MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Recording Clinical Data – From a General Set of Record Items to Case Report Forms (CRF) for Clinics

Angela Merzweiler[a], Petra Knaup[a], Ralf Weber[a], Hartmut Ehlerding[b], Reinhold Haux[a], Timm Wiedemann[a]

[a] *University of Heidelberg, Department of Medical Informatics, Heidelberg, Germany*
[b] *Medical University of Hanover, Children's Hospital, Hanover, Germany*

Abstract

Standardising a documentary language makes only sense if we use it for documentation consequently. Using an example of Paediatric Oncology in Germany we have developed a procedure that generates CRFs from a documentary language. The introduced procedure has proved to be feasible in practice. With it we can support developers of documentation systems in creating their CRFs; through the guarantied use of the documentation terminology we further achieve that information recorded with the created CRFs may be statistically analysed across different institutions.

Keywords:

CRF, Case Report Form, Terminology, Clinical Trials, Data dictionary, Structured Data Entry

Introduction

A major goal of medical informatics is the improvement of the quality of information recorded in several information systems in health care [1]. Therefore, there are many efforts to standardise documentation systems and especially the used documentary language inside organisations or even across organisational boundaries [2]. Often there is the problem that a unified documentary language is defined but not used in practice consequently. In this article we want to demonstrate a procedure by an example of multicentre clinical trials in Paediatric Oncology in Germany how we want to guarantee, that the defined documentary language is used in practice by the help of computer based development of CRFs.

Materials and methods

Clinical trials in Paediatric Oncology

In Paediatric Oncology in Germany, much information is shared between Paediatric Oncology Centres, responsible for treating children suffering from malignant diseases, and clinical trial offices, doing research on the best therapy for the children [3]. Clinical trial centres work out guidelines for the treatment of the children whom they refer to the paediatric oncology centres by means of therapy protocols. The Paediatric Oncology Centres report on administered therapies and the course of the disease process of the treated children to the clinical trial offices. For this they use standardised CRFs that are part of the respective therapy protocols. In Germany, there are 24 different clinical trials trying to improve therapy in Paediatric Oncology. Each of these clinical trials collects a set of record items on the patients. But only a part (approximately 30%), called common basic set [4], is standardised across all clinical trials. Two problems arise, because terminology is not at all standardised across different clinical trials. Documentation is made more difficult for physicians, if different clinical trials use special terms for identical concepts (i.e.: block vs. element of therapy) or similar terms for different concepts (i.e.: prephase of a therapy and pretherapy). Above all this occurs if concepts are used that are not defined in universal medical dictionaries. The second problem is, that the clinical trial offices can perform only a strongly limited number of statistical analyses across different clinical trials.

The data dictionary for Paediatric Oncology

The problems mentioned above that result from a not unified documentary language shall be solved by the project "computer-based data dictionary for Paediatric Oncology". This project wants to reach the following objectives [5]:

1. *To define a unified standardised terminology.*
2. *To design methodologies to guarantee that the unified standardised terminology is applied for the development of CRFs for clinical trials.*
3. *To develop a computer-based data dictionary that supports the management of the defined terminology.*

The individual clinical trials collect just an excerpt of the set of standardised record items. Moreover, even after the process of standardising record items, the set of possible attribute values may differ across different clinical trials. For example, each clinical trial stratifies its patients in different appropriate strata according to specific criteria. Therefore, we want to allow the use of trial-specific codes. In order to see which clinical trials are effected by changes of the uniform documentary language a manager of the unified documentary language therefore needs the following information: which clinical trials collect which record items with what possible attribute values. In addition, we want to support the clinical trial offices to create databases for storage of collected data in their clinical trial offices as well as in the computer-based documentation systems in the clinics. We can accomplish this support only if we can guarantee that the unified documentary language is used on the CRFs. For these reasons we want to offer the capability of creating standardised CRFs through computer-based composition of record items of the standardised documentary language. In this paper we describe the method to attain the formal definition of CRFs from a set of record items of a standardised documentary language.

Results

Structure of the data dictionary

As described above, the data dictionary shall store and provide two kinds of information: the standardised documentary language itself as well as information about all CRFs (Which CRFs exist and which record items are recorded on them with which possible attribute values). Therefore, the data dictionary consists of two parts: One general part provides the standardised terminology and another clinical trial specific part that contains information about the trial specific usage of record items.

