Supporting Uniform Representation of Data

Structuring medical narratives for care and research

Renske Los

The work presented in this thesis was performed at the Department of Medical Informatics, Erasmus Medical Center, Rotterdam, the Netherlands. The work was funded by a grant from the Netherlands Organization for Health Research and Development (ZonMW). The grant was part of the Information and Communication in Healthcare program (ICZ – FO: Informatie- en Communicatietechnologie in de Zorg – Funderend Onderzoek).

The author greatfully acknowledges the financial support from McKesson Nederland B.V. for the printing of this thesis.

McKESSON
Empowering Healthcare

Design and Layout: Ton Everaers – www.tonzilla.nl

Printed by PrintPartners Ipskamp, Enschede

Supporting Uniform Representation of Data - Structuring Medical Narratives for Care and Research
Renske Kirsten Los
Ph.D. Thesis, Erasmus University Rotterdam, March 2006.
ISBN: 90-9020381-8

Supporting Uniform Representation of Data
Structuring medical narratives for care and research

Ondersteuning van uniforme data representatie
Medische data structureren voor zorg en onderzoek

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. S.W.J. Lamberts

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
vrijdag 10 maart 2006 om *13.30* uur

door

Renske Kirsten Los

geboren te Roosendaal en Nispen

Promotiecommissie

Promotor:

Prof.dr. J. van der Lei

Overige leden:

Prof.dr. M. Berg
Prof.dr. A. Hasman
Prof.dr. S.E.R. Hovius

Copromotoren:

Dr. A.M van Ginneken
Dr. H.A. Moll

Table of Contents

1 General Introduction

1

'It can readily be seen that all narrative data presently in the medical record can be structured, and […] entered through series of displays, guaranteeing a thoroughness, retrievability, efficiency and economy important to the scientific analysis of a type of datum that has hitherto been handled in a very unrigorous manner.'

-Lawrence L. Weed, 1968 [1].

Introduction

To date, a substantial amount of data needed to make medical decisions are still recorded as free text in (paper-based) patient records. As early as the 1960s, researchers started to computerize the patient record. Electronic patient data are associated with many potential benefits, such as data sharing, quality assessment, research, and management of patient care [2-7]. The goal of electronic medical records is that data once stored in the context of care are readily available for secondary use [8]. To harvest the potential benefits of electronic data, the data must also be available in a structured format to enable processing by computer applications [3]. However, even in electronic records, narrative patient data are often still stored as free text or scanned documents. As a result, for secondary data use such as clinical research, researchers still have to perform the labor-intensive task of reading and interpreting free text in individual electronic medical records. *The research described in this thesis focuses on obtaining patient data in a structured format suitable for both clinical practice and clinical research.*

Medical Narratives

The emphasis of our research is on structuring medical narratives. The medical narrative consists predominantly of physician-gathered, qualitative data [9]. The medical narrative can be found in diverse sections of the medical record: medical history, family medical history, physical examination, progress notes, or reports (e.g., radiology, surgery or

pathology reports) [10]. Medical narrative data tend to be unruly, and the content and level of detail of such narratives are often unpredictable and vary per domain (and even per clinician) [11, 12].

Free text has been the ideal format to collect narratives as it has a high degree of expressiveness [13] which accommodates the unpredictability of the narratives. Free text allows clinicians to record data in whatever words, abbreviations, or codes desired. This is, however, undesirable for research purposes. For research purposes, data are preferably structured and coded [14, 15]. The challenge is to structure the medical narrative in a manner that poses no a priori limitations on detail and that structures the data in a format also suitable for research. An additional aspect that must be addressed for data sharing purposes is uniformity in data representation. An application that attempts to tackle the challenge must be generic, yet tailorable to specific domains, to allow data sharing *and* domain specific data collection.

Structuring Medical Narratives

In an attempt to support structured recording of medical narratives we have developed OpenSDE (SDE: Structured Data Entry). The goal of OpenSDE is to support structured data entry in a variety of settings, so as to have patient data available for both routine care and retro- and prospective research. Therefore, OpenSDE is designed to accommodate the structured recording of data in settings where content and order of data entry often cannot be predicted. With OpenSDE we intend to support clinicians in such a way that separate data collection for research, alongside the regular data collection for patient care, is no longer necessary. We, therefore, developed OpenSDE with the aim of providing seamless integration of data collection for patient care and research purposes.

OpenSDE has its roots in ORCA (Open Record for CAre) [16]. Since 1996 the structured data entry module has been separated from ORCA as a stand-alone application. This SDE-application underwent many changes in the subsequent years. Since March 2003 the SDE-application is available in open source as OpenSDE [17].

The aim of this research project is to investigate the fea-

sibility of using data recorded with OpenSDE, for research purposes. Consistency and accuracy of collected data are pivotal for research, and are especially challenging if data will be collected over long periods of time and by different users. *This research, therefore, focuses on pitfalls for data extraction for research purposes, and aims to formulate strategies to improve uniformity in data entry to enhance the reliability of data retrieval.*

The work described in this thesis can be divided into two parts. In the first part we focus on supporting data entry, storage, and retrieval. In the second part of the thesis we concentrate on uniformity of data representation in OpenSDE.

Part 1: Supporting Data Entry, Storage, and Retrieval

In *Chapter 2* we focus on the goal of OpenSDE as well as on the requirements of achieving flexibility and expressiveness in data entry. To provide insight into which data can be recorded we describe both data entry options as well as data constraints that can be enforced during modeling.

The goal of OpenSDE is to support SDE in a variety of settings and to support potential benefits of structured data such as research and data sharing. To meet these goals, the data entry application should allow tailoring to specific medical domains and individual preferences without the need for technical adaptation [18]. In *Chapter 3* we describe the storage method that we chose to apply for storing the structured data recorded with OpenSDE. We illustrate how the recorded data are represented and identify the differences between our storage method and similar storage approaches.

Knowing which data can be recorded and how they are consequently represented, the next step is to support extraction of data for research purposes. *Chapter 4* focuses on the possibility of extracting data recorded with OpenSDE and representing the extracted data in a manner suitable for research purposes. The data recorded in OpenSDE are conceptually hierarchical, whereas the researcher, typically, will use conventional relational tables. In this chapter we describe the tool developed to support data extraction and conversion from the hierarchical format to a data set that can be used for further analysis.

Part 2: Uniformity of Data Representation in OpenSDE

Data entry involves interpreting and consequently translating observations into a predefined structured format. Ideally, recording data using structured data entry leads to uniformly structured data. In *Chapter 5* we investigate the uniformity of recorded data when OpenSDE is used to transcribe data from the same source. In OpenSDE we respected the clinicians' need for flexibility and expressiveness, i.e. with certain degrees of freedom, to describe findings. Freedom in data entry, however, implies that the same data may be recorded differently by different clinicians [19]. For purposes such as research and decision support, on the other hand, a structured, uniform representation of the same data set is essential.

The results of the study described in chapter 5 showed that recording data using structured data entry does not necessarily lead to uniformly structured data. Consequently, *Chapter 6* focuses on the origin of differences in representation of semantically identical information. Our main focus is the impact of expressiveness and flexibility on uniform data representation. We investigate the repercussion of initial design decisions and propose measures to improve uniformity in data entry.

The proposed measures for improving uniformity are evaluated in *Chapter 7* in which we describe a second study that investigates uniformity of data transcribed from the same source.

We conclude this thesis with a summary of the lessons learned.

References

1. Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(11):593-600.

2. Grover FL, Shroyer AL. Clinical science research. J Thorac Cardiovasc Surg 2000;119(4 Pt 2):S11-21.

3. Powsner SM, Wyatt JC, Wright P. Opportunities for and challenges of computerisation. Lancet 1998;352(9140):1617-22.

4. Dick RS, Steen EB, Detmer DE, eds. The computer-based patient record: an essential technology for health care. Revised Edition ed. Washington: National Academy Press; 1997.

5. van Ginneken AM. The computerized patient record: balancing effort and benefit. Int J Med Inf 2002;65(2):97-119.

6. Safran C. Using routinely collected data for clinical research. Stat Med 1991;10(4):559-64.

7. McDonald CJ. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc 1997;4(3):213-21.

8. Mainous AG, 3rd, Hueston WJ. Using other people's data: the ins and outs of secondary data analysis. Fam Med 1997;29(8):568-71.

9. Tange H. Medical narratives in the electronic medical record - Towards a searching structure with optimal granularity [PhD Thesis]. Maastricht: University of Maastricht; 1997.

10. Tange HJ, Hasman A, de Vries Robbe PF, Schouten HC. Medical narratives in electronic medical records. Int J Med Inf 1997;46(1):7-29.

11. Tange H. How to approach the structuring of the medical record? Towards a model for flexible access to free text medical data. Int J Biomed Comput 1996;42(1-2):27-34.

12. Worster A, Haines T. Advanced statistics: understanding medical record review (MRR) studies. Acad Emerg Med 2004;11(2):187-92.

13. van der Lei J. Closing the loop between clinical practice, research, and education: the potential of electronic patient records. Methods Inf Med 2002;41(1):51-4.

14. McDonald CJ, Tierney WM. Computer-stored medical records. Their future role in medical practice. Jama 1988;259(23):3433-40.

15. Moorman PW. Towards formal medical reporting. An evaluation in endoscopy. Rotterdam: Erasmus University; 1995.

16. van Ginneken AM, Moorman PW. Self-contained patient data in ORCA to cope with an evolving vocabulary. Proc AMIA Symp 1998:190-4.

17. OpenSDE. OpenSDE (OSS). http://sourceforge.net/projects/opensde. Last accessed: August 24, 2005.

18. van Ginneken AM, de Wilde M. A New Approach to Structured Data Entry. In: Waegeman CP, editor. TEPR 2000; 2000 May 8-11 2000; San Francisco, Ca.; 2000. p. 627-35.

19. Lovis C, Baud RH, Planche P. Power of expression in the electronic patient record: structured data or narrative text? Int J Med Inform 2000;58-59:101-10.

2 Recording Data with OpenSDE

Published as:

"OpenSDE: a Strategy for Expressive and Flexible Structured Data Entry"

Renske K. Los, Astrid M. van Ginneken, Johan van der Lei

In:
International Journal of Medical Informatics 2005; 74(6):481-490

2

Abstract

Purpose: This description focuses on the expressiveness and flexibility of OpenSDE: an application that supports recording of structured narrative data.

Methods: OpenSDE enables data entry with (customizable) forms based on trees of medical concepts. The relevant scope for data entry can be tailored per medical domain by construction of a domain-specific tree. OpenSDE is intended for structuring narrative data to make these available for both care and research.

Results: The OpenSDE application is currently in use at several departments in our academic hospital, including radiology, neurology, pediatrics, and child psychiatry. OpenSDE is available for all in open source.

Introduction

Electronic patient data are associated with many potential benefits, e.g. data sharing, decision support, quality assessment, research, and management of patient care [1-4]. The degree to which patient data are currently available electronically varies. To harvest the potential benefits of electronic data, the data must also be structured to enable processing by computer applications [2]. Structuring the medical narrative poses a significant challenge: content and level of detail are often unpredictable and vary per domain (and even per clinician) [5]. In an attempt to structure medical narratives in a manner that allows for variation and unpredictability, we have developed OpenSDE (SDE: structured data entry). OpenSDE is an application that supports clinicians with the recording of structured data for use in both care and research [6, 7]; data that are till now typically recorded in free text narratives.

Other published work on support of SDE does not provide much insight in the functionality and expressiveness of the respective applications. Therefore, in our description of OpenSDE we focus on those aspects that enable flexibility and expressiveness in data recording. Since OpenSDE is based on the selection of predefined concepts, we also explain why we did not choose to directly adhere to an existing terminology standard. OpenSDE is available in open source [6].

Medical Narratives

The medical narrative can be found in diverse sections in the medical record: medical history, family medical history, physical examination, progress notes, and reports (e.g., radiology, surgery, or pathology reports) [8]. Medical narrative data tend to be unruly [5], and only predictable to a certain degree. Free text has been the ideal format to collect these data as it has a high degree of expressiveness [9]. Free text allows clinicians to record data in whatever words, abbreviations, or codes desired. The challenge is to structure the medical narrative in a manner that poses no a priori limitations on detail and that structures the data in a manner also suitable for research. An additional aspect that must be addressed for data sharing purposes (and

multi-disciplinary research, for example) is uniformity in data representation. An application that attempts to tackle the challenge must be generic, yet tailorable to specific domains, to allow data sharing *and* domain specific data collection.

OpenSDE: goal and perspective

The goal of OpenSDE is to support structured data entry in a variety of settings, so as to have patient data available for both routine care and retro- and prospective research[1]. This implies that OpenSDE intends to support two goals that have diverging requirements for data format and level of detail. For care, data are preferably entered as free text, whilst, for research purposes, one prefers coded data.

