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

Citation for published version (APA):

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 I. INTRODUCTION

1. Overview
Sciences strive to represent their domain knowledge in an unambiguous way, often relying on logical and numerical representations. Ontological knowledge is a particular form of knowledge specifying conceptualization of the domain, that is, it defines the basic concepts and relationships underlying the domain. In a sense, this knowledge is meant to function as a semantic contract between stakeholders to allow sharing knowledge of the domain. The basis of such knowledge is often formed by an enumerable set of categories (also often called classes or types). A category, for example 'person', may be associated with a code, for example a natural number, and may have various terms, for example the string “homo sapiens”. Mapping a term or a phrase to the code of its category is called the coding problem. Ontological knowledge can be represented as a simple list of codes or terms of categories, a classification showing which categories are subclasses of others, or as complex as logical statements in description logic or first order logic stating constraints on the relationships between categories. Depending on the completeness and degree of formalization of the knowledge one may use it in inference for example to conclude that a category is subsumed by another. Efforts to represent conceptualizations are old, for example the search for enumerable classes in biology and medicine and arranging them in a classification started far before the introduction of computers.

The subject of my study is medical classification and coding. Obviously a much larger field than a single thesis can cover. Therefore I restricted the scope to the problem of the actual use of classifications in clinical practice especially in the context of the coding problem. Worldwide, a tremendous amount of health related information is collected such as health care episodes, morbidity surveys, and mortality data. There are essentially two uses of health data: one is the clinical use within the primary healthcare process. The other use, which is the focus of my thesis, is the central collection of data aggregated from clinical records. These aggregate data can be used for research and for policy decisions. A great portion of these data is coded.

There are three main reasons why these data are classified in discrete categories and coded.

1) Statistical Having a limited set of categories and hence sufficient number of cases in each category is essential for sufficient inference power in statistical analysis. Also, such limited number of classes is easier to manage and present when a code is assigned to a category instead of using the names or descriptions of the classes.

2) Language independency For international comparison of data it is vital that the data can be expressed in a language-independent way. It means, however, that the coding schemes have to be available in many languages, but once this is done, one can translate codes into their own language and vice versa.
3) **Computational feasibility and tractability** Information expressed in natural language is still extremely hard to process by computers. Computers hence cannot assign meaning to phrases and in this sense coding schemes 'translate' the data into codes that indicate the meaning of the underlying concepts.

The – usually national – data repositories are very large (e.g. in Hungary all hospital cases are documented in this way since 1993) and this makes them very valuable for research and decision-making, but there are serious concerns about the usability of these data. Most often it is claimed that the validity of these data is poor. Validity, the problem of whether the represented data reflects reality, is hence an important theme that I address in this thesis.

2. **Preliminaries**

This section briefly sketches the life cycle of coded medical data. This life cycle may vary from country to country, so we use the Hungarian state of affairs as an example. We start from a patient-physician encounter (in- or outpatient healthcare episode) resulting in some clinical documentation (in the medical record) that usually contains one or more diagnostic phrases.

The coding procedure takes as input some free text extract of the medical record, such as a discharge letter. The whole record is rarely used as input, as is also the case in our experiments, where an explicit phrase (often reflecting a diagnosis) has already been extracted from the records to reflect the respective episode of care. The coding procedure ends up with a set of codes assigned to the given phrases. Subsequently, all cases of a health care service provider are reported to some central office or agency (e.g. insurance agency or statistical office) where the data are aggregated forming various statistics, that are stored in (regional and national) health databases, which usually also contain other statistical data. Institutions responsible for the regional or national data collection often report their data to some international organisations in particular the WHO (World Health Organisation), OECD (Organisation for Economic Co-operation and Development), and EUROSTAT (The statistical office of the European Union). In this way large international public health databases are created.

Coded data may be used at various aggregation levels, from the local (often at the level of the hospital) up to the international level. Aggregated data are studied and evaluated among others in order to recognise hidden correlations or phenomena that require some action to improve the health of citizens. A large amount of effort and money is spent on this whole procedure worldwide. The underlying assumption is that this investment should eventually benefit the patients. However, the use of the data is hampered by a number of factors, which form the subject of my study.

