the intercoder variability [6]. The most convincing methods are those in which trained encoders make a ‘blind’ re-coding (without knowing the original codes) after which the discrepancies are discussed with the hospital. An example is found in [6]. The chance of subjective errors is reduced but not eliminated by these methods and the systematic errors still remain. The exact determination of the accuracy seems to be extremely difficult if not impossible. There are some papers, in which the coding errors are observed from a specific clinical point of view, like [19,20,25,27]. If the analysis is restricted to one or a few codes or diseases, the kappa statistic could be used to estimate the agreement. However this measure was used only in one of these papers [25].

The absence of a gold standard makes these accuracy measurements questionable. In many countries, like Hungary, there is no official qualification in ICD encoding. Certain persons might be thought to be knowledgeable in this field, but their expertise is not certified by any means. A very simple experiment—in which hundred clinical diagnoses extracted from real patient records were coded by two different physicians experienced in ICD coding—convinced us, that the coding accuracy in Hungary is similar to the ones reported in the literature. We can not expect that performing a more complicated accuracy test will provide new insights in the reasons for making errors. Instead we will focus on encoding as a process (section 4), and identify the possible sources and
types of errors (sections 5 and 6). In section 7 the possible ways towards improvement of the quality of the coding process will be briefly considered.

4. The process of encoding

In our view the codes emerge as a result of a quite long process, starting with the creation of the discharge report and ending with the submission of the coded chart or report to the appropriate office or organisation. The problem of data distortion during transmission from the hospital to the proper organisation also might be interesting [6], but is out of the scope of this study. Errors may emerge anywhere within the whole process. In details this process varies from country to country and from institution to institution. These variations come from the different uses of coded information, from organisational and logistic issues. The way of encoding described in this section is thought to be more or less characteristic for at least Hungarian hospitals. However we are convinced, that the essential logical steps of this process are more or less the same anywhere in the world and hence the reasons of errors may also be similar.

The patient record constitutes the input information for the coding process. The validity of the data stored in the patient record depends on the quality of the entire medical service and is not part of our study. It is likely to assume that the organisation of the patient record has impact on the ease of the creation and hence the quality of the discharge report, but the proof of this assumption is also out of our scope.

The main steps of the process leading to the coded information are presented in the following.

The first step is the creation of the discharge letter. The physician has to summarise the most relevant information about the patient. The physician is responsible for the discharge letter, which usually has a pre-defined structure. Every discharge report should contain at least one explicit diagnostic statement. Whenever a diagnostic statement is missing in the discharge report, the error rate of the coded reports increases significantly [24]. Here we have to emphasise the important difference between the (clinical) diagnosis and the disease. Since these and similar terms, like ‘illness’, ‘sickness’, etc. might have different interpretations, it is important to give some definitions, in the context of this paper.

Definition 1
Clinical diagnosis is a noun phrase expressing the essence of the physician’s considerations and knowledge about the patient.

Definition 2
Disease is a name of an abstract entity, based on some common features observed in different patients grouped together for some reason.

Disease and diagnosis differ from each other in the following ways:
1. The diagnosis indicates more or less precisely the future medical activity for the given patient, and can also be seen as justification of the procedures already done. The disease on the other hand does not necessarily indicate the reasonable or necessary healthcare processes for an individual patient. Correlation between diseases and procedures can be analysed statistically, however, for which the use of diagnoses would not be appropriate because of their specificity to individuals.
2. The diagnosis might refer to a symptom instead of the disease, if the disease is unknown. Example: 'hemoptoe'.
3. A diagnosis might refer to a symptom of a known disease if the presence of that symptom is important. Example: 'dyspnoea' mentioned together with 'laryngeal cancer'.
4. A diagnosis might refer to certain conditions and facts which are not diseases themselves. Example: 'status post appendectomy'.
5. A diagnosis should refer to the stage and severity of the disease in cases where the corresponding medical activity strongly depends on them. Example: 'incipient otitis', 'inoperable gastric cancer', 'severe pneumonia'.
6. A diagnosis might refer to its certainty ('suspect diagnoses'), or the means by which it was obtained ('radiological, histological, etc. diagnosis'). Examples: 'suspected tuberculosis', 'hystologically verified cancer'.
7. A diagnosis might refer to more details (e.g. laterality) whenever they are relevant for the medical management of the patient. In textbooks of medicine, like [26] left and right are discriminated only exceptionally (e.g. in case of cardiac failure), but in patient records this information is essential. Other example: 'fracture of the 3d digit' (this a different diagnosis but the same disease as that of 2nd to 5th, but fracture of thumb is different. This is because of the different structure and function of thumb).