When developing an application like OpenSDE, one must choose a perspective from which to approach the problem [9]. The direct benefit of structured data lies primarily with the research component. However, one is dependent on the clinicians for data collection, and to motivate them to structure data, benefits, such as validity checks or report generation, must be added. Data collection for research as a separate activity from data collection for patient care would be an undesirable expansion of the clinician's task. We, therefore, developed OpenSDE from the perspective of care aiming to provide seamless integration of data collection for research.

[1]OpenSDE has its roots in ORCA (Open Record for CAre) [10]. Since 1996 the structured data entry module has been separated from ORCA as a stand-alone application. This SDE-application underwent many changes in the subsequent years. Since March 2003 the SDE-application is available in open source as OpenSDE [6].

OpenSDE: data entry

The expressiveness provided by the OpenSDE application must not pose (a priori) limits on the level of detail in which one wants to structure data. Not only should data entry be highly expressive, it should also be straightforward. OpenSDE, therefore, applies the following principle for structured data entry. Data can be entered about predefined concepts. These concepts are organized as nodes in a tree structure (we refer to this as a domain model). In this tree, every node is described by its sub tree, as shown in Figure 1. In general, the deeper one navigates into the tree the more detailed a sign or symptom can be described; the tree also holds constraints relevant for the presentation of data entry options. The essence of data entry with OpenSDE is traversing the tree of medical concepts and selecting those nodes that correspond with the medical observations. The tree is domain specific; the modeling of trees and tree characteristics is discussed in the paragraphs about domain models, further on in this paper.

Entry Forms

Figure 1 shows a screen capture of the OpenSDE data entry application. The left-hand side shows the tree that contains the predefined medical concepts. In this example, 'history of present illness' is the selected concept. The right-hand side of the screen illustrates a (standard) form for the concept 'history of present illness'. This form contains all concepts in the sub tree of 'history of present illness', i.e., all concepts that are relevant to describe in the context of the history of the current illness. If, however, the sub tree of a concept is more than three levels deep, the form becomes too large to oversee. Therefore, after the third level in depth, we use hyperlinks to subdivide a form into more detailed forms. At the bottom of the form in Figure 1, a hyperlink is presented for the concept 'patient uses anticoagulants…'.

The example in Figure 1 shows entered data for history of present illness in its corresponding form and in the tree.

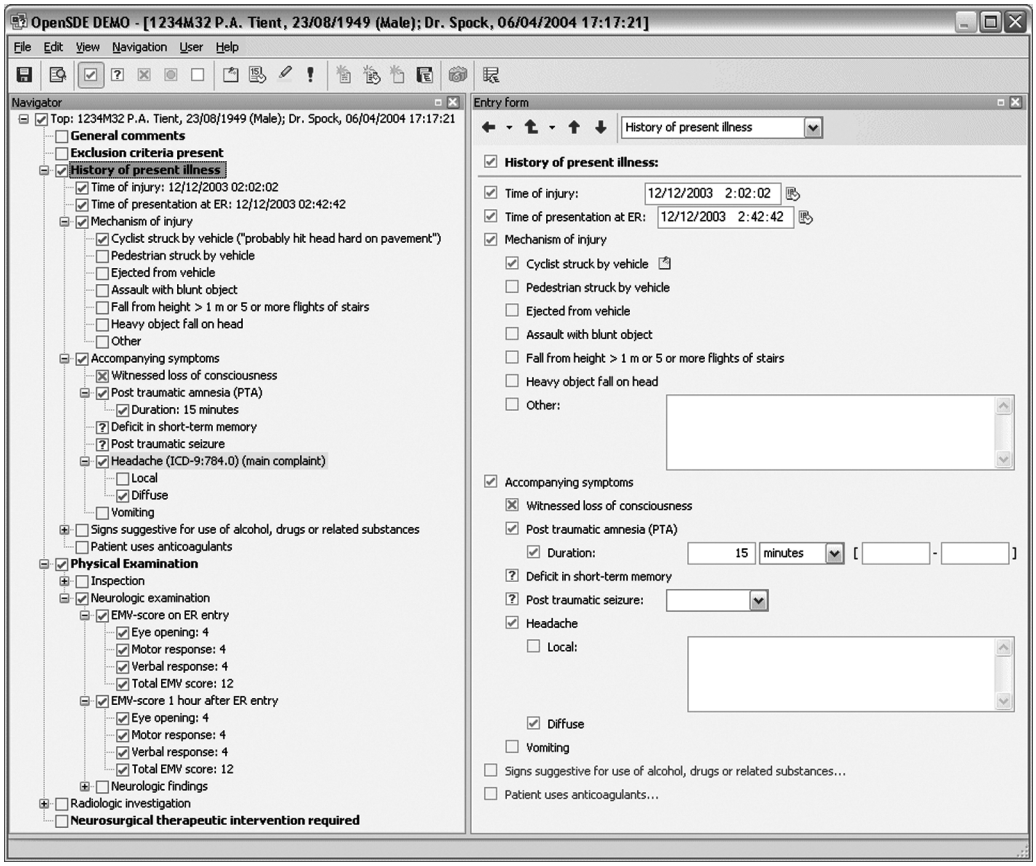

Figure 1. Screen capture of OpenSDE. The left side of the screen shows the domain model tree, which contains medical concepts. On the right is the form on which data are entered. This form is associated with the selected node, in this case: 'History of present illness'.

As one navigates through the tree on the left, the forms will change accordingly. The form always corresponds with the selected concept in the tree, and is generated by the application, based on the concepts in the tree. Making changes to the tree does not require manual adaptation of the standard forms.

Users can create custom forms that contain the medical concepts relevant to a particular scenario. In a custom form

for a particular concept, the user can select a sub set of the nodes in the sub tree of that concept, and determine the order in which the selected nodes appear on the custom form. A custom-made form for, e.g., a diabetes check-up may be defined to contain such concepts as blood pressure, weight gain and loss, eyesight, sensibility, and other relevant information. Clinicians can define the forms to accommodate specific topics and their individual preferences, which enhances the flexibility for data entry. Custom forms can be made using a form editor, which is a tool inside the OpenSDE application.

Expressiveness in Data Entry

We will use the example provided in Figure 1 to illustrate the kind of expressiveness that forms the basis for OpenSDE.

A clinician admitting to the ER a patient with a trauma to the head caused by a blunt object, will need to record data relevant to the scenario. Relevant data may include the mechanism and time of injury; any accompanying symptoms such as headache or vomiting; loss of consciousness recalled by patient, bystander, or companion; and findings from the physical examination and radiologic investigation.

In the example above, mechanism and time of injury are descriptors of history of present illness in the sense that they describe the injury in more detail. The accompanying symptoms, such as headache and vomiting can be present or absent. Furthermore, time of injury requires the recording of a date/time value, whereas the duration of headache requires a numeric value with a unit.

In OpenSDE we support these examples of expressiveness in a generic way. Besides the ordering of medical concepts as nodes in a tree, each node has a set of data items to specify, for example, presence state (absent, present, or unknown), timestamp, and value. The presence state is entered in the check box in front of the medical concept, as can be seen on the form in Figure 1. A check is present, a cross is absent, a question mark means unknown, and an empty check box implies that no data have been recorded about the concept.

An additional data item to enhance expressiveness is the main complaint which enables the clinician to label vomiting as 'main complaint' as this was the reason for encounter. It may also be necessary to describe the progression of a complaint over time, e.g. headaches have become less frequent. Besides that, a complaint or symptom may manifest itself differently in different circumstances. Headache, for example, may be local in the morning and diffuse in the afternoon. In other situations it may be necessary to record distinct data about the left and the right ear: bleeding is present in the right ear, but absent in the left. To structure such data we have enabled the user to duplicate particular sub trees to allow the recording of the chronology, different manifestations, and multiple occurrences of an observation; resulting in an array of sub trees of the same type. We refer to these data sub trees as progress descriptions, multiple descriptions, and multiple instances, respectively. The actual storage of the data sub trees and all other data recorded with OpenSDE is described elsewhere [11].

When clinicians feel the need to record data that cannot be represented by any of the data items offered in OpenSDE, they can add comments in free text.

Data Templates

In general, data entry occurs using the forms displayed on the right-hand side of the screen in Figure 1. Another option for data entry is data templates. Users can create templates to contain personalized predefined values. One may enter data typical for 'CT-scan head: normal', and save this typical data as a template. Whenever the radiologist encounters another patient with a normal head CT scan (i.e. showing no abnormalities), he can select the template 'CT-scan head: normal'. OpenSDE then copies the data, as defined in the template, as actual patient data. The clinician can then adjust these data to fit the case at hand. The use of templates is not without risk, and our advice to clinicians is, therefore, to limit use of templates to highly standard situations.

Domain Models

The OpenSDE application is generic in the sense that it can be tailored to multiple specialisms without the need for changes to the database and software. The data entry procedure with OpenSDE is the same for every user, regardless of the specialism. The only difference is that the content varies per specialism.

OpenSDE uses domain models, which are trees of medical concepts purposely developed for the application. A specific domain model is created per medical discipline. A cardiology domain model, for example, contains all the relevant concepts at the necessary level of detail for the cardiologist to record his medical narrative data. In general, these models do not contain knowledge needed for inference, such as 'fracture affects bone', 'a skull is a bone', therefore, a skull can have a fracture. The aim of domain models is to define the concepts and constraints that are relevant to record the medical narrative.

Domain models vary in content from each other but not in terms of structure, i.e., the model for cardiology will vary in content from the model for pediatrics but the representation (structure) of the content remains the same. The domain model should, therefore, be seen as consisting of a content and a structure. The content refers to the medical concepts that can be selected during data entry to create medically relevant expressions, whereas the structure refers to the tree format in which these concepts are represented.

Domain models are manually authored.

Domain Model Structure: Trees

The domain models are represented as a rooted tree structure. A rooted tree consists of nodes and arcs that connect these nodes, and has one root node. Every node, except the root, has one parent node, while every parent node may have one or more child nodes. A node without children is called a leaf. For every node, one path extends from the root to the particular node.

Domain Model Content: Medical Concepts

The developers intended the OpenSDE application to be used for the recording of medical narrative data. The content of the domain model for patient contacts, although tailored per specialism, will generally contain the sections: patient history, family history, review of systems, and physical examination. Every section contains elements that are more specific: the deeper one navigates into the tree, the greater the level of detail. The hierarchical character of domain models reflects the nature of medical descriptions.

Figure 2. This is an example of a domain model as seen by the modeler. The tree contains concepts (black words) in a hierarchical organization. The node types and properties are shown in grey. The concepts in bold face are Core Entities.

In Figure 2, the concept 'Neurologic findings' is described by concepts as 'lateralising motor weakness', 'lateralising sensory disturbances', and 'focal neurological deficit'. Lateralising motor weakness, can be either 'left' or 'right'. The path that leads from the root to a node indicates the context in which that node should be interpreted. In the case

of Figure 2, 'left' belongs to 'lateralising motor weakness', but 'left' can also be used in the context of 'lateralising sensory disturbances'.

Domain Models: Data Constraints

Creating domain models (modeling is described in the Section 'Creating Domain Models') not only encompasses defining and ordering the concepts about which data can be recorded, it also includes defining the constraints on the data. Constraints include the type of information that can be entered about a concept (presence states, numerical or free text values, etc). Constraints also include the limits and restrictions on the data (plausible values for systolic blood pressure must be in between 90 and 200, and the systolic pressure must be higher than the diastolic pressure).

Constraints are added to nodes by giving the node a 'node type' and by assigning appropriate properties to each node. A node can be one of four types: feature, option, unit, or shortcut. The properties that can be set depend on the node type of a node.

The node type 'feature' represents a characteristic that cannot be entered as absent; blood pressure and weight are features since a person always has a blood pressure and a weight. The presence state of a feature is, therefore, either present or unknown, and never absent. In the case of features, OpenSDE only accepts presence state 'present' in combination with a further description.

A concept receives the node type 'option' when it is an optional item of data: something that can be entered either as present or absent (or unknown), such as 'headache'. The main difference between a feature and an option is that the presence state of a feature cannot be 'absent' whereas for an option it can be.

Figure 2 presents an extract of the domain model used for the radiology/neurology study. The concept 'post traumatic amnesia' (PTA) can be described by 'duration', which has been modeled as child of PTA. The node PTA has the node type 'option', as it is something that is not necessarily present in a patient. The node 'duration' has been modeled as 'feature' node type. This characteristic is always applicable when a person has suffered from PTA; if there is PTA, it always has a duration.

The 'unit' node type is always a child node of a node with a numeric value property, and indicates the possible unit(s) of the value that must be entered. Figure 2 illustrates that the units of 'duration' in this example, are seconds, minutes, or hours. A unit node can be further specified by 'default unit' and 'unit factor'. In Figure 2, the default unit is set to 'minutes'. The unit factor enables calculations between the default unit and the other possible units.