3. **Scope definition and limitation**

This thesis concentrates on disease coding, more specifically on the use of the worldwide used International Classification of Diseases (ICD). Chapter III and VII put this in a wider context. In Chapter III disease classifications are discussed in a historic and cultural context e.g. in relation to biological classification of species. In
Chapter VII I study public health databases, where disease incidence and prevalence data (coming from disease coding) are integrated with other statistical health data (procedures, health care resources, risk factors, demography etc.) The thesis does not deal with the fine-grained classifications (often called nomenclatures) that are designed and used to support the clinical primary care process of individual patients. SNOMED (Systematized Nomenclature of Medicine) is probably the best known example for medical nomenclatures.

Being interested in practical challenges I do not discuss in detail the theoretical feasibility and limitations of medical knowledge formalisation, and the philosophical debate pertaining to concepts and categories in terms of being products of human thinking or entities that exists in reality independently of any human knowledge. In addition, this thesis can not give an account of all existing computer assisted coding methods. Not only because they are quite numerous but also because many of them are too costly and difficult to implement or adapt to various languages (most of them were developed for the English language).

4. Research questions

In this thesis, I address the following research questions.

Question 1
How serious is the validity problem, and what are the main causes of coding errors?

While all human activity is error-prone, it is also claimed quite often that in many countries, where coded reports of hospital or outpatient cases are used for reimbursement, the financial incentives distort the data in order to maximise income of health care institutions. But most often in various professional communities this issue is only discussed in a general way, without providing information about the percentage of coding errors, the reliability of statistical data, or the sources and causes of coding errors.

Question 2
What are the cultural determinants of current biomedical classifications and how do these determinants challenge the use and reuse of coded data?

Apart from the problem of validity it is clear that no classification is able to represent the whole reality. Classifications are products of human intelligence and by necessity, every classification is an abstraction of reality: some features are emphasised, while others are neglected. And as such they are influenced by a number of cultural factors that developed during centuries. Systematic data collection started centuries ago, and there is a clear need for data that are comparable over time. This is an important reason why analysts are reluctant to change the structure of their coding schemes. In that sense coding schemes are conservative. But even a brand-new classification would be determined by the actual knowledge, culture and development of contemporary science. So all current classifications are the result of a continuous cultural process, and many of their features stem from history. How these cultural factors evolve in time and how they affect the usability of classifications are problems that are rarely investigated.
Chapter I - Introduction

**Question 3**
**How to compare and evaluate the performance of various statistical computer-assisted coding methods? What are the factors that influence their performance?**

The current internationally used standard classifications were typically developed for manual use. Even today fully automated coding does not seem to be feasible, and in fact, most existing data collections are based on human coding. However, this human work can effectively be assisted and improved by computers. The implementation of such systems in practice requires careful evaluation. A large variety of computer assisted coding methods have been developed during the past decades. Some of them are quite sophisticated and hard to implement, others are simple and easy to implement. Up till now, there is no standard and generally accepted evaluation method that could help to choose the best or at least a satisfactory method for a given task in a given environment. In case of the widely used corpus based approaches the performance of the coding algorithm might depend not only on the method but also on the quality of the used corpus especially in terms of the cardinality relationship between phrases and codes (that is, how many phrases relate to one code and vice versa). This aspect is rarely addressed in the literature. A further problem of the evaluation is that there is no absolute gold standard: based on the noisy data alone, there is no way to decide objectively if a code that was chosen in a given case was the correct one or not.

**Question 4**
**To which extent a corpus based algorithm is able to recognise compound diagnostic expressions and decompose them into single diagnostic entities?**

In many cases the mentioned evaluation in Question 3 above is carried out on a one by one basis: one code is assigned to a diagnostic expression and this assigned code is compared to some gold (or silver) standard. But often a diagnostic expression comprises more than one diagnostic entity. A system that assigns one code to one diagnostic expression necessarily will fail to represent one or more diseases in case of compound expressions. If we are able to create a corpus that consists only of "elementary" diagnostic entities, it is reasonable to use this corpus to recognise compound expressions and then to decompose them into single ones.