This may be not an exhaustive list, but it demonstrates that 'diagnosis' represents a lower level of abstraction than 'disease'. Symptomatic diagnoses might be less specific than the underlying (unknown) disease(s), but can not be more abstract (the statement 'headache' is a result of an observation rather than of an abstraction process).

It might be confusing that sometimes the same phrase may refer to a diagnosis and a disease as well. 'hypertension' is a good example. [The latin term 'morbus hypertonicus' (hypertension disease) can be used to differentiate the hypertension disease from the diagnosis of hypertension.]

The number of diagnoses might not match the number of diseases for a given patient. Many-to-one, many-to-many, one-to-many, one-to-one and even one-to-zero relations may occur as well.

As an example, consider a patient having a disease called otosclerosis. Suppose both ears were operated, and now he has noise sensations and suffers from moderate deafness. This patient might have the following diagnoses:
1. Otosclerosis, both sides.
2. Status post stapedectomy, both sides.
3. Tinnitus.
4. Moderate combined hearing loss on both sides.

The first diagnosis refers to the pathological process (the underlying disease) itself; the second expresses that both ears of the patient were operated already (i.e. case history statement). The third and the fourth express signs of these conditions. Even these four diagnoses do not predict exactly what treatment is needed, but tell much more about the possible choices than a simple designation of otosclerosis as a disease.

All this is a borderline problem of medicine and medical informatics. This may be the reason why the literature pays not much attention to this question. Only a
general reference to this problem was found in the medical informatics literature [28]. The question is also mentioned from the special point of view of psychiatry [29].

Obviously no one but the physician should establish the clinical diagnoses: the necessary knowledge is purely medical.

The next step towards the coded report would be the recognition of the disease or diseases behind the established diagnoses. This step is nearly always missing since the majority of the physicians are not aware of the differences between disease and diagnoses. This is because they are nearly always working with diagnoses and not diseases in clinical practice. They meet the latter mainly in education and epidemiological research.

The carelessness about this step seems to be one of the main reasons for ambiguities in ICD coding. This step again requires medical knowledge, neither medical record personnel nor computers can take over the full responsibility. However physicians should be supported here, since medical knowledge alone might not be sufficient. Some knowledge on categorical structures is required as well.

One may ask, why diseases have to be coded and not the diagnoses of the patients. In epidemiological databases the interest is on diseases. This is because epidemiology is dealing with populations and diseases are related to populations, as we saw. Therefore the use of ICD is more or less obvious for epidemiology. The usefulness of ICD in financial systems it is not self-evident. For clinical research and quality assurance purposes the use of ICD or other disease classification system is usually not sufficient. The scope of our study is ICD coding, however.

The following step is the indexing of the established diseases into ICD categories.

Indexing is defined by Wingert [30] as follows:

"Indexing is... a process which translates a language utterance into a target language representation while preserving the meaning."

In our context the "language utterance" is the name of the disease, the "target language" is the set of the names of ICD categories. We remark here, that the ‘preservation of meaning’ is not absolute. Some part of the meaning is always lost by indexing, which makes this step irreversible.

This step serves statistical and financial purposes, and requires knowledge about the structure and rules of the ICD, the given statistical purposes and the financial system applied in the given situation. Relatively less knowledge is required about the nature of the different diseases. Physicians generally are not expert in this task.

After the indexing the next step is the transformation from the ICD category to the ICD code, which is relatively simple and does not require too much specific knowledge. This step easily can be done by computers alone.

In a narrower sense ‘coding’ is this transformation, while in medical informatics jargon the word coding is usually used for indexing and coding together. This difference is well known and described in the literature many years ago [30,31].

The last step is the labelling of the coded disease according to the specific regulation, especially in case of multiple diseases. This concerns the determination of the primary disease, and the labelling of the other disease types as co-morbidities, complications, etc. The knowledge required here is a mixture of medical, epidemiological and statistical expertise. The relevant regulations vary
amongst different countries or data collecting systems. If coding serves reimbursement, some knowledge about the financing system and the case mix system used in the given environment is also useful to achieve an optimal income for the hospital [32].