Sometimes a finding is relevant in more than one medical context, i.e., should be offered for description in more than one place in the tree. Edema of the extremities, for example, may be relevant in the context of cardiovascular, renal, or endocrine disorders. It is, however, not desirable to describe the same finding in more than one branch in the tree. Instead of describing the same edema in more than one place, only one of the edema nodes contains all relevant describing child concepts, and the other edema nodes become references to this node. We refer to this reference as a 'shortcut'. The node type 'shortcut' is conceptually different from the node types 'feature' and 'option'. The feature and option node types represent certain constraints on the data that can be recorded about a node. Shortcuts were added for the convenience of data entry, and to prevent the same concept from being described twice in a structured manner in a tree.

The node type determines which properties can be assigned to the nodes. The two most frequently occurring node types in the domain models are 'feature' and 'option'. The properties that can be set for features and options are listed in Table 1, together with a brief explanation of the implications of these properties for data entry.

Table 1. Properties applicable to features and options. + Signifies that the property is applicable (i.e., it can be set by the modeler); + 'By default' means the property is set by default; NA stands for property not applicable.

Property	Effects on data entry	Feature	Option
Core Entity	Represents the main entity of interest in one particular path	+	+
Description mandatory	The presence state of at least one of the concept's children must be entered, or a comment must be added	+ By default	+
Description requires evidence	If concept is described further, at least one of its children must be present, or a comment must be added	+	+
Multiple instance	Applies when multiple occurrences of a concept can be described (e.g., fingers, warts)	NA	+
Multiple description	Allows more than one description to be added to describe the concept in different circumstances	+	+
Value	A value is any one of the following:	+	+
Numeric	Concept is a numeric value of the type: single numeric value; value that lies within a range in the form of x-y; or value has a margin x ± y	+	NA
Calculated field	A numeric value may contain a calculation based on values of other nodes. This enables the calculation of scores (e.g., APGAR or GCS)	+	NA
Free text	Allows entry of free text data	+	+
Moment	Date or date-time value	+	NA
One child present only	Only one of the child nodes may be 'present' (for mutually exclusive children)	+	+
Condition	Data must conform to a specific condition (systolic pressure> diastolic pressure)	+	+
Picture	A picture can be added to illustrate a specific concept	+	+
Codes	Concept to which a classification code, for example an ICD-10 code, can be added	+	+

There are two properties that require more explanation: core entity and codes. In every path, one node is assigned the core entity property. This identifies it as the main node of interest in this path. As mentioned in the paragraph on expressiveness in data entry, we enable the user to duplicate particular sub trees to allow the recording of the chronology, different manifestations, and multiple occurrences of an observation. This duplication of sub trees is only allowed at the level of

core entities and, if more than one type of duplication applies, in a predefined order. This limitation to the expressiveness was introduced to increase predictability. If these sub trees are allowed anywhere in the tree and can be nested in any order, the way in which data will be recorded becomes highly unpredictable, which for our purposes is undesirable.

The codes property allows a code to be assigned to a particular node. This enables a link to a classification or terminological system. If desired, codes can be shown in the OpenSDE interface (see Figure 1). Codes are described in more detail in the discussion of this paper.

Creating Domain Models

Domain models are created by experts in particular fields of medicine, using a specifically designed tool (Domain Model Editor), which creates a visual representation of the concepts as they are ordered in the tree structure. Modeling is described in more detail elsewhere by Doupi and van Ginneken [12]. When creating a new domain model, the domain model editor will display only one node: the top (or root) node. The domain model is expanded by first adding a child node to the top node. New nodes are then added to this node as siblings, meaning that they are on the same level (as 'history of present illness' and 'physical examination' are in Figure 2), or as children of the node (in Figure 2, 'local' is a child node of 'headache').

When adding a new node to the tree, the node must be assigned a node type and the applicable properties, as described above.

Discussion

Since OpenSDE domain models are trees of predefined concepts, domain models intuitively resemble a terminology. Therefore, we often receive the question what the difference is between domain models and a terminology, or why we did not use a terminological system instead of our own, manually authored, domain models.

Standardization is essential for the aggregation and pooling of data for clinical research, as well as for the sharing of data between applications that need to process the data [13]. Data structure is important for research and decision support. With the OpenSDE application, we want to enable data collection for research purposes, as well as enabling the use of the application for data collection during routine medical practice. We want to support the clinicians with SDE, but standardization currently poses restrictions that are difficult to adhere to when the goal includes SDE for patient care.

Instead of creating domain models specifically designed for OpenSDE, we could have decided to use a coding scheme or a terminological system. Although this could perhaps have facilitated data interpretation and exchange, we purposely chose not to commit ourselves to a terminological system because of the following three aspects:

1 There is an essential difference between the goal of a terminological system and the goal of OpenSDE domain models.
2 Terminological systems have less granularity than the requirements that patient care poses on OpenSDE domain models, because terminology standards usually support granularity at levels appropriate for aggregation of data.
3 Standards are rigid, which is of course part of their purpose.

The first reason for not committing to a terminological system is the difference in goal between terminological systems and OpenSDE domain models. Terminological systems are mainly intended for semantic matching of medical concepts so as to enable data aggregation, exchange, and reasoning about concepts. For example, the aim of the GALEN Project was to construct a reference terminology in a formal representation that allows reasoning with general knowledge about what 'can be said' [14], as well as semantic matching involved in pooling of data. According to the National Library of Medicine, UMLS is intended to 'facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health,…, UMLS is not optimized for particular applications'

[15]. If we were to choose a subset of the UMLS we would currently not be able to fully meet the requirements for documentation of narratives. Using such tools would, therefore, still require much manual adaptation. OpenSDE is specifically intended to document the patient's signs and symptoms in detail. As a result, the information contained in OpenSDE domain models differs from the information contained in a terminological system.

Secondly, terminological systems have less granularity than the requirements that patient care poses on OpenSDE domain models, because terminology standards usually support granularity at levels appropriate for aggregation of data. Treatment decisions and care providing in general requires more detail than the terms used in many terminological systems [16]. Besides, few terminologies contain all relevant concepts to describe the medical narratives of *all* domains.

The last reason for not committing to a terminological system is that standards are rigid, which is of course part of their purpose. If one wants to support data entry for multiple specialisms and accommodate new insights and procedures in a flexible manner, it must be possible to add, change, or remove concepts (no longer) necessary for data collection. Altering a standard can take years, which in a practice setting is undesirable.

If a modeler insists on using a terminological system when creating an OpenSDE domain model, three strategies could be followed. The first would be to limit the concepts in the OpenSDE domain models to the relevant terms from a terminological system. As described above, this would be unsatisfying due to the limited level of granularity; data collection is limited to particular concepts which may not suffice for specific clinical research. A second option would be to choose one terminological system and expand it with those concepts, details, and relations needed for OpenSDE in a care setting. This not only requires adding new terms or a greater level of detail, it may also require expanding the formalism of the terminological system with new relations or concepts types that are needed for supporting structured data entry [17]. However, due to the rigidity of standards, this option is far from ideal. A third option is to create the domain models with concepts to suit those that will use them and, where relevant, associate these concepts with concepts of a terminological system, for example, by using codes. It will not, however, be possible to map all concepts from the domain models to one particular terminological system, unless a termino-

logical system is created that contains all concepts necessary for describing the medical narrative in a particular domain. For OpenSDE we chose the third option as this gives the modeler the freedom to map concepts to relevant concepts from a specific terminological system and to combine these with other concepts to create a domain model suitable for his purpose.

OpenSDE and its relation to standards proposed by CEN and HL 7 has been addressed elsewhere [17].

As mentioned, OpenSDE is designed and intended to support structured data entry in a variety of settings, in such a manner that no redundant recording is required to have patient data available for both routine care and retro- and prospective research. OpenSDE is not the only development that aims at enabling the above. However, documentation of expressiveness and functionality of similar SDE applications is scarce or outdated, making it difficult to obtain a good overview of state-of-the-art SDE applications. From the available literature, it is difficult to distill whether or how systems such as UltraStar [18] or PenIvory [19] deal with different manifestations of complaints in different circumstances and the extent to which clinicians are free to choose the level of detail in which they structure their data. An opportunity to compare applications like Medcin [20], Pen & Pad [21], IMR-E [22], Pure MD [23], UltraStar [18], PenIvory [19], and Purkinje [24] in terms of expressiveness, functionality, and use of standards may be very useful to the SDE community.

OpenSDE is currently being used by several departments at the Erasmus MC, including pediatrics, immunology, and child psychiatry, and is available in open source [6]. Alongside, five pilot studies in a clinical practice setting are being undertaken to evaluate OpenSDE in terms of: completeness of domain models, uniformity of data representation, and acceptance by end users. Depending on the outcomes of these studies, a decision will made about whether OpenSDE will be made available throughout the academic hospital. The pilot studies are in the domains of venerology, ear, nose and throat, pediatrics, liver disease and transplants, and anesthesiology. The study performed in conjunction with the departments of radiology and neurology started in February 2002 and ended in November 2004. Data have been recorded for over eighteen hundred patients. Both the collected data and the data format were suitable for evaluating the decision rule under investigation in the study [25].

Conclusion

Approaching structured data entry from the care perspective places emphasis on approaching the expressiveness of free text. We chose this perspective because we wanted to ensure that data collection corresponds as much as possible to the needs of the clinicians who are actually recording the data. Having spent effort on enabling data entry in a manner that suits clinicians, the next step is to approach the challenge from the perspective of research. Is it possible to use data that are unpredictable and potentially diverse? How can the hierarchically organized data be extracted for scientific analysis? Such questions are addressed in the paper entitled: "Extracting Data Recorded with OpenSDE: Possibilities and Limitations" [26].

Acknowledgements

The authors would like to thank Cobus van Wyk, Georgio Mosis, and Jan Talmon for their helpful comments on previous versions of this manuscript.

The work presented in this paper is funded by a grant from ZonMW.

References

1. Grover FL, Shroyer AL. Clinical Science Research. J Thorac Cardiovasc Surg 2000;119(4 Pt 2):S11-21.

2. Powsner SM, Wyatt JC, Wright P. Opportunities for and Challenges of Computerisation. Lancet 1998;352(9140):1617-22.

3. Dick RS, Steen EB, Detmer DE, eds. The Computer-Based Patient Record: An Essential Technology for Health Care. Revised Edition ed. Washington: National Academy Press; 1997.

4. van Ginneken AM. The Computerized Patient Record: Balancing Effort and Benefit. Int J Med Inf 2002;65(2):97-119.

5. Tange H. How to Approach the Structuring of the Medical Record? Towards a Model for Flexible Access to Free Text Medical Data. Int J Biomed Comput 1996;42(1-2):27-34.

6. OpenSDE. Opensde (Oss). http://webserver.mi.fgg.eur.nl/opensde/. Last accessed: March 31, 2005.

7. van Ginneken AM, de Wilde M. A New Approach to Structured Data Entry. In: Waegeman CP, editor. TEPR 2000; 2000 May 8-11 2000; San Francisco, Ca.; 2000. p. 627-35.

8. Tange HJ, Hasman A, de Vries Robbe PF, Schouten HC. Medical Narratives in Electronic Medical Records. Int J Med Inf 1997;46(1):7-29.

9. van der Lei J. Closing the Loop between Clinical Practice, Research, and Education: The Potential of Electronic Patient Records. Methods Inf Med 2002;41(1):51-4.

10. van Ginneken AM, Moorman PW. Self-Contained Patient Data in Orca to Cope with an Evolving Vocabulary. Proc AMIA Symp 1998:190-4.

11. Los RK, van Ginneken AM, de Wilde M, van der Lei J. Opensde: Row Modeling Applied to Generic Structured Data Entry. J Am Med Inform Assoc 2004;11(2):162-65.

12. Doupi P, van Ginneken AM. Structured Physical Examination Data: A Modeling Challenge. Medinfo 2001;10(Pt 1):614-8.

13. Cimino JJ. From Data to Knowledge through Concept-Oriented Terminologies: Experience with the Medical Entities Dictionary. J Am Med Inform Assoc 2000;7(3):288-97.

14. Rector AL, Nowlan WA. The Galen Project. Comput Methods Programs Biomed 1994;45(1-2):75-8.

15. NLM. http://www.nlm.nih.gov/research/umls/about_umls.html. Unified Medical Language System. Last accessed: April 6, 2005.

16. Cimino JJ. Review Paper: Coding Systems in Health Care. Methods Inf Med 1996;35(4-5):273-84.

17. van Ginneken AM. Considerations for the Representation of Meta-Data for the Support of Structured Data Entry. Methods Inf Med 2003;42(3):226-35.