**Question 5**
**To which extent is it possible to transform traditional coding schemes, like ICD, into a formal representation?**

By definition, the performance of corpus based methods depends on the quality of the used corpus. This quality can be improved in various ways, but since all corpus based methods start from some collection of manually coded diagnoses, the possibility of human error is always present. Another way of coding is the formal approach where the principle is to create a formal representation of the ICD categories, and use some Natural Language Processing (NLP) tool to translate diagnostic phrases or medical records into some formal terminology after which an automated reasoner can classify the formally represented diagnoses to ICD classes.
However traditional coding schemes like ICD might contain a number of arbitrary classes that are not formally defined, and hence they perhaps can not be precisely represented in a formal language. The formal approach has other serious drawbacks: it requires a comprehensive formal medical terminology for representing diagnoses or medical records and the NLP tools are language dependent. These are unresolved problems, yet the possibility of formal representation of coding schemes is a relevant issue: it could help develop terminology services (e.g. to collect all ICD codes that are related to renal diseases – these codes are scattered in various chapters of ICD and hard to collect manually) and also could help to compare and map different coding schemes.

Question 6
To which extent is it possible to improve comparativeness of public health indicators by an ontological framework?

The ultimate reason of the coding procedure is that eventually we get a set of aggregated clinical data or indicators that describe the morbidity of a population or the activity of the health care system to help us improve it. Many national databases are created with this goal in mind, and several international organisations collect national data to incorporate them into various international health statistics databases. But analysis of these data across various countries is a challenging enterprise; since the categories used to describe the health status or the health care system differ from country to country. Some categories are based on administrative, legislative definitions that are also subject to change over time. It seems to be a likely assumption, that if we are able to represent the underlying categories and their definitions in a formal system, this can help to compare data from different sources or data that span a longer period of time. As a minimal result, we should be able to help the user to decide whether the requested data are comparable or not.

5. Organisation of thesis
The above described research questions form the structure of this thesis.
Chapter II gives an overview of the problem of the validity of ICD coding of diagnostic expressions. The chapter includes a literature review and then provides a logical framework of the coding process. It is important to emphasize that this is not a real workflow description: coders rarely follow the described logical steps. But this framework appears to be useful to understand how and why so many coding errors are made.
The third chapter describes the cultural-historic context of classifications, tries to explain that classification systems have some historical continuity and that many of their structural features – that also often influence the data quality – are inherited from the past, and it is hard to ignore or get rid of them entirely.
The history of classifications in a wide context draws our attention to some inherent limitations of all formal approaches. The coding procedure itself starts from an informal free text description and ends with some formal (coded) representation. When this process is assisted by computers, basically two different
paradigms seem to be possible. One is called the statistical approach that is based on a corpus of manually coded data. The approach uses some "learning" algorithm to guess the best codes for the free text input. The other paradigm is the formal approach that starts with some natural language processing to translate the free text input into formal structures and then uses a classification inference algorithm to map the input information to classes of a predefined system of categories. Chapters IV and V focus on statistical methods. Chapter IV deals with performance evaluation problems and chapter V describes an algorithm that identifies single diagnostic entities in compound diagnostic expressions.

Chapters VI and VII turn to formal methods. It is a pre-requisite of all formal approaches that the classes of the target classification should be formally described. But a formal representation of e.g. ICD is also important if we do not want to use formal methods for computer assisted coding. For example, if one wishes to retrieve all cases of lung diseases from a database, the first step is to collect all ICD codes that represent lung diseases. Such codes are scattered all over the various chapters of ICD (infections, tumours, respiratory tract diseases, injuries etc.) Their manual collection is repetitive and error prone. Once we have a formal representation of ICD then a terminology server application can facilitate doing this work and with the potential of a drastic reduction in the number of errors. This is the theme of Chapter VI.

Chapter VII provides a formal ontological framework for public health indicators that could improve comparability of various national or international public health data. Chapter VIII discusses and concludes the thesis.