It is also possible to label the diseases before indexing, since the labels represent the role of the diseases in the given health care episode. The problem then is that the number of diseases not necessarily matches the number of the codes, as was shown in the section 2. So if a primary disease is represented by two or more ICD codes, only one of them should be labelled as primary. Perhaps the best way is that the physicians should qualify the diseases first when they establish them from the diagnoses, and then encoders should perform a re-labelling on the level of ICD codes. The ICD-10 rules support this method.

Summing up this section, Figure II-1 summarises the steps of the entire coding process. As the process moves away from the input information, the required medical knowledge becomes less and less. The labelling step is an exception, and this step is missing from the figure, because it might be done at different levels, as it was described above. The information will become increasingly formal and abstract after each step of the process. For this reason, the potential role of computers increases, because they can act on formal structures more effectively.

The process as described here is an abstraction of the real process: the logical steps are described, which may be performed more or less subconsciously in practice.

The next section deals with the possible types of errors and tries to explain them using the context of the above described model of the coding process.
5. Types of coding errors

5.1. Overlooked diagnoses

It might happen that one or more relevant clinical diagnoses do not appear in the discharge report. Since physicians are highly specialised nowadays, they focus on their own field, and tend to disregard some details not belonging to their speciality. This concerns of course their documentation activity and not their medical decisions. If one asks a physician what the diagnoses of a particular patient are, he likely will mention only the diagnoses he works on, and skip some other known co-morbidities. Chronic diseases like diabetes or hypertension are typical examples if the treatment was performed for other, more acute diseases, like appendicitis. The descriptive part of the discharge letter more likely will refer to such conditions than the explicit diagnostic statements. If the coded report is prepared by properly trained professionals, they may discover such overlooked diagnoses. This task in principle also can be performed by computers—at least partially.

No explicit reference to the existence of such errors was found in the literature. However it was shown that the 'clarity of the diagnosis'-i.e. the presence of a well formulated explicit diagnostic statement—has significant impact on the coding accuracy [9]. In a workshop on natural language understanding at the MEDINFO 98 conference Zwingenbaum emphasised that the task of coding is twofold: how to code, and what to code. So the entire patient record and not the explicit diagnostic statements has to be searched for codable items.
5.2 Errors of induction

As was shown in section 4, the diagnosis represents a lower level of abstraction than the disease, and the number of diagnoses does not necessarily match the number of diseases. This should not be confused with the fact that in the ICD some single diseases have to be described by two or more codes according to different aspects, as was shown in section 2. The latter problem belongs to the indexing phase and not to the abstraction phase.

Possible types of error in the induction phase are the followings:
1. The disease is confused with one of its more or less characteristic symptoms, e.g. ‘congenital malformation of the middle ear’ might be described as ‘conductive hearing loss’, or ‘vestibular neuritis’ might be described as ‘vertigo’.
2. A single disease is described by more diagnoses. (See the example of otosclerosis in section 4). If someone does not perform the step of induction, these diagnoses can be regarded as different diseases.
3. The disease is over-specified. The clinical diagnosis might be extremely specific. Let us take ‘atticitis’ as an example. This is inflammation of the atticus, a minute compartment of the middle ear. This condition is a form of otitis media. Such very detailed concepts are not intelligible for most physicians, only for specialists. If the induction phase is skipped an error may occur in the indexing phase, because such diagnoses are likely indexed to a generic class, such as ‘other disorders of ear’ or something similar.

According to the personal experience of the author this kind of error is responsible for a significant part of the coding errors. This impression can not be confirmed by the available literature, however, since the underlying conceptual difference of disease and diagnosis is scarcely dealt with in the literature.

5.3 Indexing errors

Indexing errors may occur separately when all and only relevant diseases are identified, but their mapping to ICD categories is wrong. These are described in subsection 5.3.1. Indexing errors are most commonly encountered in combination with induction errors, in which case it is difficult to say in what phase of the process the error really emerged. Such situations are presented in subsection 5.3.2. In other cases indexing errors are just due to disregarding the ICD rules. This situation is discussed in subsection 5.3.3.

