A semantic clinical data repository – how the work on DRGs can serve clinical medicine, too

Hans Rudolf Straub, Michael Lehmann
Semfinder AG, Kreuzlingen

Summary

Most of the clinically relevant information on a patient is documented in freetext. Diagnoses and procedures in particular are usually written in free wording. They are therefore not structured and cannot be analysed statistically without prior manual processing. When diagnoses and procedures are coded (ICD-10, CHOP), only a small amount of the primary information is retained and the structuring is poor.
However, with a fully automatic semantic coding programme the primary documentation's input text is cast in an internal form which is well structured, keeps the detailed information of the original text and reconstructs implicit meanings. This internal representation is usually discarded in the process of coding and only the final codes are outputted. It could, however, be exported to a data base (CDR = Clinical Data Repository). Thanks to its structured and detailed representation of clinical facts, diagnoses, procedures, medications and others, very precise statistics could be calculated. The CDR could thus serve scientific purposes as well as clinical management of the patients (e.g., alerts when prescribing contraindicated drugs). In addition, easy-to-perform online queries in natural language of the full patient base's clinical information are possible.
The precise semantic structure of the internal form, the preservation of the full original information and the reconstruction of implicit information allow much more definite answers than queries of free text or of the poorly structured and less detailed codes. Because the internal form originates automatically when coding, no additional work has to be done by clinicians or coders.

DRGs: A big part of the workload falls on clinical personnel

SwissDRGs will be implemented and they will have implications it is already worthwhile to consider [1]. However much the expectations and fears of those concerned may differ, one prognosis can be regarded as certain: a considerable workload awaits the clinics in Switzerland.
The workload will fall not only on staff concerned with billing and controlling. As DRGs are based on the complex field of patient diagnosis and treatment, the primary information must come from those who actually see and treat the cases: the doctors in the clinical departments must provide the necessary data. This means that their documentation must be precisely formulated, but also, in point of fact, that clinicians must relearn and document not only the clinically relevant facts, but pay particular attention to those characteristics which are relevant for assigning the appropriate DRG – in providing, for example, the correct and honest arguments for a more costly DRG (right-coding). The accurate coding (ICD-10 and CHOP) can only be done when all the necessary facts are known, and this means that the clinical documentation must be complete *and* DRG-oriented. The clinician must know what information is crucial for the coding and grouping of the billing section.
In view of the additional workload to be expected for clinicians, one may enquire whether the clinical department does not directly recover some value from which the actual work of the physicians with the patients could benefit. It is the opinion of the authors that such a return is not only possible but could be earned with modest effort.

What information on a case is documented?

We should be aware that only structured information can be analysed reliably. Structured means that the scale of a variable and its possible values are well defined. In this way we can compare costs (scale = CHF) or the age of patients (scale = years). Unfortunately, this well-structured comparability is applicable to only a few patient characteristics. And the clinically interesting data, i.e., diagnoses and procedures, have no common scale and not even a mandatory set of standard values. Diagnoses and procedures are therefore not statistically analysable without prior editing by hand.
Clinicians traditionally note diagnoses and procedures in their documents in free wording. The wordings in use have stood the test of centuries and colleagues can take all essential information concerning the case from them. But this is only true if one looks at just one case.

Correspondence:
Hans Rudolf Straub
Semfinder AG
Hauptstrasse 53
CH-8280 Kreuzlingen
straub@semfinder.com

When cases are grouped, statistically analysed or their content processed by computers, they must be edited in advance by hand. This task is time-consuming and error-prone.
To be sure, SwissDRGs offer, together with a thorough coding of diagnoses and procedures, a systematic structuring of medical cases. Unfortunately, this structuring is inappropriate for clinical use. We will show why this is the case and how the structuring can be made suitable for clinicians with very little effort.

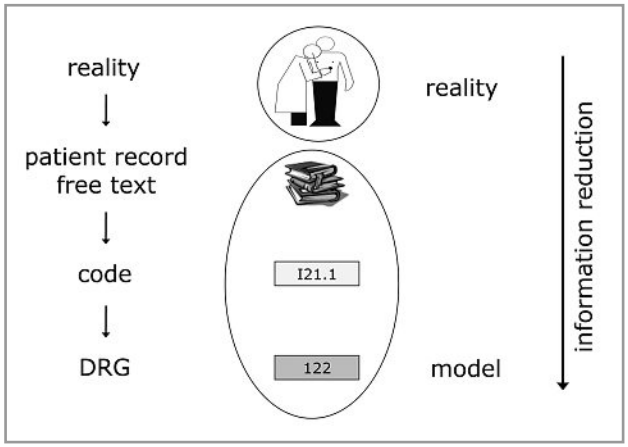

Figure 1
Information reduction.