The general part of the data dictionary

The general part of the data dictionary is a collection of record items that are at disposal for documentation purposes. These record items are composed of three concepts: An object class, whose data should be recorded, an attribute type specifying the characteristic, that shall be recorded, and a set of possible attribute values. An example is shown in Table 1.

Table 1 – Examples for record items

Object class	Attribute type	possible attribute values
Disease	Space of time	{past, present}
Disease	kind of	{syndrome, malignant disease, ...}

We distinguish 5 types of record items due to their data types:

- *unlimited (except length) text*
- *numbers*
- *date/time*
- *set of possible options*
- *references to other object classes*

For each concept used for the definition of record items the following information is stored: (For an example see figure 1.)

- *definition as a text or per genus et differentium*
- *preferred term and synonymous terms*
- *generic, partitive and associative relations.*

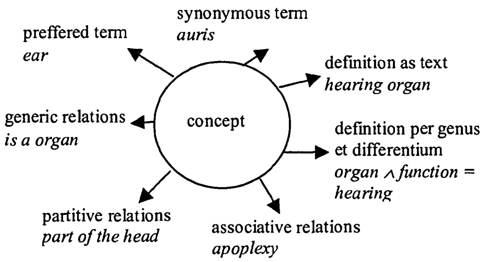

Figure 1 – example for information about the concept ear

The structure of CRFs

Figure 2 shows the typical structure of a CRF. It consists of a CRF term (e.g. first report form) and a set of contexts (e.g. master data, prior diseases). The contexts are strictly monohierarchical structured. The context "prior diseases" is a subcontext of "mother" for example. To these contexts attributes are assigned. Attributes consist of an attribute type (e.g. date of birth) and a set of possible attribute values (e.g. years between 1900 and 2000). Each attribute is assigned to exactly one context.

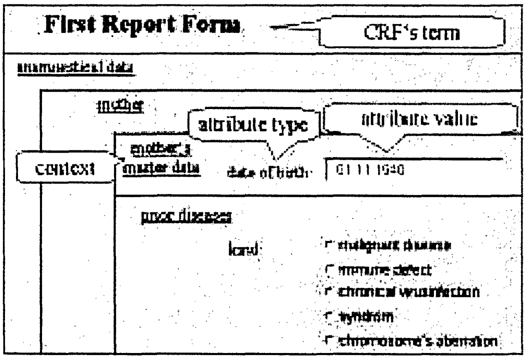

Figure 2 – Structure of CRFs

Construction of CRFs by the use of a standardised documentary language

According to the above-described structure of CRFs the creation of CRFs consists of three steps:

1. *definition of the CRF*
2. *structuring the CRF into a hierarchy of contexts*
3. *filling the contexts with attributes*

We describe these steps in the following section.

Definition of CRFs

Defining a CRF one firstly has to assign a term to the CRF. This term tells us, when the CRF should be filled out and which kind of attributes should be reported. Typical terms are "First report form" or "course of the disease form". Although, down the road, documentation will be mostly done with the help of computer–based documentation systems, paper–based CRFs must still be available since some clinics don't want or can not use the specific documentation system. Paper-based forms do not necessarily need to exactly match the computer-based forms. Therefore each CRF should be characterised whether it should be established as a computer-based or as a paper-based CRF.

Building contexts

If one wants to build contexts, one has to assign a term as a heading to this context that subsequently will appear on the CRFs. The chosen term (e.g. patient's master data) specifies the class of objects whose attributes should be recorded in this paragraph of the CRF. So the heading of the context corresponds with the object class of record items of the general part of the data dictionary. As described before the general part of the data dictionary contains information on which attributes can sensibly be reported for each object class. Therefore the chosen term controls the selection of attributes that can be recorded in this context. The example in Table 2 shows record items that can be attached to the context "patient's master data".

