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.
SUMMARY

The subject of my study is the generation and utilisation of coded medical 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 public health research and for policy decisions. A great portion of these data is coded. This thesis concentrates mainly on disease coding, more specifically on the use of the worldwide used International Classification of Diseases (ICD). The coding procedure takes as input some free text extract of the medical record, such as a discharge letter or just the stated clinical diagnoses. This procedure is error prone; therefore the use of coded data in reimbursement, research and policy making requires some precautions. 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.

Chapter I provides a general introduction, briefly describes the "life cycle" of coded data from their production to their aggregation into national and international statistics. Then the following research questions are addressed:

1. How serious is the validity problem, and what are the main causes of coding errors?
2. What are the cultural determinants of current biomedical classifications and how do these determinants challenge the use and reuse of coded data?
3. How to compare and evaluate the performance of various statistical computer-assisted coding methods? What are the factors that influence their performance?
4. To which extent a corpus based algorithm is able to recognise compound diagnostic expressions and decompose them into single diagnostic entities?
5. To which extent is it possible to transform traditional coding schemes, like ICD, into a formal representation?
6. To which extent is it possible to improve comparativeness of public health indicators by an ontological framework?

These questions provide the organisation of the following chapters.

Chapter II is a literature based overview of the nature, causes and magnitude of coding errors. The validity of the coded information is unsatisfactory in general, however the ‘correctness’ is purpose and environment dependent. This chapter gives a general framework of the coding process for detecting potential error sources. 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 (such as principal disease and complications). Each step corresponds to 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, 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.

Since ICD itself is a source of errors, it is worthwhile to look at its development as a historical process. **Chapter III** gives an outline of the history of medical classifications in a general cultural context. Classification is a general phenomenon in science and has an outstanding role in the biomedical sciences. Its general, principles were already developed in ancient times, while domain classifications, particularly medical classifications have been constructed since the 16th and 17th centuries. I demonstrate with several examples that all classifications reflect an underlying theory. The development of the notion of disease during the 17th-19th centuries markedly influenced disease classifications. Development of classifications currently used in computerised information systems started before the computer era, but computational aspects reshaped the whole picture. A new generation of classifications is expected in biomedicine that depends less on human classification effort and uses more the power of automated classifiers and reasoners.

Even very simple computer applications can improve the quality of coding. Since there is a wide range of computer assisted coding methods, many of them are based on some corpus of manually coded cases that is used as a learning set for some learning algorithm. **Chapter IV** deals with the evaluation problems of such methods, aiming at the investigation of the influence of corpus quality on the performance of corpus-based coding methods. Two corpora of manually ICD-coded diagnostic expressions differing in language (German and Hungarian), size (93,863 vs. 53,198 expressions), heterogeneity, and quality were subjected to two different coding methods (based on Naïve Bayes and vector-space). The methods were evaluated using 5 fold cross-validation on both corpora in terms of their ability to place the correct code in a specified rank range. In the German corpus the performance was also determined per ICD chapter. In the Hungarian corpus the performance was also evaluated against a silver standard of a dataset of 300 expression-code pairs, where codes emerged by consensus of additional human coders. The corpora differed markedly in terms of redundancy (multiple occurrence of the same expression) and inconsistency (the same expression is assigned to different codes). On the German corpus the vector space method appeared to be slightly superior to Naïve Bayes. The difference was statistically significant but of no apparent practical relevance (52 vs. 50% precision for the first ranked code). On the Hungarian corpus Naïve Bayes drastically outperformed the vector space method (71% vs. 21%). Performance varied strongly per ICD chapter but differences among chapters were higher than differences between the methods. The performance of the Naïve Bayes method improved on the silver standard but this was not the case for the vector space method. As a conclusion we state that superiority of one method over another strongly depends on specific quality characteristics of the used corpora. The observed performance discrepancies were attributable to differences in redundancy and inconsistency of the corpora and in performance evaluation one should hence
Summary

inspect the characteristics of the expression-code relationship. An inconsistent sample, however, might have the advantage of offering different options to a human coder focusing his or her attention to alternative codes that might fit best a given expression.

A common problem in evaluation of computer assisted coding methods is that the cardinality between diagnoses and codes is often not one to one. One of the reasons is that clinicians often use compound diagnoses: one expression may refer to more than one condition that have to be coded. Chapter V presents a study to find an easy to implement method to detect compound medical diagnoses in the Hungarian medical language and decompose them into expressions referring to a single disease.