5.3.1 Pure indexing errors

Let us say that B is the set of the assigned ICD codes for a given case, and C is the set of the correct ICD codes for that case. We assume here, that the set of the diseases for the case is given. An indexing error is present whenever $B \neq C$. When we are trying to categorize the errors, it does not matter how many elements of B and C agree (the number of elements of $C \cap B$ is irrelevant here). Two basic types of error are possible:
1. Type 1. Error of omission: B is a proper subgroup of C.
2. Type 2. There is a code $b \in B$ where $b \notin C$. Theoretically all other possibilities are combinations, however for practical consideration a third type should be
defined: 3. Type 3. There is a code $b \in B$ given instead of $c \in C$. This is a combination of type 1 and 2 error, but practically it means that a disease was indexed to a wrong category. These types are illustrated graphically in the Figure II-2.

![Figure II-2 Types of pure indexing errors](image)

Type 1 errors are mainly due to laziness or the lack of attention of the coder. This error is common in cases where more than one code should be used for one disease (see the examples in section 2).

Type 2 errors just occur accidentally, while type 3 errors are very common. This is the case in which a ‘wrong’ code appears instead of the ‘correct’ one. There may be many reasons for this. One of them is that the semantic content of the categories is not defined exactly, even in the printed versions. In the machine-readable versions the inclusions, exclusions and explanations are sometimes missing (this is the case with the publicly available versions in Hungary) which makes the semantic identification of certain items more difficult in computer assisted systems.

There is a special case of type 1 error, where indexing is incomplete because a disease that should be represented by two or more categories is assigned only to one.

As was mentioned in section 3, some papers categories the errors according to their severity. However, the ‘severity’ of an indexing error is sensible only in case of type 3, where the ‘semantic distance’ of the correct and the assigned code makes sense.

### 5.3.2. Indexing errors in combination with induction errors

As mentioned in section 5.2, the over-specification of a diagnoses may lead to indexing errors, if the induction step is missing. The relations among diagnoses, diseases, and ICD categories sometimes are rather complicated.
Figure II-3 Discrepancy in number of diagnoses, diseases and ICD codes

Figure II-3 illustrates a situation in which several diagnoses belong to a single disease, which has to be represented by several ICD codes. It is essential to emphasise, that ICD1 or ICD2 not necessarily correspond directly to any of the given diagnoses, but they represent different aspects of disease 1.

Take as an example a patient presenting himself for shortness of breath due to a recurrent laryngeal cancer with metastases in the regional lymph nodes. Suppose that he previously received radiotherapy. The clinical diagnoses together with the TNM classification in this case might be:

- **Dg1.** Recurrent laryngeal cancer with cervical metastases left side T3N2M0.
- **Dg2.** Dyspnoea.
- **Dg3.** Status post radiotherapy.

The induction step leads to just one disease, i.e. laryngeal cancer. However this can not be indexed precisely into ICD by a single code, since the most specific entry C32.9 (malignant tumour of the larynx) does not represent sufficiently the extension of the process, so C77.0 (secondary malignant tumour of lymph nodes in head and neck) has to be added. According to this reasoning three clinical diagnoses and two ICD codes belong to a single disease. The two ICD codes correspond to the diagnosis 1, while diagnosis 2 and 3 should not be coded into ICD, because they are not diseases but just facts about the patient. The result of this reasoning is presented in Figure II-4.

Figure II-4 Representation of one diagnosis by two ICD categories (Dg2 and Dg3 disregarded)

If the induction step is missed there is a temptation to seek one ICD code for each diagnosis. In this case the codes would be different, as is shown in Figure II-5.

This latter way of representation is incomplete, since the entity of the cervical metastases is missing-and inaccurate, because R06.0 is contradicting the rule of most specific code (see section 0), and 292.3 should only be used for cases in which the history of irradiation is the cause of the medical service.
On the other hand one might feel that the representation according to Figure II-3 is also incomplete, because it does not refer to two important facts about the patient (dyspnoea and history of irradiation), and C77.0 should be used only if the primary tumour is unknown.

According to a fourth opinion nothing else but C32.9 should be used, since this is the only disease of the patient, the other diagnoses are consequences and details. In addition, one might argue that the histological diagnosis also is required, so the appropriate M-code should be added (e.g. M8071/3 = squamous cell carcinoma, nonkeratizing).

