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.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Chapter VIII. DISCUSSION AND CONCLUSIONS

My aim in this work was to better understand the generation and utilisation of coded medical data, and find ways for computers to improve both. Generation of codes (using an existing coding scheme) is error prone, and use of computers can reduce, although not totally eliminate, the errors. The coding problem, i.e. the assignment of a code to an expression, is much influenced by the coding schemes. These schemes are based on a cultural and historic process. It is for historic reasons that even the International Classification of Diseases (ICD), the most widely used coding scheme, itself is often inconsistent, a fact that contributes to coding errors. One should, however, recognize that a certain amount of errors does not necessarily lead to wrong inferences.

Morbidity and mortality data gain additional value when they are integrated into public health databases that contain other health related information. However, the utilization of these data has limitations. Comparison of data originating from various sources can be hindered because of the use of different concepts and definitions especially in case of large international databases. For this reason I tried to set up an ontological model that can represent these data – often called public health indicators – in a formal way.

In this final chapter I first address each research question posed in the Introduction, briefly summarise the conclusions from each chapter and then discuss them in the context of the whole work. Finally, I delineate future work.

1. Answers to the research questions

For each question I provide a short context, the key findings and the answer to the question.

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

Manual ICD coding has a minimal error rate that cannot be avoided. This error rate is approximately 10% for three character codes and 30% for the fully detailed four character codes. Very often the distorting effect of the health care reimbursement incentive is mentioned as the most important reason for validity problems. But the reported error values vary among the various publications, and errors may be committed also in countries where the reimbursement does not depend on disease coding.

In this thesis we have set up a logical framework of the coding process as a multi-step abstraction from the patient record to the ICD codes. All coding errors are related to one or more steps in this logical process. In our view the coding process is affected by a number of factors including clinical documentation, human factors on the part of encoders, the consistency and
Chapter VIII – Discussion and conclusions

usability (among others clarity of category definitions) of the coding scheme as well as the distorting effects of the reimbursement system or other incentives.

The error rate of disease coding is obviously too high for the medical coded data to be used in the clinical context, but ICD was designed with aggregate analysis in mind. Since most national and international statistics are reported with the three character system, the error in estimating the frequency of a disease group in the population will roughly amount to 10 percent. This means that these data – with some precaution – can be used in decision making at the health policy level. One should, however, still strive to reduce these errors.

Some of the error sources can be reduced by using computer systems. According to the literature, even very simple tools may improve data quality. On the other hand, some systems that oversimplify the coding process may result in unsatisfactory quality of data. Designing a more consistent classification scheme obviously could improve data quality, but at the expense of incompatibility with all previously collected data. Mapping from one coding scheme to another cannot solve this problem: if we create a new scheme so that the old scheme can be mapped onto the new one without significant distortion, then we have to preserve the essence of the old inconsistent structure. If we create an essentially new consistent structure, then we cannot map our data without distortion.

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

Classification – the way of arranging things in various groups - forms the basis of nearly all scientific activity. In biomedical sciences the huge variety of phenomena results in such a large number of classes that it requires systematisation. This means that categories are arranged into hierarchies with some 'top level' categories. These 'top level ontologies' (an ontology specifies the classes as well as relations among classes in the domain) come from philosophy and hence reflect the general knowledge of mankind about reality and obviously have cultural determinants that change over time. But not everything in the world can be classified into hierarchies.

The modern view of diseases as entities that – at least in principle – can be defined and classified is a result of a long development. Classifications and coding systems are evolving. In this thesis we discussed how the following factors affect the way things are classified at a given time:

- Development of the domain knowledge
- Development of the philosophy and knowledge about rules of building categorial structures (e.g. using exclusive and inclusive hierarchies)
- Development of the technology used for the representation of categories. An evident example is the difference of paper-based
and digitally represented classifications. Modern classifications are based on formal languages that have enjoyed much attention in the last decades.

The world cannot be represented by an enormous single robust ontology. All classifications have some domain restrictions and can be considered as modules that preferably should be compatible and coherent with each other. To ensure this, we propose to start from a top level ontology that describes the basic structure of reality, and then create a number of reference ontologies that comprise universally used categories of medicine (e.g. anatomy), while specific classes used for a given task are to be represented in application ontologies.

The above mentioned determinants are changing over time rendering the maintenance of disease classification an endless process. This makes the reuse of data collected in the past challenging. The problem grows with the elapsed time. E.g. data conversion from ICD9 to ICD 10 is possible with some more or less tolerable distortion. But most of the categories used in the London Bills of mortality are rather obscure today. One can expect that the currently used ICD categories will also become obscure for the people of the 24th century.