18. Bell DS, Greenes RA. Evaluation of Ultrastar: Performance of a Collaborative Structured Data Entry System. Proc Annu Symp Comput Appl Med Care 1994:216-22.

19. Poon AD, Fagan LM. Pen-Ivory: The Design and Evaluation of a Pen-Based Computer System for Structured Data Entry. Proc Annu Symp Comput Appl Med Care 1994:447-51.

20. MedicompSystems. Medcin. www.medicomp.com. Last accessed: March 31, 2005.

21. Nowlan WA, Rector AL, Kay S, Goble CA, Horan B, Howkins TJ, et al. Pen & Pad: A Doctor's Workstation with Intelligent Data Entry and Summaries. In: Miller RA, editor. 14th SCAMC; 1990; Los Alamitos, California: IEEE Computer Society Press; 1990. p. 941-42.

22. Trace D, Naeymi-Rad F, Haines D, Robert JJ, deSouza Almeida F, Carmony L, et al. Intelligent Medical Record—Entry (Imr-E). J Med Syst 1993;17(3-4):139-51.

23. Lussier YA, Maksud M, Desruisseaux B, Yale PP, St-Arneault R. Puremd: A Computerized Patient Record Software for Direct Data Entry by Physicians Using a Keyboard-Free Pen-Based Portable Computer. Proc Annu Symp Comput Appl Med Care 1992:261-4.

24. Purkinje. Purkinje Dossier. www.purkinje.com. Last accessed: March 31, 2005.

25. Smits M, Dippel DWJ, De Haan GG, Tanghe HJG, Hunink MGM. Indications for Ct in Patients with Minor Head Injury: Generalization of a Published Decision Rule. In: RSNA 2003.; 2003; Chicago; 2003.

26. Los RK, van Ginneken AM, van der Lei J. Extracting Data Recorded with Opensde: Possibilities and Limitations. Int J Med Inf 2005;Accepted for publication.

3 Storing Data Recorded with OpenSDE

Published as:

"OpenSDE: Row Modeling Applied to Generic Structured Data Entry"

Renske K. Los, Astrid M. van Ginneken, Marcel de Wilde,
Johan van der Lei

In:
Journal of the American Medical Informatics Association 2004;
11(2):162-165

3

Abstract

Clinicians generally record medical narrative data, such as current complaints, physical examination, and progress notes, as free text in paper-based medical records. The medical narrative involves heterogeneous and detailed data that includes the description of (multiple) occurrences of medical findings or symptoms that may progress over time. Structured, electronic recording of narrative data would facilitate the use of these data for research. Our OpenSDE application supports clinicians with the structured recording of narrative data in both a research and care setting. Data entry is enabled using forms that are generated using domain specific trees of medical concepts. For data storage we have expanded the traditional row modeling methodology with additional columns that allow structured representation of medical narratives including descriptions of findings, multiple occurrences of findings, and the progression of findings over time.

Introduction

The medical narrative section of the patient record comprises the medical history, physical examination, progress notes, and reports on additional tests and interventions. Medical narrative data vary per discipline, per patient, and over time. Besides the heterogeneity of the data, the level of detail in recording varies greatly amongst clinicians. The unruliness and large variation in the collected data have made it difficult to support structured recording of the medical narrative [1]. Clinicians convinced of the potential benefit of electronically available data (e.g. greater availability, data sharing, data analysis, or use of decision support) have launched efforts to develop dedicated systems to accommodate their data needs. Such attempts are far from ideal [2]; over time, adaptation and expansion of databases results in haphazard collections of tables and data. New tables will make older tables (partially) obsolete, and data redundancy is frequent. Performing research on one or more of such databases is on the verge of being (un)manageable especially for clinicians or researchers who are relatively unfamiliar with database management [2].

Our objective is to support structured recording of narrative data in the form of an application that allows tailoring to specific medical domains and individual preferences without the need for technical adaptation [3]. Furthermore, we want to support structured recording of data with a high degree of expressiveness. We developed an application called OpenSDE [4] (SDE: Structured Data Entry) that supports structured data entry in a variety of settings, thus facilitating the use of data for both care and research. OpenSDE supports data entry using customizable entry forms based on domain specific trees. In this paper we will describe how we implemented row modeling to enable structured recording of medical narrative data.

Row Modeling

Row modeling is a methodology that is suitable for storing heterogeneous and evolving data sets [5]. In essence, row modeling involves a column-to-row transformation: the attributes (or column headings) of the conventional column-modeled table are stored as data in the row-modeled table.

A column-modeled table contains a column for every attribute. A row-modeled table contains *one* column that holds all attributes and one column that holds the values of the attributes. In a column-modeled table, one record holds a set of facts about a patient, whereas in a row-modeled table, every record holds one particular fact about a patient [6]. A row-modeled table only holds those attributes for which a value has actually been recorded.

In row modeling, the data definition is not defined in the data tables themselves. The data definitions are stored separately and are often referred to as "metadata". The advantage of separating the metadata from the physical data schema is that one eliminates the need to change the physical data structure when the data set changes: only the metadata content needs change. In a conventional column-modeled approach metadata are held in table definitions and relations between tables. Changes to a column-modeled table would involve adding or removing columns from tables, i.e. changing the database structure.

Row modeling can be used as a generic structuring technique for diverse and changing data sets. Metadata hold the information necessary for the correct semantic interpretation of the data held in the row-modeled table. Metadata, therefore, need to be edited and adapted for different disciplines, and constitute an important area of research [7].

Method

In OpenSDE metadata are represented as discipline-specific domain models. The domain model defines the content of the medical narrative in a specific discipline. Domain models vary in content but not in structure. The content consists of concepts and constraints organized in a rooted tree structure. The nodes of the tree structure represent the concepts and are connected to each other via one-directional arcs: a node at the end of an arc represents a descriptor of the node at the beginning of an arc. For every node, one path extends from the root to the particular node.

We developed a toolset that uses a graphical interface to define domain models; using this toolset, clinicians can

define their own domain models [8].

OpenSDE uses the domain models to generate an interface for data entry. Figure 1 is a screen capture of OpenSDE. The domain model tree (metadata) is presented on the left of the figure, whilst the right shows the dynamically generated entry form with all nodes detailing the node selected in the domain model. The forms can be customized by clinicians themselves.

To accommodate expressiveness for the recording of medical narratives, OpenSDE supports a number of general items that can be recorded for each concept in the domain model. Every instance of a concept has a 'presence state' which states whether a concept is present, absent or unknown. Numerical values can be a single value (with a deviation), a range, or a date/time value; each value has a unit. Domain models, however, have their boundaries: clinicians may encounter narrative that cannot be expressed using the domain model. To deal with this limitation of the domain model, clinicians may add free text to any node in the tree, i.e. each recorded finding may be supplemented by free text.

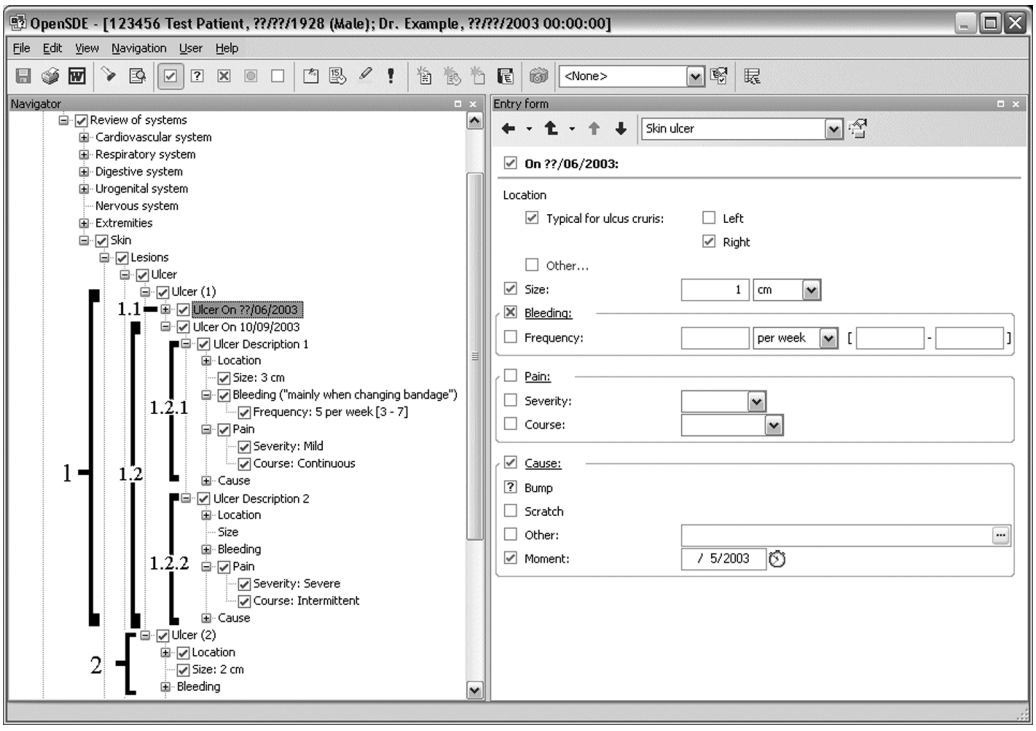

Figure 1. Screen capture of the OpenSDE data entry application. The left-hand side shows the domain model tree, which contains medical concepts. On the right is the form on which data are entered. The form is associated with the selected node, in this case 'skin ulcer'. The brackets on the left (included in this figure as example) indicate that two different ulcers are described: ulcer 1 on the right shin (see entry form) and ulcer 2 on the left shin (location is hidden in this view). Ulcer 1 consists of two descriptions over time (progress descriptions); the first description (1.1; shown on entry form) is of June 2003, describing the probable cause of the ulcer in May 2003; progress description 2 (bracket 1.2) shows the progression of the ulcer on September 10, 2003. Progress description 2 contains two descriptions of pain to indicate that pain is continuously mild (Ulcer Description 1/bracket 1.2.1) and intermittently severe (Ulcer Description 2/bracket 1.2.2).

OpenSDE uses an extended row-modeled table to support the complexity of the medical narrative. The example shown in Figure 1 illustrates that complexity: the patient reports that he has several skin ulcers; one of the ulcers is located on the right shin and the other on the left shin. The ulcer on the right shin was possibly caused by bumping into a table several months earlier; in the past few weeks this skin ulcer has grown, started to bleed, and is increasingly painful. In

OpenSDE, the row-modeled table has been extended with columns for *multiple instances*, *progress descriptions* and *multiple descriptions*. *Multiple instances* represent findings that can occur more than once (in Figure 1, the patient describes two skin ulcers: one on the left shin and one on the right shin). *Progress descriptions* represent findings that evolve over time (in Figure 1, the patient describes that as of September 10th, 2003 the skin ulcer on the right shin has started bleeding, mainly when the bandage is changed). *Multiple descriptions* represent findings that present themselves differently under different circumstances (in Figure 1 the patient complains that the ulcer is always a little painful, but that the pain is sometimes severe).

The data presented in Figure 1 are represented in Table 1. Every concept for which data have been entered (both in the tree and on the form in Figure 1) corresponds to one record in Table 1.

Table 1. The table represents an excerpt from the row-modeled table that we use to store data collected using OpenSDE. The first row contains the column headings. The following 31 rows contain patient data. The first 'key' column is shortened for this example, it normally consist of a reference to the patient, the event, and the domain model version and discipline. The column 'Node' is actually a code but for this example we have used the associated text. 'Node' refers to the node in the domain model associated with the recorded data. The following 11 columns are the data items. PresSt= Presence state (1= present, 2= absent, 3= unknown). The columns that include 'val' are used to represent the values (primary value, min, max, and margin) and unitId refers to the unit of the value. The 'comment' column holds free text values, and 'DateTime' refers to date applicable, i.e., data entry date unless otherwise specified by clinician. The last three columns are index columns: MIIx for multiple instances, PDIx for progress descriptions, and DIx for description index. The brackets at the right side of the table correspond to the brackets in Figure 1.