All these opinions are defensible. The main problem in deciding which coding is correct, is the lack of exact rules and definitions.

These different opinions agree only in one aspect, namely, that the principal disease is laryngeal cancer. However, if the main reason of hospitalisation was asphyxia, which required an urgent tracheotomy, the representation could be different again. The problem of the selection of principal disease will be discussed later in section 5.4. The lesson to be learned here is that there is no correct use of ICD in general. The ‘correct’ coding is purpose (for example getting overviews of reasons for interventions or getting an complete overview of diseases) and environment dependent (hospital, GP, country). Diseases are artificial abstractions, which are not observable directly in the real world [31]. In medicine there is not always agreement on what are independent disease entities and what are different forms or manifestations of the same disease. In our example one might regard the lymph node metastases as a disease entity just because it may be the reason for an operation or it may be observable with or without known primary neoplasm.

Furthermore ICD classifies diseases according to different attributes, and the relevancy of these disease attributes is purpose-dependent. In the previous example the ambiguity was a consequence of the lack of definition of the purpose of coding.

\[
\begin{align*}
\text{Dg 1} & \quad \text{C32.9 Malignant tumour of the larynx} \\
\text{Dg 2} & \quad \text{R06.0 Dispnoe} \\
\text{Dg 3} & \quad \text{Z92.3 Personal history of irradiation}
\end{align*}
\]

Fig. II-5. One by one representation of diagnoses

5.3.3. Conflict with ICD rules

As was mentioned in section 2, ICD makes use of several coding rules. When one is not aware of these rules one can make indexing errors, e.g. a ‘dagger’ code is recorded without the corresponding ‘asterisk’ code, or a generic category is used when there is a more specific category. Violation of rules pertaining to the selection of the primary disease leads to labelling and not indexing errors, so this problem is discussed in the next section. We do not deal with the errors of transformation from ICD category to the ICD code value in detail. This is just a
technical problem, like typing errors etc. It can not be fully eliminated, but it is responsible only for a minor proportion of the errors.

5.4. Errors in labelling

In addition to ICD rules many countries have some special regulations for coding [4,25,33]. These rules pertain mostly to the labelling of certain diagnoses, and rarely regulate the indexing step. Usually the regulations allow one and only one principal disease-in accordance with the original ICD philosophy. The improper selection of principal disease frequently occurs. As Corn found, the 35% total error rate would decrease to 17% if this kind of error could be avoided [7]. More than half of the errors was due to mislabelling. Doremus et al. found that the improper selection of a principal disease occurred only in about 22% of the total number of errors [14]. Local regulations can strongly affect the frequency of this kind of error. The restriction which allows only one condition to be specified as principal may be reasonable from an epidemiological or financial point of view, but is not sensible from a clinical point of view (see the critical remarks in section 2). Moreover, one condition sometimes has to be expressed with more than one code (see Figure II-3 and Figure II-4 and the corresponding example). This problem was mentioned already in section 4. If the regulations only permit to assign one code for the principal condition, additional rules are required to avoid further ambiguity. ICD regulates this question only partially. Since there are many reasons to have more than one ICD code for one disease, these rules might be too complicated to follow them perfectly.

6. Potential sources of errors

Considering the process of encoding as described in section 4 the first source of coding errors is the physician. We do not mean here misdiagnosis in a medical sense but errors in the formulation of explicit diagnoses, i.e. merely documentation mistakes. In principle physicians have some interest in the quality of medical documentation, since they are one of the main users of it. There are many other uses of medical documentation in which physicians do not have direct interest. Moreover, documentation is a tool by which the work of physicians can be controlled, what they perhaps dislike. Many doctors are convinced, that documentation is a just a burden [24]. Therefore the quality of primary medical documentation is not always satisfactory.