**Question 3**

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

A large number of different computer assisted coding tools has been developed and described in the literature. Due to the lack of a gold standard for ICD coding, performance evaluations of computer assisted coding methods are based on some manually coded set of diagnoses or documents. However, sufficiently large and reliable corpora are hard to collect. The codes suggested by the software are compared with codes that were assigned by humans to the same diagnoses. In other words we use a noisy learning sample, and use another part of the noisy corpus for evaluation. So in essence one measures how smart computers are in reproducing human errors. For corpus based methods it is evident that the performance measures depend on some features of the learning set. Several publications compared different methods among each others and stated that some are superior to others.

In this thesis we showed that comparing the performance of two corpus-based methods on two different corpora demonstrated that superiority of one method over another is not absolute: one method performed slightly better in one corpus but drastically worse in another, depending on the corpus characteristics. On one corpus we could additionally demonstrate that performance differences between ICD chapters where much larger than differences between the two methods.

Comparison of the methods and selecting the best available solution have, hence, to be performed in the given environment, where we plan to implement a computer assisted coding system. The quality of the used learning set should be always considered. In particular cardinality and consistency of the expression-code...
relationship should be considered to understand coding performance. Interestingly, inconsistent learning samples might still help the human coder choose between close alternatives, thereby improving the quality of coding. Our experience with the inconsistent sample, where the same diagnostic terms appear with different codes is that the right code used to be more prevalent than the erroneous ones. For this reason the Naïve Bayes method might perform quite well with such samples as it could exploit the frequency of the correct code to find it. The vector-space method for coding could not exploit the majority code and was hence distracted by the other “close” codes.

**Question 4**

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

Physicians sometimes comprise two or more clinical condition into one diagnostic expression. Such compound diagnostic expressions have to be recognised and decomposed before coding. This can be based on a linguistic approach but also on a corpus based method provided that we have a corpus entirely consisting of expressions that refer to single disease conditions.

The n-gram tree algorithm that we proposed in this thesis appears to be useful not only for decomposing compound diagnoses into single entities but also to suggest codes to the user. Because the algorithm was tested on a sample different from the samples used for performance testing of the Naïve Bayes and vector-space methods, we cannot compare their performances. The obtained recall of more than 80% that was measured on a small sample is close to the results reported when using various corpus based-methods. But it has one serious drawback compared to other corpus-based methods. It is the difficulty of setting up a reliable corpus. I argued that for statistical methods even non error-free corpora could still be useful in certain situations. This is not the case with the n-gram method. First, the corpus should not contain any compound diagnoses and only a careful manual evaluation can ensure this. Second, if some diagnoses in the corpus are erroneously coded, this error will always appear when the corresponding full n-gram is matched with the query diagnostic expression. The n-gram method does not provide a metric that could be used to rank the candidates: all proposed codes are equally probable.

The N-gram method is an efficient algorithm both for decomposing and coding but it is more sensitive to the quality of the learning set than the other methods. Since this method cannot rank the candidate codes, it cannot take advantage of the higher prevalence of the correct codes, in contrast to methods such as the Naïve Bayes.

**Question 5**

*To which extent is it possible to transform traditional coding schemes, like ICD, into a formal representation?*
Legacy coding systems such as ICD were developed for human use, and formal representation of the categories did not seem necessary when it was created. The meaning of the categories was conveyed by natural language. In order to ensure consistency of the permanently growing system including its use in modern computer systems, the formal representation became increasingly more important. There are a number of re-usable reference ontologies that can be used to build up a formally represented ICD. In this thesis we tested reference ontologies on the first two chapters of ICD. We found that these reference ontologies and the expressive power of description logic languages seemed to be sufficient to express the meaning of ICD categories. Some features, like role propagation (e.g. a disease of a part of heart is a heart disease) require workaround solutions (since the publication of our paper, however, a new version of the Web Ontology Language has been published which solves this problem.) The frequently used modifiers ('other', 'not elsewhere classified' and 'not otherwise specified') caused many difficulties. The formal representation that we created was tested on an NLP based computer assisted coding system yielding promising results.

It is not only possible but also advantageous to translate ICD to a formal language. This can be used not only to assist disease coding but also to help to improve the consistency of ICD during the maintenance of the system.

Question 6

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

The final goal of coding clinical cases is to create public health databases that contain aggregate data reflecting the health status of a population and/or the activity of the healthcare system. Apart from this type of disease-specific public health indicators, a number of other indicators are used in such databases. Many of them are also aggregations. This data aggregation is based on methodologies that vary from country to country. The different methodologies, different definitions and observation techniques make cross-country comparisons challenging.