Table 2 - record items that can be attached to the context "patient's master data"

Object class	**Attribute type**	**Possible attribute values**
master data	name	Text
person's master data	first name	Text
person's master data	date of birth	Date

In the general part of the data dictionary we have defined object classes (e.g. master data) as general as possible. Opposite to that the terms of the contexts should be as specific as possible so that the set of objects whose data should be recorded is specified precisely. For example: Object class: "master data", context's term: "patient's master data" or "master data of the insured person". By the selection of a more specific term it's possible to limit the set of possible attribute values inside a context. For example it is reasonable to define a context "prior diseases" as shown in Figure 2, although there is just an object class "disease" in the general part of the data dictionary. With this more specific term the user realises that he should only document prior diseases. By that way the set of possible attribute values of the record item "disease – space of time" is limited to "past" automatically. Precondition for this is that the general part of the data dictionary defines how attribute values are limited by more specific context's terms.

The term chosen for the context has another relevance: It fixes, which subcontexts can be built.

Building Subcontexts

Due to the strict monohierarchical structure of contexts in a CRF, contexts can just be built as a root context of a CRF or as a subcontext of an already defined context. There are two kinds of subcontexts:

1. *subcontexts associated with superordinated contexts and*
2. *subcontexts that are in an is- a relationship to the superordinated contexts.*

Ad 1: With the aid of the record items of data type "reference to other object classes" defined in the general part of the data dictionary the hierarchy of contexts can be extended. If there are record items of the data type "reference to other object classes" a new subcontext can be built on the basis of the referenced object class. Example: The subcontexts "mother's master data" and "prior diseases" can be defined as subcontexts to the context "mother", because the object classes "person's master data" and "diseases" are associated with the object class "person" in the general part of the data dictionary.

Ad 2: In a context, characteristics of a special object class are recorded. If additional characteristics have to be recorded for a subclass of this object class, it is possible to build subcontexts of a context referenced by a is-a relationship. Example: The response to a specific therapy block should be recorded for all patients by the attribute "grade" of the object class "response" with the possible values "good, bad, not existing, not measurable". For a response classified as a good response, the clinicians should further be asked whether they could reach a complete remission of the tumour. In this case the clinician has to fill out the attribute "CR reached" of the object class "good response" with the possible values {yes, no}. This attribute cannot be captured in a context "response", because it is an attribute of the more specific object class "good response". On the other hand one can not capture both attributes in a context "good response", because all patients would be classified as good responders on therapy automatically due to the more specific context heading. So we need two contexts: the context "response" and its subcontext "good response". This is only possible, if the subcontext is build on the basis of a concept being defined as a more specific concept (per genus et differentium) of the heading of the more general context.

Adding new attributes to contexts

Now, the existing structure of contexts must be filled with attributes. It was explained above which attributes fit to a context. Here we just want to emphasize that it is impossible to add record items of the data type "reference to other object classes" since these record items only serve to build subcontexts.

By adding attributes to a context the term for the attribute type is automatically filled up with the corresponding preferred term, in order to assure, that terms on CRFs are uniform.

For record items of the data type "option" it is possible to adapt the choices to the specific clinical trials. This works only in the way, that granularity of the choices is adapted to the particular clinical trial. Furthermore, a code must be assigned to all options.

Discussion

We have developed a procedure for computer–based defining of CRFs based on a given documentary language. The creation of contexts plays a central role in this procedure, because they clearly arrange CRFs. Moreover, we want to achieve that no attributes are added to contexts whose presence depends on other attributes of the same context. At the moment, we use this procedure in the field of multicentre clinical trials optimising therapy in Paediatric Oncology. Until now, we have defined CRFs that are typically for Paediatric Oncology in Germany (e.g. first report form, course of the disease form) for capturing record items of the minimum basic set. For this the described procedure proved to be feasible.

In the field of Paediatric Oncology this procedure could be optimised by using the template CRFs of the minimum basic data set. These template CRFs can be used as a basis of the CRFs by adding clinical trial specific contexts and record items and adapting already available attributes to their needs.