The second source is the person doing the indexing, who either is a physician-like in [21-23,28]-or some specially trained health professional [24,34], with or without supervision of physicians. In Hungary the physicians should take the responsibility, it frequently happens however that the indexing is done routinely by uneducated clerks, who can not do more than select one item from a prepared short list of commonly used diagnoses. The personal errors highly depend on organisational and even on logistic issues. Mechanni et al. for example reports an extremely low data capture in case of patients who died in hospital. This was because coders facing a huge backlog in their work thought that charts of living patients must have strong priority [9].
Chapter II – Questions on validity of ICD coding

The third important source of errors is the ICD itself. Using it for indexing clinical diagnoses is not the original objective of ICD. Its logic and structure differ from the clinical way of thinking. Even the terminology is strongly different from the clinical jargon (especially in the case of the official Hungarian translation of ICD-9 and ICD-10, [35]). The language used in coding systems is usually different from the clinical jargon. This is the case not only with ICD but some other coding systems, for instance HICDA-2 [36].

However the most serious problem is the internal inconsistency of the structure of ICD. In section 2 we criticised the consistency of ICD as a categorical structure. Beyond objective consistency problems, it is important to also emphasise the psychological effect, when practising physicians are faced with a categorisation that does not make sense to them. They became convinced that this system is totally vague, hence there is no reason to make too much effort in finding the perfect codes: the result will be unsatisfactory any way.

It is often suggested that a fifth (sixth etc,) character subdivision of ICD categories would refine the system and move it closer to the world of physicians. However, there is no evidence of this. Depending on the approach it may even happen that the error rate will increase. Errors that already occurred on the third or fourth character level would be not rectified, but additional errors at the fifth level may occur even in those codes, which were correct when using the system of four characters. There is not too much chance to create a consistent subdivision in an already inconsistent system.

The internal inconsistency of ICD is emphasised by many authors. Chancellor et al. found consistency problems between the tabular and alphabetic list of the English edition of ICD-9 in case of bulbar and pseudobulbar paralysis [24]. The alphabetic list of the Hungarian edition of ICD-10 [35] only contains the expressions appearing in the tabular list without synonyms. Search for expressions used in the clinical jargon therefore usually fails. This gives the false impression that the required concept is missing from ICD, and that some ‘similar’ concept has to be found instead. Encoders who are not physicians prefer to use the alphabetic list instead of the tabular one.

Figure II-6 displays the main potential sources of errors on top of our framework describing the entire coding process. The question emerges: what can we do to ameliorate the situation? In Hungary there are two ‘official’ opinions. The first suggests more bureaucratic rigour [13]. The second suggests a multi-purpose usage of the data (until now they are used nearly exclusively for reimbursement). Fully automatic encoding seems not feasible in the near future, so it is not suggested. Both opinions originate from the belief that all the errors are caused mainly by simple human mistakes or by intentional manipulations of the data, and there are no theoretical obstacles for an entirely correct coding. Neither the international literature nor the personal experience of the author supports this belief. What the literature agrees about is the distorting effect of reimbursement on the codes [12,18], since hospitals strive to achieve higher incomes. Humans are obviously error prone, and it is a likely assumption, that the level of human errors depends on scrupulosity of the coders and the administrative rigour. No estimation was found in the literature what is the achievable maximum improvement by these two factors.
The author is convinced that bureaucratic rigour would result in never ending disputes. Without objective definitions of the ICD categories everything is disputable. It is a likely assumption that the multi-purpose utilisation of the data may reduce the distorting effect of reimbursement. Manipulations good for achieving a higher income for the hospital might be disadvantageous in quality assurance for instance. But this is true for intentional manipulations only and would not solve all the problems related to accurate coding.

In the next section we are considering some possible ways toward better quality ICD codes.

**Figure II-6**

Main causes of coding errors within the framework of the coding process

### 7. Possible ways towards improvement of International Classification Diseases coding

Computers offer a series of methods to support the difficult work of encoding. However, automatic indexing is still a matter of research rather than everyday reality. There is a temptation to simplify the coding problem in practice. There are some systems for instance used for writing discharge summaries in certain Hungarian hospitals offering a list of diagnoses together with a one-to-one reference to the corresponding ICD code. This means that for instance in the example of section 5.3 the coding can not be done anyway else than according to Figure II-5. The method has some further drawbacks. The maintenance of the
internal consistency of such thesauri is a difficult task, which requires a well-designed methodology. Secondly, in such an environment the physicians will not express what they want to say, but only what is found in the list. Of course physicians could add new terms to the list, but then they also have to find the correct ICD code, what they usually dislike. According to the ‘law of energy minimum’ (roughly saying laziness) people neglect the updating process and simply use those terms that are already in the system.