A corpus of clinical diagnoses extracted from discharge reports (3079 expressions, each of them referring to only one disease) was represented as an n-gram tree (a series of n consecutive words). A matching algorithm was implemented in an algorithm that identifies sensible n-grams that exist both in test expressions and in the n-gram tree. A test sample of another 92 diagnoses was decomposed by two independent humans and by the algorithm. The decompositions were compared to measure the recall and the precision of the method.

There was no full agreement between the decompositions of the humans (a fact that underlines the relevance of the problem). A consensus was arrived for all disagreements by a third opinion and open discussion. The resulting decomposition was used as a gold standard and compared to the decomposition produced by the computer. The recall was 82.6% the precision 37.2%. After correction of spelling errors in the test sample the recall increased to 88.6% while the precision slightly decreased to 36.7%.

The proposed method seems to be useful in decomposition of compound diagnostic expressions and can improve the quality of diagnostic coding of clinical cases. Other statistical methods (like vector space methods or neural networks) usually offer a ranked list of candidate codes either for single or compound expressions, without warning the user about compound diagnoses that may require multiple codes. Our method is especially suitable in situations where formal NLP techniques are not available, as is the case with scarcely spoken languages like Hungarian.

The use of coded data can be limited not only because of their questionable validity but also because of data retrieval problems. Researchers may formulate questions that first have to be translated to a set of codes. For example, one might be interested in the number of cases residing in a database that are related to some lung disease. Then the set of all ICD codes that express some lung disease has to be defined before the actual number can be retrieved. In Chapter VI I present our experience with the formal representation of ICD that can be used to solve such terminological problems but also can be used in knowledge-intensive coding support tools. In this study the meaning of the ICD10 categories is represented using the GALEN Core Reference Model. Due to the deficiencies of its representation language (GRAIL) the ontology was transformed to the quasi-standard OWL. A test system which extracts disease concepts and classifies them to ICD10 categories has been implemented in Prolog to verify the feasibility of the
approach. The formal representation of the first two chapters of ICD10 (infectious diseases and neoplasms) has been almost completed. The constructed ontology has been converted to OWL. In future work FMA (Foundational Model of Anatomy) will be used as anatomical reference ontology.

The last stage of the "life cycle" of coded data is their aggregation into national or international public health databases. Such databases aim to provide time serialised data (called indicators) to compare health conditions in different countries or territories. Their data sets are more or less overlapping but data from different databases and different countries are hard to compare due to different definitions and interpretations. The aim of Chapter VII was to create a core ontological model that is able to represent public health indicators. We assumed that by such representation comparability and quality of the data could be improved.

Three indicator sets were taken, and a core ontology was built from information objects, describing their top level entities. The Protégé ontology editor with RDF backend was used for building the ontology. The used indicator sets included those of the Health for All Database of the World Health Organisation (HFA), the OECD Health Data of (Organisation for Economic Co-operation and Development), and the set of indicators proposed by the European Community Health Indicators (ECHI) European project. Then 19 indicators selected from HFA were represented using the core ontology. Strengths and weaknesses of the descriptive capability of the model were studied.

The drafted core model seems to be useful in representing many of the public health indicators. In some cases it really helps improve comparability. However some of the semantic details can not be sufficiently expressed by the used ontology representation language. There is a need of merging other domain ontologies to represent indicators related to other domains, such as economy, social and environmental sciences.

Finally Chapter VIII summarises the answers to the questions listed in the Introduction. ICD coding cannot be totally free of errors. While the usual error of manual coding (about 30%) does not mean that the data are totally useless, it is worthwhile to try to reduce the error rate. To do so it is essential to understand the causes and sources of errors that can be made with the help of the proposed logical framework of the coding process. For example understanding the difference between the notion of diagnosis and disease is an important step. As a consequence of the fact that classifications are culturally determined creations, classified and coded data from the past become more and more difficult to interpret as time goes on. Formal representation of classifications might be a tool that helps to solve this problem. The methods proposed to improve the quality of coding should be evaluated with caution. We often measure how good computers are in reproducing what humans do (including errors). Even doing so, superiority of one method over another may depend on some quality features of the used corpora. While the studied corpus-based methods and the investigated ontological approaches really can improve the quality and usability of coded data, there is still a lot to do in this field in the future including:
Summary

- 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.