With the described procedure we have defined contents of CRFs. In order to get a particular real usable CRF, further steps are necessary: like layout, printing (in the case of paper-based forms), creation of a database for saving the recorded data within the documentation system of the clinic, creation of a database for saving the recorded data within the clinical trial office and the programming of the functionality of the documentation system (in the case of computer-based forms). Nevertheless, the procedure can already support the heads of clinical trials in Paediatric Oncology in Germany. If they want to plan CRFs they are supported by defining them on the basis of the standardised documentary language. If they use the data dictionary we can guarantee that terminology of the resulting CRFs is conform to the standardised documentary language. But they have to do further steps as described above (e.g. layout, programming of the functionality) manually. We can also tell them, how a database schema can be derived for storing the recorded data: For that they should use the record items of the general part of the data dictionary. One table has to be created for each object type. The columns of the tables correspond to the attribute types. References between different object classes must be represented by foreign key attributes in the referenced tables (possibly mere reference tables must be added). The procedure shall be expanded by a component for summarising tables reasonably. If the database schema is just derived from contexts of CRFs there is the risk that this procedure would create a confusing number of tables. As an alternative to this method of deriving a database schema the EAV/CR representation is introduced by Nadkarni [6]. We have chosen a conventional design, because we are afraid that the EAV/CR (Entity attribute value model with classes and relationships) model would not be accepted by the heads of clinical trials in respect to two drawbacks of EAV Design: "It is less efficient ... for bulk retrieval of numerous objects at a time" and "The process of performing complex attribute-centric queries ... is technically more difficult." [6].

The EU funded GALEN –project has set itself the goal to create CRFs for capturing data by a computer–program from a general concept model [7]. With the aid of an example of capturing data for a DIABCARD it was proven that this is possible [8]. As far as the authors know, the method has not yet been made available for common use.

Conclusion

In this article we have introduced a number of steps, that are necessary for the creation of CRFs based on a uniform documentary language. By applying this procedure, we can make the documentation easier for clinics that participate on several clinical trials, and we enable analyses across different clinical trials.

Acknowledgments

The development of the data dictionary has been funded by the German Leukaemia Research Aid (DLFH). It is currently part of the health care research network paediatric oncology and is funded by the German Ministry of Education and Research.

References

[1] Tang PC, LaRosa MP, and Gorden SM. Use of Computer - based Records, Completeness of Documentation, and Appropriateness of Documented Clinical Decisions. *Journal of the American Medical Informatics Association* 1999: 6(3 Suppl) piii: 245-251.

[2] Silva JS. An Architecture for National Scale Clinical Trials. *MD Computing* 1999: 3 piii: 43-44.

[3] Winkler K. Multizentrische Therapiestudien in der pädiatrischen Onkologie - Forschung und Qualitätssicherung. In: Michaelis J, Hommel G and

Wellek S, eds. *37. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS) in Mainz Sept. 92.* München: MMV,1993; pp. 159 - 167.

[4] Sauter S, Kaatsch P, Creutzig U, and Michaelis J. Erstellung eines einheitlichen Basisdatensatzes für den Bereich der pädiatrischen Onkologie. *Klinische Pädiatrie* 1994: 206 piii: 306-312.

[5] Merzweiler A, Knaup P, Creutzig U, Ehlerding H, Haux R, Mludek V, Schilling FH, Weber R, and Wiedemann T. Requirements and Design Aspects of a Data Model for a Data Dictionary in Paediatric Oncology. In: Hasman A, Blobel B, Dudeck Jet et al, eds. *Medical Infobahn for Europe Proceedings of MIE 2000 and GMDS 2000.* Amsterdam: IOS Press, 2000; pp. 696 - 700.

[6] Nadkarni PM, Marenco L, Chen R, Skoufos E, Shepherd G, Miller P. Organization of Heterogeneous Scientific Data Using the EAV/CR Representation. *Journal of American Medical Informatics Association* 1999: 6 piii: 478-493.

[7] Rector AL, Solomon WD, Nowlan WA, Rush TW, Zanstra PE, and Claassen WM. A Terminology Server for medical language and medical information systems. *Methods of Information in Medicine* 1995: 34(1-2) piii: 147 - 157.

[8] Ingenerf J, Diedrich J. Notwendigkeit und Funktionalität eines Terminologieservers in der Medizin. *Künstliche Intelligenz* 1997: 3 piii: 6-14.

Address for correspondence

Angela Merzweiler

University of Heidelberg

Institute for Medical Biometry and Informatics

Department of Medical Informatics

Im Neuenheimer Feld 400

69120 Heidelberg

http://www.med.uni-heidelberg.de/mi

angela.merzweiler@gmx.de