Key	Node	PresSt	PrimVal	MinVal	MaxVal	ValMarg	UnitId	Comment	DateTime	MIIx	PDIx	DIx	
4	ulcer	1							20031027	1	0	0	
4	ulcer	1							200306??	1	1	0	
4	location	1							200306??	1	1	0	
4	typ.ulcus cruris	1							200306??	1	1	0	
4	right	1							200306??	1	1	0	
4	size	1	1				cm		200306??	1	1	0	1.1
4	bleeding	2							200306??	1	1	0	
4	cause	1							200306??	1	1	0	
4	bump	3							200306??	1	1	0	
4	moment	1							200305??	1	1	0	
4	ulcer	1							20030910	1	2	0	
4	ulcer	1							20030910	1	2	1	
4	size	1	3				cm		20030910	1	2	1	
4	bleeding	1						mainly when changing bandage	20030910	1	2	1	
4	frequency	1	5	3	7		week		20030910	1	2	1	1.2.1
4	pain	1							20030910	1	2	1	
4	severity	1							20030910	1	2	1	
4	mild	1							20030910	1	2	1	
4	course	1							20030910	1	2	1	
4	continuous	1							20030910	1	2	1	
4	ulcer	1							20030910	1	2	2	
4	pain	1							20030910	1	2	2	
4	severity	1							20030910	1	2	2	1.2.2
4	severe	1							20030910	1	2	2	
4	course	1							20030910	1	2	2	
4	intermittent	1							20030910	1	2	2	
4	ulcer	1							20031027	2	0	0	
4	location	1							20031027	2	0	0	
4	typ.ulcus cruris	1							20031027	2	0	0	2
4	left	1							20031027	2	0	0	
4	size	1	2				cm		20031027	2	0	0	

Discussion

Row modeling is a technique frequently used for representing heterogeneous data sets. In a row-modeled table, every record ideally holds one particular fact about a patient [6]. Although applying the same underlying principle, different researchers have developed alternative approaches. Salgado et al. [9] use a combination of conventional and row-modeled tables for their clinical-trials information system COATI. Their approach was to create a row-modeled table per separate entity for those entities that are either trial specific or have attributes that vary between trials. Nadkarni et al. use an entity-attribute-value model with classes and relationships (EAV/CR) for the Human Brain Project [10] and clinical trials data. In addition, many researchers (e.g. [6, 11]) have separate tables for each data type; a change, for example, in data type from free text to a numeric value implies that from then on the attribute will be stored in a different table. This relocation of attributes is not necessary when hybrid data types are allowed in one column. In general, the use of multiple tables requires a decision about where to store which data, which implies the possible need for changes to the data structure when the data set changes. In OpenSDE all items are stored in a single table. That is, in OpenSDE we use an extended row-modeled table to hold extra data items in pre-assigned columns rather than introducing new tables. A row in our row-modeled table, therefore, corresponds to one fact about a patient but allows more detail about this fact to be described in one row.

A difference between the extended tables in Friedman's model [12] and OpenSDE is that Friedman represents context of data using nested rows, i.e., internal row reference. OpenSDE represents the context of each row with a reference to a unique node in the domain model.

The extensions we made to the row model fall in two categories. The first category deals with data types. Other researchers introduce different tables to deal with different data types, OpenSDE extends the row model with additional columns to reflect the data type. The second category deals with the complexity of the medical narrative (e.g., repeated descriptions over time of multiple lesions). OpenSDE extends the row model to represent descriptions of (multiple) occurrences of findings or symptoms that may progress over time.

OpenSDE does not model an ontology. At first sight, modeling an ontology in, for example, Protégé may seem similar to domain modeling in OpenSDE. Protégé, however, supports modeling for various purposes, such as decision support and data entry [13]. OpenSDE domain models are currently only used to support structured data entry; to use the domain models for inference would require adding more knowledge to our domain models. Investigating whether the expressiveness of OpenSDE can be achieved using Protégé, would be an interesting study.

OpenSDE is currently being used in several pilot projects within the Erasmus MC University Medical Center and is used by several commercial vendors of hospital information systems. OpenSDE is used in different disciplines including neurology, radiology, immunology and pediatrics. OpenSDE, written in Delphi, is available in open source [4].

References

1. Tange HJ, Hasman A, de Vries Robbe PF, Schouten HC. Medical Narratives in Electronic Medical Records. Int J Med Inf 1997;46(1):7-29.

2. Pierik FH, van Ginneken AM, Timmers T, Stam H, Weber RF. Restructuring Routinely Collected Patient Data: Orca Applied to Andrology. Methods Inf Med 1997;36(3):184-90.

3. van Ginneken AM, de Wilde M. A New Approach to Structured Data Entry. In: Waegeman CP, editor. TEPR 2000; 2000 May 8-11 2000; San Francisco, Ca.; 2000. p. 627-35.

4. OpenSDE. Opensde (Oss). http://webserver.mi.fgg.eur.nl/opensde/. Last accessed: March 31, 2005.

5. Nadkarni PM. http://ycmi.med.yale.edu/nadkarni/db_course/ (Select "Patient records"). Last accessed: April 5, 2005.

6. Nadkarni PM, Brandt C. Data Extraction and Ad Hoc Query of an Entity-Attribute-Value Database. J Am Med Inform Assoc 1998;5(6):511-27.

7. van Ginneken AM. Considerations for the Representation of Meta-Data for the Support of Structured Data Entry. Methods Inf Med 2003;42(3):226-35.

8. Doupi P, van Ginneken AM. Structured Physical Examination Data: A Modeling Challenge. Medinfo 2001;10(Pt 1):614-8.

9. Salgado NC, Gouveia-Oliveira A. Towards a Common Framework for Clinical Trials Information Systems. Proc AMIA Symp 2000:754-8.

10. Miller PL, Nadkarni P, Singer M, Marenco L, Hines M, Shepherd G. Integration of Multidisciplinary Sensory Data: A Pilot Model of the Human Brain Project Approach. J Am Med Inform Assoc 2001;8(1):34-48.

11. Ganslandt T, Mueller M, Krieglstein CF, Senninger N, Prokosch HU. A Flexible Repository for Clinical Trial Data Based on an Entity-Attribute-Value Model. Proc AMIA Symp 1999:1064-67.

12. Friedman C, Hripcsak G, Johnson SB, Cimino JJ, Clayton PD. A Generalized Relational Schema for an Integrated Clinical Patient Database. In: Proc 14th Symp Comput App Med Care.; 1990: IEEE Computer Society Press; 1990. p. 335-9.

13. Musen MA. Modern Architectures for Intelligent Systems: Reusable Ontologies and Problem-Solving Methods. Proc AMIA Symp 1998:46-52.

14. The Protégé website. http://protege.stanford.edu/ Last accessed: 19-11-2003.

3

4 Extracting Data Recorded with OpenSDE

Published as:

"Extracting Data Recorded with OpenSDE: Possibilities and Limitations"

Renske K. Los, Astrid M. van Ginneken, Johan van der Lei

In:
International Journal of Medical Informatics 2005; 74(6):473-480

4

Abstract

Purpose: OpenSDE is an application intended to support structured data entry in a variety of settings, such as routine care and clinical research. The past years development has focused on data entry to support expressiveness and flexibility. The focus is now shifting to data extraction: what are the possibilities for extracting the data and does the adopted strategy pose limitations?

Methods: Data extraction is supported by presenting the concepts for extraction in the same tree structure as for data entry. Users can select all or a sub selection of these concepts for extraction. Selected concepts are extracted and converted to a table format that can be queried using conventional tools.

Results: The extraction tool (Entity Export) provides a successful technical solution for data extraction. Using the extracted data, however, leads to obstacles that are a result of a fundamental design principle of OpenSDE.

Introduction

A medical record, whether paper-based or electronic, contains information recorded for patient care. The potential value of these data extends well beyond the use of data for patient care and includes use of the data for research or management [1-4]. The promise of electronic medical records is that data once stored in the context of care are readily available for secondary use [5]. However, even in electronic records, patient data are often still stored as free text or scanned documents. As a result, researchers still have to perform the labor-intensive task of reading and interpreting free text in individual electronic medical records.

In recent years, researchers have developed software that supports the recording of structured data [6-12]. We developed OpenSDE (SDE: Structured Data Entry): a data entry application designed to support structured recording of narrative data [13, 14]. The application is domain independent: OpenSDE can be applied for any domain and tailored to individual data collection needs and preferences. The goal of OpenSDE is to support structured data entry in a variety of settings, so as to have patient data available for both routine care and retro- and prospective research.

The data recorded in OpenSDE is conceptually hierarchical, whereas the researcher, typically, will use conventional relational tables. For researchers, the data collected in OpenSDE, therefore, need to be transformed to a data set that can be used for further analysis. In this paper, we describe the tool we developed for extracting data for research from the data recorded with OpenSDE. We first briefly describe the model underlying OpenSDE, we subsequently discuss the extraction tool and provide examples. We finally discuss the limitations of the tools we developed.

OpenSDE

The principle of OpenSDE is that clinicians can traverse a tree of medical concepts and select those concepts that correspond with the relevant medical observations. In this tree structure (or "domain model"), the nodes represent medical concepts and the path from the top of the tree to a particular node represents the context of a node.

For every node in the domain model, the OpenSDE application generates an entry form, as shown in Figure 1 [15]. For the concepts presented on the entry forms, users may indicate whether or not the concept applies (present, absent, or unknown) or, when relevant, record a specific (numerical, temporal, or free text) value. There are many more aspects concerned with data entry; these are described in more detail elsewhere [16].

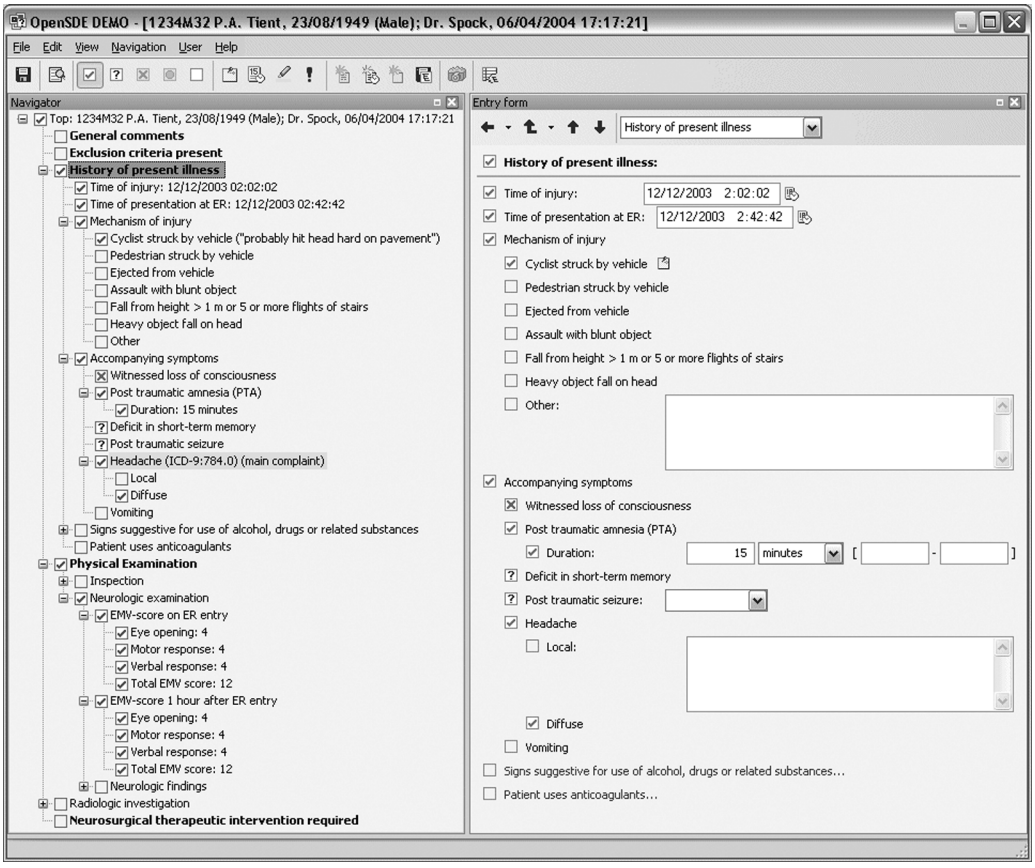

Figure 1. Screen capture of OpenSDE. The left side of the screen shows the domain model tree, which contains medical concepts. On the right is the form on which data are entered. This form is associated with the selected node, in this case: 'History of present illness'.

The strength of OpenSDE is its generic design and the resulting flexibility: OpenSDE allows tailoring of a domain model to specific medical content, and the content coverage can be expanded and altered without the need for technical adaptation of the software or the physical data structure [17]. The domain models are customized to the degree of expressiveness required by users during data collection.

Row modeling in OpenSDE

OpenSDE was developed from the perspective of data collection. The focus was, therefore, on supporting clinicians with flexible and expressive data collection. To accomplish flexibility as well as to enable the application to be generic, i.e. domain independent, OpenSDE uses row-modeling instead of conventional relational tables for data storage. In conventional relational tables the semantics of the data are held in the tables themselves, as well as in the relations between the tables in a database. In general, there is a direct mapping between the user interface and the attributes in a table. A change in content coverage will require a change in the database structure (e.g., addition of new tables or columns in tables), as well as a change in application software and user interface. Ideally, changing the content coverage does not require changing the database structure or the (interface of the) application software [17]. To enable this, we chose to apply row modeling for the storage of the data recorded with OpenSDE [18].