Figure 1 shows the flow of information from the real situation (patient and physician) to the documentation, the coding and finally the DRG. It is easy to understand that the quantitative information content diminishes continuously thereby. It is difficult to overstate how drastic this reduction of information is. While for a real patient, e.g., a case of acute appendicitis, in principle each of the approx. 25 thousand billion red blood cells could be counted, localised and described in detail, the laboratory documentation merely mentions Hb = 15 g/dl. In the discharge letter the diagnoses, the ICD-10 codes and the DRG contain no reference to the red blood cells, due to lack of relevance. This, of course, is not only the case for the red blood cells. However we look at the patient, there is always a radical reduction of information between what can be found in the real life case to the notes in the documents and further to the diagnosis codes and DRGs. This radical reduction of information is intended and nothing but sensible: to gain an overview of the facts, we leave out the less important details [2]. This is true for all levels of documentation: while a diagnosis in text form is a choice among more than 1000 million possibilities, an ICD-10 code is one out of 15 000 and a DRG one out of 1000.
The question is, which information is omitted by each step in this process? Which information is kept will depend on the goal of the coding and grouping, and so will the information to be dropped (fig. 2). There is no natural or mandatory way to select the information to be dropped when classifying and grouping diagnoses. There is no natural hierarchy in which diagnoses and procedures can be arranged in an orderly manner, such as we find in the systems of zoology and botany [2].

The facts of the case become more evident when we look at an example. In figure 3 the coding of a free text diagnosis is shown. In the field at the top the input text from the patient record (EPR) is entered: "E. coli cystitis". Below that we see a construction of several boxes, a "concept molecule", which represents the content analysis of the diagnosis text by the semantic coding programme [3, 4]. Each box represents an atomic concept obtained from the text by the programme. These concepts are not accidentally chosen words but well-defined nodes of a carefully elaborated semantic net. Various wordings of the same diagnosis always lead to the same atomic concepts. The arrangement of the boxes shows their conceptual relations, and in doing so the "semantic space" in which the semantic net is spread. The representation is unambiguous, structured and complete, i.e., the complete information of the input text is retained. For the ICD codes this is not the case. In figure 3 the two ICD-10 codes which together code "E. coli cystitis" are shaded in grey. But the two codes do not contain the text's full information. For various reasons, an ICD-10 type of code can only represent a reduced information content of a diagnosis in a structured way (fig. 2, see also [2]). While the input text carries the full information content, but not in a structured form, the ICD code is roughly structured but

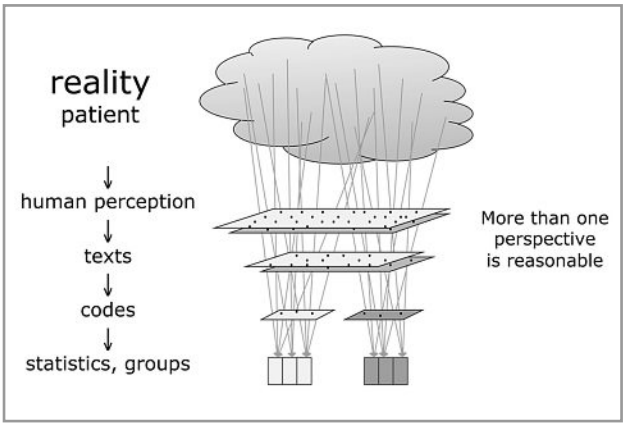

Figure 2
Depending on our goals, codes and groups retain different aspects of the primary information.

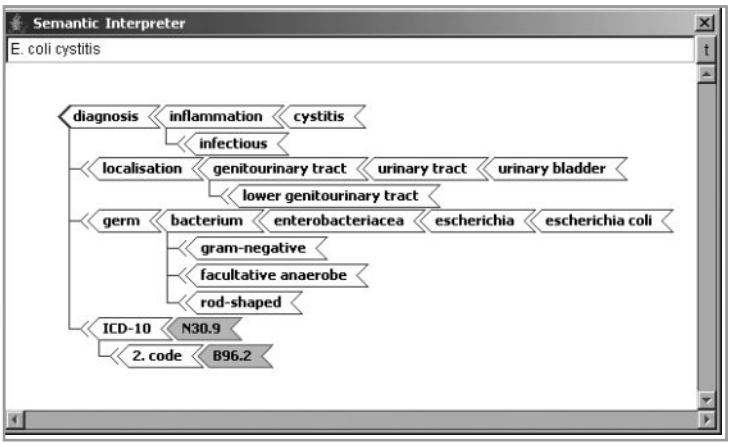

Figure 3
Coding of the diagnosis "E. coli cystitis" – and the semantic analysis behind it.

contains only a part of the original information. The internal representation, however, which develops automatically during the semantic coding, is both complete and well structured. Due to its inherent systematic structure, it could easily be analysed in statistics and queries. But because the internal representation is used only for the coding process, it is cleared immediately after creation and only the codes are issued.