Another possibility is to produce a hierarchic browser, which does not allow the physicians to express clinical diagnoses directly. Instead they are guided by the browser through the hierarchy entirely until they arrive at a leaf of the concept-tree, fitting sufficiently with the diagnoses arising in their mind [37]. This is a more advanced solution, but the drawbacks are similar. In both cases, the expressive power of human language is sacrificed.

To preserve this expressiveness and flexibility the physicians have to be allowed to use their own language and computers should be able to represent their utterances in any required form, like coded entries. To achieve that, one way is the ‘natural language understanding’ approach. For this we should have a canonical representation capable to represent everything what can be uttered, and a natural language processor, which translates the language input to this canonical representation. I would call this a structural approach, since it is postulated here, that the language information can be expressed in formal structures. This is the aim of many projects throughout the world, like MENELAS [38] and GALEN [39,40]. The canonical representation is language independent, so it can be used anywhere, but the processing of the natural language strongly relies on the character of the different languages.

For Wingert (in his early work [30]) and for many of his followers computer assisted coding is usually restricted to the problem of indexing alone. The question how to find the concepts to be indexed is disregarded. Recent experiments try to use discharge summaries as input information, and extract the concepts to be coded by natural language processing [38,41]. This can be helpful in avoiding errors of overlooked diagnoses.

All methods based on natural language understanding are at least partially language dependent. It might be difficult to find resources for adaptation of such methods in small countries speaking a unique language.

There is another way, which also allows physicians to use their language without restrictions, and maps the utterances into a computationally tractable space. This is what we call the statistical approach [42]. The underlying postulate here is that the appearance of words in a given text or expression is not a random event, but a semantically driven process. If this postulate is true the statistical analysis of the word occurrences might be used for mapping language information into a latent semantic space. This space is latent, because the language information does not have to be understood by the system, but it is semantic, because the ‘co-ordinates’ of a text in this space correlates with the semantic content of the text.

The latent semantic approach is already used for a series of indexing tasks [43-45]. By its statistical nature the performance depends on the size of the samples, i.e. the length of the texts to be indexed. Medical discharge reports seem to be suitable but diagnostic statements might be not long enough (sometimes it can be a single word
only). Therefore the approach should be adapted for such short pieces of information. This statistical approach also seems to be useful for capturing implicit diagnostic information from medical reports, e.g. the term ‘insulin’ appearing in the discharge summary is likely to indicate diabetes, even when this diagnosis is missing from the explicit list of diagnoses. The adaptation of latent semantic methods is easier than natural language understanding methods.

8. Discussion

The framework of the coding process presented here is only one of the possible models. It cannot be generalised to every medical coding system. The process of coding of surgical procedures for example might proceed differently. However, encoding always is a compound process consisting of several successive steps. Consequently any effort to improve the quality of coding should take into account the entire process and should not consider it as a single operation on the input information.

I wanted to show that ‘correctness’ of an ICD code is not absolute. Correct coding depends on the environment and also for a large part on the purpose of the coded data. Consequently computer systems assisting the coding process have to be designed in a flexible way to be able to satisfy different uses. This means that not only the knowledge about diagnoses, diseases and the categorisation should be represented in such systems, but also some knowledge about the utilisation of the coded information.

Both the structural and the statistical approach mentioned in this paper have their benefits and drawbacks. Their relation seems to be complementary and not exclusive. Still now it is not clear, how they will be able to cooperate in the future.

Since the patient record is the source of input information of the coding process, the spread and development of electronic patient record systems will open new perspectives in both directions.

Acknowledgements

The author wants to express his thankful ness to Professor Arie Hasman for his kind help and critical remarks to the manuscript of this paper.

References

Chapter II – Questions on validity of ICD coding

[37] D. Deimel, H.J. Hesselschwerdt, J. Heisel, Einfache und valide ICD9/-10- und IKPMKodierung mit Hilfe des EDV- gestutzten Verschlusselungssystems ‘do
Chapter II – Questions on validity of ICD coding

it’-Erfahrungen nach einjähriger Anwendung. (Simple and vaud ICD9-/10- and IKPM coding using an electronic data-assisted coding system ‘do it’-experiences after 1 years use), Z Orthop Ihre Grenzgeb 135 (6) (1997) 528-534 (in German).