Row modeling is a methodology that is suitable for storing heterogeneous and evolving data sets [19]. In essence, row modeling involves a column-to-row transformation: the attributes (or column headings) of the conventional column-modeled table are stored as data in the row-modeled table. As a result of this transformation the table structure itself no longer reveals the semantic information needed for data interpretation and interface generation [17]. The semantics must be explicitly specified either with or without internal row referencing. *With* internal row referencing each row holds a reference to its parent, and each parent holds a reference to its parent to retain the hierarchy needed for representation of context. Retrieving the entire context of a concept thus requires a complex procedure of recursive queries [17]. *Without* internal row referencing, on the other hand, context representation involves separating the context from the data and defining the context as metadata. Every row in the row-modeled table holds a reference to the metadata which represents the unique context of a particular concept. In her paper on considerations for the representation of metadata, van Ginneken discusses representation of semantics in more detail [17].

For OpenSDE we chose to represent semantic information without internal row referencing. The patient data entered in OpenSDE are stored in a row-modeled table. The row-

modeled table holds only those nodes for which data were recorded, and every row contains a reference to a unique node and its context in the domain model. Changes to the domain model, i.e. changes to the context, lead to a new version of the domain model with new references for each node. Data recorded with a new version of the domain model will refer to these new nodes, whilst data recorded with the old domain models retain the references to the version with which they were recorded, ensuring correctness of context over versions. Changes to the domain model have no impact on the table that stores the patient data, and the same database structure, application software, and interface can be used for many medical domains. The obvious disadvantage is that data are represented in a format that differs from the conventional relational format accepted by most data analysis software.

Extracting row-modeled data

Row modeling is a technique frequently used for large scientific databases [20] that hold data that will be queried. However, querying row-modeled data is less straightforward than querying conventionally represented data, because of the separation of data and context. Simple operations such as AND or OR queries involve many self-joins to the same table [20]. Querying is complicated for researchers as the representation of the row-modeled data does not match their conceptual perception of the data. Querying, therefore, requires support.

In general, there are two approaches to querying row-modeled data. The first approach is to build a tool that supports querying of row-modeled data. Often, the goal of such tools is to create the illusion for the researcher that he is querying a conventional relational database. The tool then translates the conventional queries to a format suitable for querying row-modeled data. Although suitable for basic ad hoc queries, this approach requires extensive programming and addition of metadata to support complex statistical analysis of data [21]. A second approach is to convert the row-modeled data to a conventional relational format suitable for querying with conventional analysis software. Exporting data has as advantages that analysis on data can

be performed with available tools often known by researchers, and does not require development of functionality that already exists. The challenge for this approach is to reintegrate the semantic information with the actual data. Nadkarni et al. [22] have created a tool which supports both ad-hoc (run-time) querying as well as extraction of data for analysis. The best approach is dependent on the intended use of the data: for ad-hoc, simple queries a dedicated tool is perhaps the best option, whereas for extensive statistical analysis, data conversion is preferred.

Extracting data recorded with OpenSDE

The intention of OpenSDE is to support clinical research involving statistical analyses; we, therefore, chose to export the data to a conventional format to enable data analysis by conventional analysis software.

An important goal for data extraction is transparency of semantics, i.e. that the semantics of the data, as intended during data entry, remain clear. Often, data extraction is performed by researchers or data managers who did not record the data themselves. The main challenge is to develop a transparent method that permits selection of data in the same context as data entry, and that transforms conceptually hierarchical, row-modeled data to a conventional format without losing the important contextual information held in the hierarchy.

The transparent method that we developed to select and transform the conceptually hierarchical data to a conventional format is realized as an application called Entity Export. Entity Export supports selection of concepts from a domain model and converts the corresponding data to corresponding columns in one or more newly created conventional tables suitable for analysis purposes. The original row-modeled table remains intact; Entity Export duplicates the data for output in conventional relational tables (see also Example of Entity Export Use).

One of the properties of OpenSDE domain models, is that in every path one node has been assigned as the principal node of interest in this path. This node is assigned during the modeling process, and is labeled as the 'core entity'. The con-

cepts leading to the core entity represent the context of the core entity and the concepts below this core entity are all descriptors of the core entity itself. Complex descriptors of core entities may involve sub trees. In Figure 1, for example, "history of present illness" is a core entity, and "accompanying symptoms" is a detailed sub tree of this core entity. Core entities and other node properties are described in more detail elsewhere [23].

Core entities represent a natural level for grouping data into one table. Every core entity becomes a table; data pertaining to all concepts below the core entity in the domain model are then presented in the table for that core entity. Every column in a table represents a data item associated with a node in the sub tree of the core entity (see also Example of Entity Export Use).

The user of Entity Export first selects the domain model that was used for recording the data. Entity Export then displays that domain model in exactly the same manner in which OpenSDE displays it for data entry (Figure 2). From this tree the user selects the medical concepts of interest. Certain concepts are further described by more detailed concepts (data items); for example, for measurements the date of measurement, the date of recording, the unit of measurement, and the actual value are available. After selecting the concepts for export, the user can tailor the data to be exported, for example, for a measurement only the value and unit are to be exported.

Once selection has been completed, the data are exported to a database and can be queried with the appropriate tools (e.g. SQL).

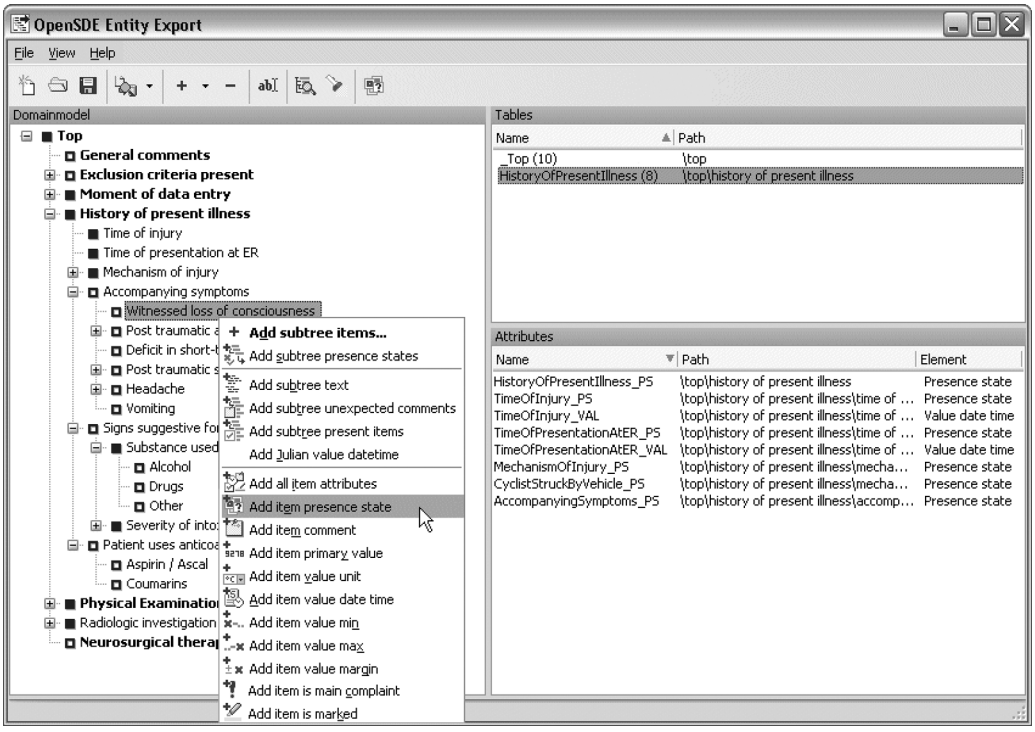

Figure 2. Screen capture of Entity Export. The domain model is presented on the left side, with a pop-up menu showing the optional data items associated with a node. The tables made for export are shown at the top right of the screen. The bottom right shows the attributes selected for the table 'History of present illness', alongside the paths (context) of the attributes. These attributes can be associated with any node in the sub trees of the selected core entity.

Example of Entity Export Use

The Entity Export tool has currently been tested in a few domains such as pediatrics, immunology and the combined (sub-) domains of neurology and radiology. For now we will focus on the latter setting. The Departments of Neurology and Radiology at the Erasmus MC are working together in the CHIP study (Computer tomography of Head Injured Patients). In this study data are collected on patients who

received a CT scan because they were submitted to the Emergency Room with trauma to the head caused by blunt objects. The purpose of the study is to evaluate criteria by which one can assess whether or not a patient (with a head injury caused by a blunt object) must receive a CT scan [24]. Data collection started in February 2002 and ended in November 2004. Data have been collected on over eighteen hundred patients.

Although the main goal of the study is to evaluate the clinical decision rule, the Department of Medical Informatics is involved in the data collection and extraction process, as data are collected with OpenSDE. The informatics component of the research involves investigating the OpenSDE application and the Entity Export tool in a clinical research-based setting.

The data collected for the CHIP study were exported for analysis purposes. The export resulted in seven exported tables, one for each core entity. The tables were analyzed for completeness of data. The conventional format of the exported table made it very clear which data were omitted during data entry; this was very difficult to oversee in the original row-modeled table. The conventional tables were exported to SPSS for statistical analysis. Figure 3 shows an excerpt of the exported data recorded for the core entity 'History of Present Illness' and all the nodes below this core entity.

Data analysis involved investigating aspects such as mean age, gender distribution, mean Glasgow Coma Score, and percentage of patients that suffered from loss of consciousness. The interface of Entity Export and the format of the extracted data posed no problems in the query process. All analyses needed to evaluate the clinical decision rule were successfully performed on the extracted data [25].

HistoryOfPresentIllness : Table

PatientId	EventId	TreeNodeTimeS	HistoryOfPresen	TimeOfInjury_PS	TimeOfInjury_VAL	TimeOfPresenta	TimeOfPresentatio	MechanismOfInjury	CyclistStruckByVehicle
Patient1	20030111311195	20030111311195	1			1	13/1/03 11:20:10	1	
Patient1002	20030319195 10	20030319195 10	1	1	19/3/03 18:45:00	1	19/3/03 19:15:00	1	
Patient1003	20030320090 13	20030320090 13	1	1	13/3/03 11:00:00	1	13/3/03 11:30:00	1	
Patient1004	20030320090 44	20030320090 44	1	1	13/3/03 14:00:00	1	13/3/03 14:30:00	1	
Patient1005	20030320091 55	20030320091 55	1	1	6/3/03 14:00:00	1	6/3/03 14:30:00	1	

Figure 3. Screen capture of the exported table for the core entity 'History of Present Illness'. The first three columns represent references to the patient, contact moment and the moment at which a sign, symptom or observation apply. The following columns represent the data for the extracted concepts.

Discussion

The first step in enabling extraction of data recorded with OpenSDE, was to ensure that it was technically possible. Our focus, therefore, was on the conversion of the hierarchical data stored in row-modeled tables to a conventional format suitable for querying using conventional tools. Entity Export, the tool that we developed to enable this conversion, was tested using the CHIP study data set. The conversion of the hierarchical data to conventional tables per core entity succeeds without problems. In that regard, Entity Export is successful.

However, using the extracted data leads to obstacles that are a result of a fundamental design principle of OpenSDE. In the design of OpenSDE we purposefully chose not to infer data beyond what is actually recorded by the clinicians. This implies that OpenSDE does not make inferences or reason about data that were not explicitly recorded. Inference and reasoning are left to the users of the application. The following two examples illustrate the type of problems that result from the design principles of OpenSDE and that users can encounter when querying data.

When querying, one may be interested, for example, in certain concepts which are absent. Figure 4 shows two examples which are extracts of the domain model presented in Figure 1. In the example on the left the user has recorded that there is no headache, but there are accompanying symptoms (in this case vomiting). The example on the right illustrates that there are no accompanying symptoms.

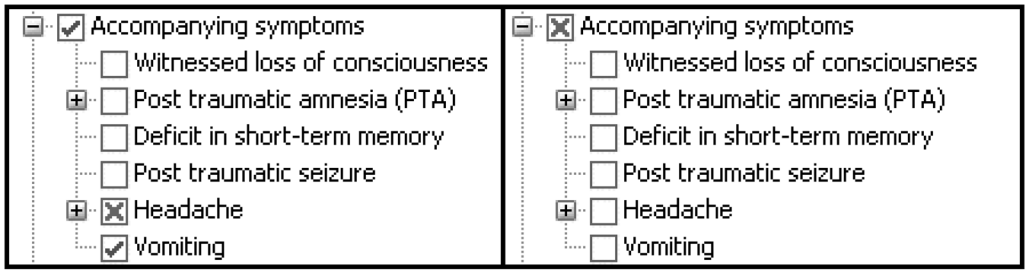

Figure 4. Two excerpts of data recorded with the CHIP study domain model. In the example on the left 'Accompanying symptoms' applies (indicated with a checkmark) and headache does not apply (indicated by cross). In the example on the right, there are no accompanying symptoms.