A semantic Clinical Data Repository (CDR)

The well-structured and complete information (the fabric of boxes in fig. 3) which emerges automatically during the semantic coding process could be given to a data base without any additional effort on the part of physicians or coders. Thus the complete semantic information content of the diagnoses of all the clinic's patients would be directly accessible for computer evaluation.

While the DRGs represent each case only from the economic point of view and the ICD-10 codes lead to an uncertain and faulty evaluation, the structured semantic representation of figure 3 allows the clinicians a direct, targeted and precise analysis of the full set of patients' diagnoses. Thus a search for "gram-negative infections of the lower genitourinary tract" matches directly with the diagnosis "E. coli cystitis", although none of the words of the diagnosis are mentioned in the query. But the internal composite semantic representation formed during the coding process of the diagnosis already lists all the searched concepts. The query can also directly access all those concepts which are only implicitly contained in the diagnoses, such as "gram-negative", because the semantic analysis of the coding programme makes them explicit. If the atomic concepts of the internal representation are not directly expressed in the query, they are found, too, e.g., in a search for "catarrhs of the bladder": the programme's semantic machine translates such expressions to its unique atomic concepts – as it reduces synonymous expressions in the patients' diagnoses to its structured basic concepts when coding.

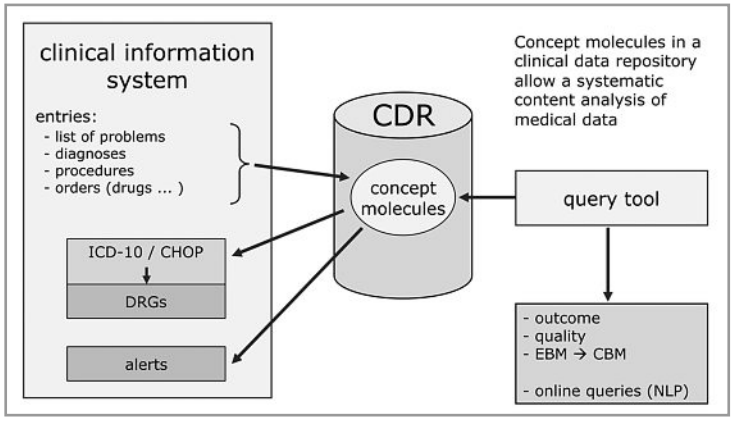

Figure 4
Benefits of a semantic Clinical Data Repository (CDR).

Efforts for and benefits of a semantic clinical data repository

The workload of building and maintaining a clinical data repository does *not* fall on the physicians and coders. Because the semantic representation of figure 3 is automatically built by the programme during the coding process, it involves no human work.

The effort is only a technical one: the internal representation of the coding programme must be stored in a database. For this purpose a unidirectional interface between the programme and the database must be implemented. Then a query programme for searches of the database must be written. This programme most reasonably uses the existing semantic interpretation machine of the coding tool, so that queries in natural language (NLP) are possible (fig. 4). This allows clinicians with scientific intent to search the content of the CDR without training, simply using their own medical language. The full set of the clinic's patients' routine medical data can be queried easily and precisely for its semantic content. Apart from these online searches, preformed queries are possible, e.g., for precise and detailed operation statistics and for alerts in the clinical routine, warnings for contraindications when prescribing drugs and hints for useful procedures in particular clinical situations. Due to the complete and well-structured semantic representation of the clinical data, such warnings and hints will be very precise and without the many irritating false positive alarms of the conventional solutions based on vocabularies or ICD-codes, which discourage their use in practice.

The effort of acquiring the data in the clinical routine is so minimal because the structured semantic representation of the medical data is gained automatically during the routine coding process. Apart from that the programme can build the structured concepts in the background from problem lists or routine diagnosis entries in the electronic patient record. The flexible structure of the semantic concept representation allows it to link the diagnoses' semantic information systematically with information on operations, medications, therapies and laboratory results.

References

1 Berchtold P, Schmitz CH. Eine Zukunft für Spitäler. Schweiz Ärztezeitung. 2010;91(48):1914–6.
2 Straub HR, Duelli M, Mosimann H, et al. From Terminologies to Classifications – the Challenge of Information Reduction, in: Proceedings of the European Federation for Medical Informatics Special Topic Conference, Timisoara, Romania, 2006: 341–52. See also: http://www.semfinder.com/fileadmin/Daten/Dateien/Publikationen/From_terminologies_to_classification.pdf [Last visited 28. Jan. 2010].
3 Straub HR, Frei N, Mosimann H, et al. Simplified Representation of Concepts and Relations on Screen, Proceedings of MIE 2005: pp. 799–804. See also: http://meditext.ch/texte/Simplified_Representation_on_Screen.pdf [Last visited 4.Feb.2010].
4 Oertle M. Natural Language Processing: Real-time-Struktur aus Freitext im Klinikalltag? SMI. 2007(61):15–7.