In this thesis, we developed a pilot ontology model of public health indicators and found that since public health indicators are information objects their formal representation is quite different from the representation of real world entities. An information object is about something, but the subsumption relations between information objects do not follow the relations between the real world entities that the information objects describe. The formal representation of differences in methods of data collection and calculation complicates the ontology further as they pose requirements on ontologies not addressed before.

Development of top level ontologies were recently undertaken. They are increasingly more focusing on real world entities (this is the traditional field of ontology). Public health indicators are social artefacts, and the formal representation of all of their aspects requires more theoretical foundations. For this
reason some compromise seems to be necessary between rigorous rules of formal ontology and practical applicability.

2. General considerations
When, in the early 1990-s as a young physician, I first met the problem of classification I realized that ICD formed a heavy heritage with all its oddities, inconsistencies and peculiar categories. I was really enthusiastic to get rid of this “heritage” and it took quite a long time before I understood that changing to a brand-new classification scheme could be beneficial perhaps for the future but obviously would be a disaster for the past. Since then a tremendous amount of information was collected not only in Hungary but also worldwide. I realized then that this data may contain a treasure even though these data are noisy and often ambiguous. Indeed, when I first had the possibility to analyse a larger dataset of hospital cases with coded diseases and surgical procedures, I realised that quite often it was hard to infer – at a clinical level – what the real problem of the given patient was and what exactly happened. The inconsistency, the coding errors, and the vaguely defined categories will remain for a long time. While all efforts to reduce these problems are welcome, in the foreseeable future we will not be able to totally get rid of them. Besides, these efforts may result in shifting the problems elsewhere.

The human knowledge about classifications (how to build classifications and how to use them) cumulated through many centuries. Nowadays we try to represent this knowledge in computer systems. One way to do this is the use of formal ontologies and logical inference engines. Aside from this explicit knowledge representation there is an alternative: to collect huge amounts of data in free text that implicitly represent the human knowledge and find tools that are able to exploit this implicit representation. This is what corpus-based coding tools do. Building consistent classifications is obviously much easier if the classes are described in a formal ontology. However the use of classifications in practice (the task of coding) can be effectively supported by corpus-based methods not necessitating much effort in formalizing the classification system. Again, when coded data are aggregated in various statistical databases, and when one wishes to retrieve data for analysis and discover new knowledge, formal methods can be useful (e.g. by means of an ontology-based terminology server or an ontology of public health indicators). In this sense implicit and explicit knowledge representations have a complementary relation.

3. Further research
Finally, I would like to discuss the necessary further research in a closer, ‘narrow angle’ and a ‘wide angle’ view.
Narrow angle view
I attempted to show in this thesis that classifications are results of a culture-dependent development process that is still in progress. Emergence of formal ontologies and computer systems that are able to process them is quite new in this history, and expectably will essentially reshape future biomedical classifications. A tremendous work in this direction is currently done, and it can be strengthened by foundational and top level ontologies.

Such ontologies are becoming available; and many of them are already in the public domain. Yet, important components are missing, like a practically useful ontology of information objects that could serve as a basis for a full blown ontology of public health indicators.

In concrete terms, to overcome current barriers and meet the challenges in the use of biomedical classifications, further research seems to be necessary in following directions:

- Testing the performance of statistical computer assisted coding tools in real environments and developing friendly user interfaces.
- Developing consistent classifications based on formal ontologies.
- Seeking categories that can be used to describe the health of a population instead of aggregating data on health conditions of individuals.
- Developing formal descriptions of the categories of the legacy coding systems, and developing applications (terminology servers) that can exploit them.
- Developing ontologies that support a formal description of public health indicators.

Wide angle view
While a continuous effort is necessary to continuously improve the quality of our data and classifications, we have to look in parallel at our data from a new perspective. New methods of data analysis have to be developed in order to be able to study populations as systems. Most of the currently used indicators are simple statistics: sums, averages and ratios. But these indicators consider populations just as sets of individuals. It is, however, likely that populations have certain features that cannot be derived directly from the properties of individuals. This is similar to elementary particles and atoms. When electrons, protons and neutrons together compose an atom, this emerging structure depicts features that none of the elementary components has. Those particles have e.g. mass, electric charge but no chemical valence. The mass or electric charge of the atom is (more or less) the sum of the mass and charge of the constituting particles. But chemical valence is something that simply does not exist at the level of elementary particles. Similarly, populations can be considered as systems composed by individuals. Our currently used classifications are based on concepts used to describe individuals and we can sum up those properties. Currently we do not have any tool to grasp the emerging features of populations. But further study on the tremendous amount of data that we have collected might lead to the invention of new, emerging features of societies (or populations) that are not simple sums or averages of some properties of individual human beings.