If one performs a query to select all patients that do not have a headache the answer will include all patients where headache has actually been recorded 'absent', as shown in the left example in Figure 4. However, for some patients as shown on the right in Figure 4, accompanying symptoms (which is one level above headache in the tree) will have been set to 'absent' as they do not have any accompanying symptoms. This implies that headache is also 'absent', but according to the fundamental design principles of OpenSDE, it is not explicitly stored as data. Therefore, these patients will not be extracted with the above query. The user of Entity Export must decide whether to include only those patients where headache is explicitly recorded as absent, or whether implicit information (obtained by explicit querying of this information) also needs to be included. The query items and the level at which a concept is queried in the tree must be carefully selected.

The necessity for insight into the context of data, as well as insight into the possibilities for data entry becomes obvious in situations in which nodes can be mutually exclusive. OpenSDE offers the possibility to define mutual exclusivity in the form of a property ("one child present only" see [23]), but this does not guarantee that a modeler has actually set this property. This can have consequences for querying. For example, a sub tree consists of a grandparent node (gp), a parent node (p), and three child nodes (node 1 - node 3). A researcher is interested in extracting those patients where child node 3 is absent. Depending on whether the child nodes are mutually exclusive or not, and whether the mutual exclusivity has been explicitly modeled or not, three situations can occur.

Figure 5 illustrates how these three varieties can be represented in OpenSDE.

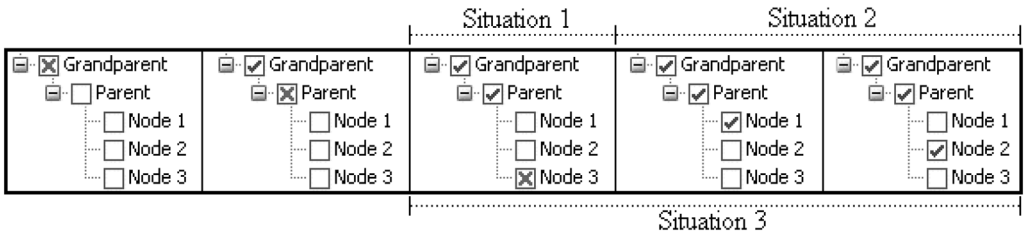

Figure 5. This figure is a graphical representation of data entry varieties that may occur when mutual exclusivity is not applicable (Situation 1), is applicable and modeled (Situation 2) and is applicable but *not* modeled (Situation 3). The first two rectangles apply to all three modeling alternatives. The third rectangle represents situation 1. The last two rectangles represent situation 2. The last three rectangles apply to situation 3.

Each situation requires a corresponding query (see below).

Situation 1:	**Query 1**:
Child nodes *are not*	NOT gp OR NOT p OR NOT node 3
mutually exclusive	
Situation 2:	**Query 2**:
Child nodes are mutually	NOT gp OR NOT p OR node 1 OR node 2
exclusive (modeled)	
Situation 3:	**Query 3**:
Child nodes are mutually	NOT gp OR NOT p OR node 1 OR node 2 OR NOT node 3
exclusive (NOT modeled)	

It is thus essential that insight is provided into both the context of data and the properties of nodes.

During data analysis it became apparent that the format of the recorded date/time values was not appropriate for calculations. The researchers were interested in the age of the patient, which was not explicitly recorded but had to be derived from the date of birth and the date of admission to the hospital. To enable calculations on these values we added a

Julian[1] date/time function to Entity Export. During the extraction process Julian date/time values are now added to our initial representation of date/time values.

We mentioned that the challenge in designing a tool for data extraction is to reveal the semantics of data, as intended during data entry, to the person performing the extraction. The examples described illustrate how essential this presentation of input semantics is. By presenting the domain model (i.e. the semantics) used for data entry, the user of Entity Export may envision how data may have been recorded. It is essential that the user is aware of both the possibilities for data entry and the implicit information to optimize the semantic coverage of his queries.

Another design principle of OpenSDE that has consequences for data recording, as we concluded from the exercise with the CHIP study, is the freedom the users have to record data. The data collected for the CHIP study were less complete than we had anticipated. Although the data to be recorded with OpenSDE consisted of a small set, clinicians did not record data for all nodes that they were expected to record data for. From this we can conclude that offering predefined options for data entry *alone* does not guarantee data completeness, unless, of course, completeness is explicitly enforced. To ensure data completeness, a reminder function such as a data checklist is available in OpenSDE and can be activated. Checklists influence the completeness of data by stimulating users to enter particular data items [26]. Checklists can be used to enforce data entry, but also as reminders that the user may ignore.

[1] Julian date/time values are a format in which one point in time is taken as a reference point and all dates are represented as a time period from that point.

Future Research

Entity Export has been tested with the straightforward and well-defined data set of the CHIP study. The results of Entity Export are promising: data extraction and representation in a conventional format is *technically* possible. The challenges that remain on a semantic level are our next focus.

OpenSDE is currently in use at the Erasmus MC Sophia pediatric hospital to collect routine patient data. Once a large and varied enough data set has been collected with OpenSDE, we will analyze the effects of the design principles of OpenSDE for data extraction and use of data for research purposes.

The goal of OpenSDE is to support structured data entry in a variety of settings, so as to have patient data available for both routine care and retro- and prospective research. Especially for research purposes a uniformly represented data set is highly preferable. Future research, therefore, focuses on investigating whether structure invites users to represent data uniformly.

Acknowledgements

The authors would like to thank Cobus van Wyk, Georgio Mosis, and Jan Talmon for their helpful comments on previous versions of this manuscript.

The work presented in this paper is funded by a grant from ZonMW.

References

1. Safran C. Using Routinely Collected Data for Clinical Research. Stat Med 1991;10(4):559-64.

2. McDonald CJ, Hui SL. The Analysis of Humongous Databases: Problems and Promises. Stat Med 1991;10(4):511-8.

3. Grover FL, Shroyer AL. Clinical Science Research. J Thorac Cardiovasc Surg 2000;119(4 Pt 2):S11-21.

4. Dick RS, Steen EB, Detmer DE, eds. The Computer-Based Patient Record: An Essential Technology for Health Care. Revised Edition ed. Washington: National Academy Press; 1997.

5. Mainous AG, 3rd, Hueston WJ. Using Other People's Data: The Ins and Outs of Secondary Data Analysis. Fam Med 1997;29(8):568-71.

6. Cheung NT, Fung V, Chow YY, Tung Y. Structured Data Entry of Clinical Information for Documentation and Data Collection. Medinfo 2001;10(Pt 1):609-13.

7. Webster C, Copenhaver J. Structured Data Entry in a Workflow-Enabled Electronic Patient Record. J Med Pract Manage 2001;17(3):157-61.

8. McCullagh PJ, McGuigan J, Fegan M, Lowe-Strong A. Structure Data Entry Using Graphical Input: Recording Symptoms for Multiple Sclerosis. Stud Health Technol Inform 2003;95:673-8.

9. Duftschmid G, Wrba T. A Tool for the Design of Clinical Forms Supporting End-User Integration. Med Inform Internet Med 2004;29(1):29-41.

10. Mansson J, Nilsson G, Bjorkelund C, Strender LE. Collection and Retrieval of Structured Clinical Data from Electronic Patient Records in General Practice. A First-Phase Study to Create a Health Care Database for Research and Quality Assessment. Scand J Prim Health Care 2004;22(1):6-10.

11. Lei J, Stetson PD, Chen ES, McKnight LK, Mendonca EA, Cimino JJ. Structured Data Entry of Cross-Coverage Notes Using a Pda. Medinfo 2004;2004(CD):1712.

12. Rosenbloom ST, Kiepek W, Belletti J, Adams P, Shuxteau K, Johnson KB, et al. Generating Complex Clinical Documents Using Structured Entry and Reporting. Medinfo 2004;2004:683-7.

13. van Ginneken AM, Stam H, van Mulligen EM, de Wilde M, van Mastrigt R, van Bemmel JH. Orca: The Versatile Cpr. Methods Inf Med 1999;38(4-5):332-8.

14. van Ginneken AM, de Wilde M. A New Approach to Structured Data Entry. In: Waegeman CP, editor. TEPR 2000; 2000 May 8-11 2000; San Francisco, Ca.; 2000. p. 627-35.

15. van Ginneken AM, Verkoijen MJ. A Multi-Disciplinary Approach to a User Interface for Structured Data Entry. Medinfo 2001;10(Pt 1):693-7.

16. Los RK, van Ginneken AM, van der Lei J. Opensde: A Strategy for Expressive and Flexible Structured Data Entry. Int J Med Inform 2005;74(6):481-90.

17. van Ginneken AM. Considerations for the Representation of Meta-Data for the Support of Structured Data Entry. Methods Inf Med 2003;42(3):226-35.

18. Los RK, van Ginneken AM, de Wilde M, van der Lei J. Opensde: Row Modeling Applied to Generic Structured Data Entry. J Am Med Inform Assoc 2004;11(2):162-65.

19. Nadkarni PM. http://ycmi.med.yale.edu/nadkarni/db_course/ (Select "Patient records"). Last accessed: April 5, 2005.

20. Nadkarni PM, Brandt C. Data Extraction and Ad Hoc Query of an Entity-Attribute-Value Database. J Am Med Inform Assoc 1998;5(6):511-27.

21. Deshpande AM, Brandt C, Nadkarni PM. Metadata-Driven Ad Hoc Query of Patient Data: Meeting the Needs of Clinical Studies. J Am Med Inform Assoc 2002;9(4):369-82.

22. Nadkarni PM. Qav: Querying Entity-Attribute-Value Metadata in a Biomedical Database. Comput Methods Programs Biomed 1997;53(2):93-103.

23. Los RK, van Ginneken AM, van der Lei J. Opensde: A Strategy for Expressive and Flexible Structured Data Entry. Accepted for publication in Int J Med Inf 2005.

24. Haydel MJ, Preston CA, Mills TJ, Luber S, Blaudeau E, DeBlieux PM. Indications for Computed Tomography in Patients with Minor Head Injury. N Engl J Med 2000;343(2):100-5.

25. Smits M, Dippel DWJ, De Haan GG, Tanghe HJG, Hunink MGM. Indications for Ct in Patients with Minor Head Injury: Generalization of a Published Decision Rule. In: RSNA 2003.; 2003; Chicago; 2003.

26. Wyatt J. Quantitative Evaluation of Clinical Software, Exemplified by Decision Support Systems. Int J Med Inf 1997;47(3):165-73.

4

5 Are Structured Data Structured Identically?

Investigating the uniformity of pediatric patient data recorded using OpenSDE

Renske K. Los, Jolt Roukema, Astrid M. van Ginneken,
Marcel de Wilde, Johan van der Lei

Published in:
Methods of Information in Medicine 2005; 44(5)631-8

5

Abstract

Objective: OpenSDE is an application that supports structured recording of narrative patient data to enable use of the data in both clinical practice and clinical research. Reliability and accuracy of collected data are essential for subsequent data use. In this study we analyze the uniformity of data entered with OpenSDE. Our objective is to obtain insight into the consensus and differences of recorded data.

Methods: Three pediatricians transcribed 20 paper patient records using OpenSDE. The transcribed records were compared and all recorded findings were classified into one of six categories of difference.

Results: Of all findings 22% were recorded identically; 17% of the findings were recorded differently (predominantly as free text); 61% was omitted, inferred, or in conflict with the paper record.

Conclusion: The results of this study show that recording patient data using structured data entry does not necessarily lead to uniformly structured data.

Introduction

Many potential advantages of electronic patient records (EPRs), such as availability of patient data for clinical research, decision support, or quality assessment [1, 2], require data to be represented in a structured manner [3, 4]. Structured Data Entry (SDE) is a method by which clinicians record patient data directly in a structured format. SDE involves predefined fields for data entry. Advantages of this approach are: data are structured at the source, without requiring intervention or correction rounds; data are more uniform; predefined entry fields may predispose users to record data in more detail; and SDE offers the possibility of enhancing the quality of data [5].

SDE remains challenging to apply for medical narratives, as data vary per domain, per patient, and over time [6-8]. The medical narrative comprises the medical history, physical examination, progress notes, and reports on additional tests and interventions [9]. The narrative is a combination of patient narrated and clinician-observed data.

Our objective is to support structured recording of narrative data (in multiple medical domains) to enable use of the data in both clinical practice and clinical research. Therefore, we developed OpenSDE [10] as an application that offers structured data entry in a variety of settings. OpenSDE supports data entry using customizable entry forms based on domain-specific trees. OpenSDE is available in open source [11].

Although OpenSDE supports structured data entry, suggesting that data are structured uniformly, the actual concordance in data representation has not yet been explored. Reliability and accuracy of collected data are pivotal if data will be collected over long periods of time and by different users [12, 13].Therefore, in collaboration with our hospital's pediatric department, we analyzed the uniformity of recorded data when OpenSDE is used to transcribe data from the same data source. Of interest in this qualitative analysis is whether recording data using OpenSDE by definition leads to uniformly structured data. Obtaining insight into the consensus and differences in data recorded with OpenSDE is particularly important when retrieving routinely collected data for clinical research purposes [14]. Uniformity in data entry facilitates data extraction and lookup: if the same data are recorded in different manners by different clinicians the

chance of finding the data (in a particular place in the record) becomes smaller. If, for example, one clinician records a penicillin allergy in a structured manner and another clinician records this as free text comment in patient history, both places must be checked to see if a patient is allergic to penicillin. This problem becomes even larger when data can be recorded as free text anywhere and one does not know in advance where to expect particular data. Data can easily be overlooked and the chance for duplicate data recording also increases. Look up may take more time and increase the workload on clinicians which can lead to a decrease in the quality of patient care and a lower success rate of the implementation of OpenSDE [15].

The purpose of this study is to provide qualitative insight into how data are recorded. It is important to understand how to format information to make data easier to find and clearer to interpret [16]. We need to understand if the current format that we offer clinicians to record data leads to uniformity. If OpenSDE invites users to record the same data in exactly the same manner, retrieval and look up will be more predictable and easier to do for the user. If OpenSDE does not lead to uniform data representation we need to investigate what differences occur and how these differences can be minimized.

Materials

OpenSDE

OpenSDE is an application for structured recording of narrative sections of the patient record. The principle of OpenSDE is that clinicians can traverse a tree of predefined medical concepts and select those concepts that correspond with the relevant medical observations. The content of such a tree is domain specific and we refer to the tree of medical concepts as a domain model. In this tree structure, the nodes represent medical concepts and the path from the top of the tree to a particular node represents the context of a node [10].

Clinicians can select a node in the tree, and the application will display a form associated with this node alongside the tree, as shown in Figure 1. Each form presents the selected concept and the corresponding descriptors (branching nodes) of the concept [17]. For the concepts presented on the entry forms, users may indicate whether or not the concept applies (present, absent, or unknown) or, when relevant, record a specific value (numerical, temporal or free text). Symptoms can be described more than once in the context of progression over time, different circumstances, or multiple occurrences. OpenSDE also supports the use of free text for particular details not covered by the content of the domain model. Users can create custom entry forms (using an integrated form editor) to suit their individual data entry preferences [10, 18].

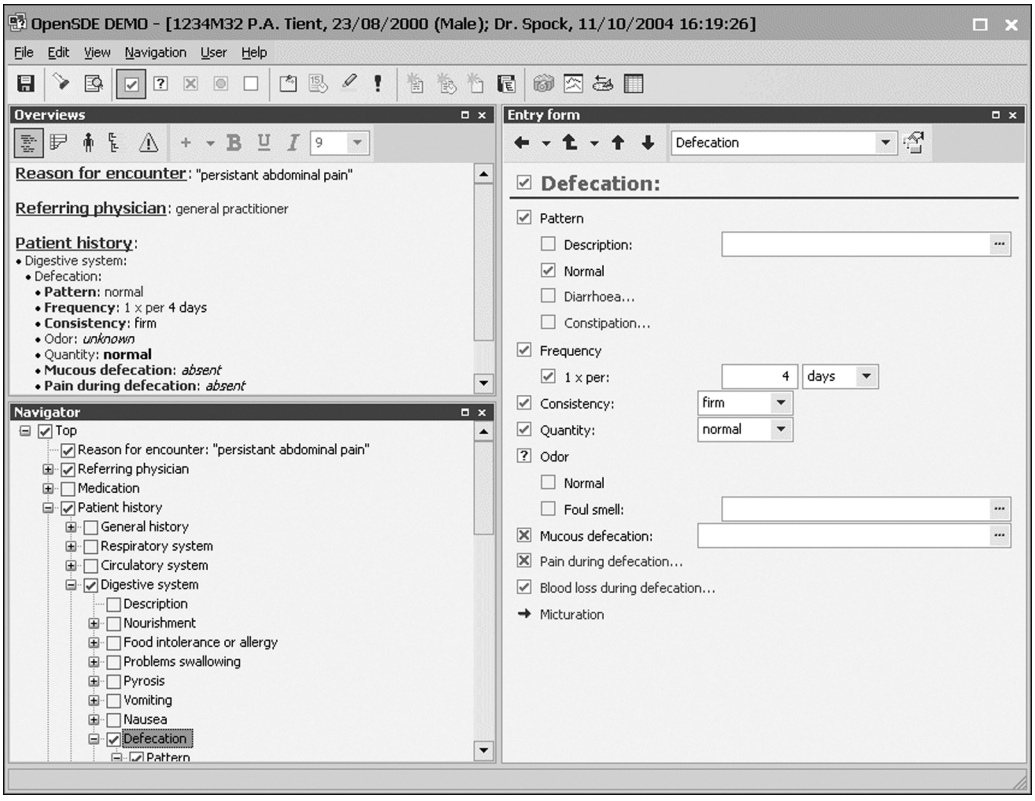

Figure 1. Screen capture of the OpenSDE data entry application. The top left of the screen shows an overview of the data recorded for the patient in the current session. The bottom left shows the domain model tree with medical concepts. On the right is the form on which data are entered. The form is associated with the selected node, in this case 'defecation'.

The pediatric domain model used in this study was created by modelers with a background in pediatrics (second author) and medical informatics (third author). Prior to this study, experienced OpenSDE users recorded data of over 100 pediatric paper records in OpenSDE to evaluate the ordering and coverage of the pediatric domain model. The model was then altered to improve both ordering and coverage, as well as to facilitate data entry [19].

Methods

Data entry from a common data source

At our pediatric outpatient department we recruited three pediatricians. Prior to this study, the experience of these pediatricians consisted of a standardized course on the use of OpenSDE in general pediatrics and documentation of ten first-contact patients in OpenSDE.

We randomly selected 20 handwritten paper patient records created for first-contact patients at the pediatric outpatient department. These records belonged to patients that were not under care of any one of the three pediatricians involved in this study. We chose first-contact patients as intake and physical examination data for these patients are fairly standardized. Although the patient data are recorded as free text in the paper records, the data are written on semi-structured forms that contain headings such as 'family medical history', 'birth history', 'allergies', and 'neurological examination' at which the corresponding medical findings can be recorded. Due to this 'structure' the paper records were comparable in format, degree of detail, and in amount of data content. We expected that data entry by three clinicians would provide good insight into the nature of differences in data representation.

The three pediatricians transcribed the 20 paper records in OpenSDE, creating a data set of 60 transcribed records. The pediatricians were informed about the goal of the study, and knew that the transcribed records would be analyzed.

Consensus and differences

Our main interest in this study is the consensus and differences in the representation of structured patient data. Therefore, we conducted detailed analyses of the transcribed records to identify the types of differences in the transcribed records. We identified six categories of consensus and differences. To classify the findings into one of the six categories we developed the algorithm described below and presented in Figure 2.

Per patient we created a list of all findings recorded in OpenSDE. For every finding we analyzed how it was represented by each of the three pediatricians. If the finding was represented in exactly the same manner in all three transcribed records, the finding was classified as *identical*. If there was a difference in at least one of the transcribed records, we searched through that entire transcribed record to establish whether the finding was recorded elsewhere. If the finding was represented in a structured manner elsewhere, the finding was classified as *structured differently*. If the finding was recorded as free text, at the same place or elsewhere in the tree, the finding was classified as *free text*.

For those cases where the finding was not represented in all three transcribed records (either identically, as free text, or structured differently) we consulted the paper record. If the finding was present in the paper record, we classified the finding as *omitted*. If, however, the description of the finding in the paper record conflicted with the description of the finding in the transcribed records, we classified the finding as *conflicting*. A last possibility is that (one of) the transcribed records contained a finding not present in the paper record. In this scenario the finding was classified as *inferred*. The classification process was repeated for all findings.

The findings were subsequently classified as normal findings (e.g. 'no cardiac murmur') or abnormal findings (e.g. 'constipation'), and were split into patient history and physical examination findings.

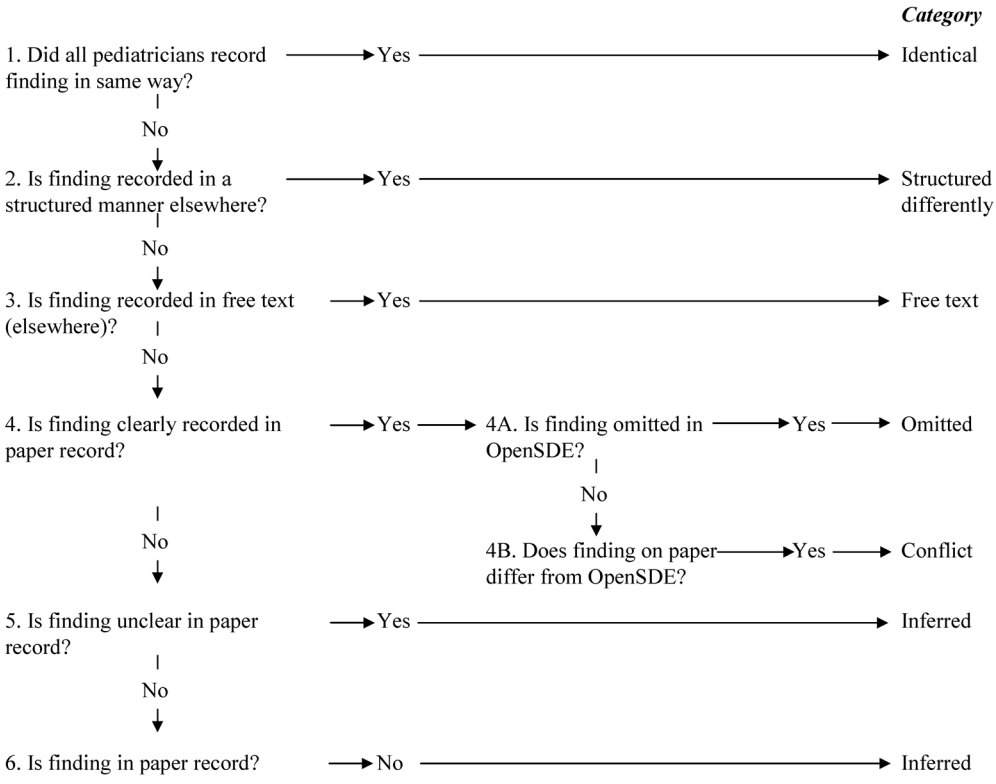

Figure 2. Algorithm used to categorize findings.

Results

The findings recorded in the transcribed records were divided into patient history and physical examination findings and then subdivided into normal or abnormal findings. These (sub) divisions are shown in Table 1: the rows hold the type of finding (normal or abnormal) and the columns represent the part of the record (patient history or physical examination). A total of 1764 findings were recorded for all 20 patients.

Of these findings, 495 are normal patient history findings and 867 normal physical examination findings (totaling 1362 normal findings). In total, 77.2 % of all findings are normal findings, which corresponds to a mean of 68 normal findings per patient (range 23-117).

Table 1. Normal and abnormal findings in the patient record
The results presented in this table represent the number of findings per part of the patient record (patient history or physical examination) and per type or finding (normal or abnormal). The percentage corresponds to the percentage of the total number of findings.

Part of patient record / Type of finding	Patient History	Physical Examination	Entire Record
Normal Findings	**495** (28.1%)	**867** (49.1%)	**1362** (77.2%)
Mean per transcribed record	24.8	43.4	68.1
Range	3-59	20-61	23-117
Abnormal Findings	**351** (19.9%)	**51** (2.9%)	**402** (22.8%)
Mean per transcribed record	17.6	2.6	20.1
Range	5-38	0-6	5-43
All Findings	**846** (48%)	**918** (52%)	**1764** (100%)
Mean per transcribed record	42.3	45.9	88.2
Range	8-88	20-63	28-151

In Table 2 we present the findings as classified per category of consensus/difference. All findings could all be classified into one of the six categories of consensus/difference according to the algorithm. The first row of the table shows that we encountered 90 normal and 79 abnormal patient history findings which were recorded identically for all three patients. For the physical examination, we counted 198 normal and 21 abnormal findings that were transcribed identically by all three pediatricians. Of all findings 22% (or 388 findings) were recorded identically by all three pediatricians. In total, 4.9% of all findings were structured differently, and in 12.2% of all findings one or two pediatricians recorded the findings as free text. Almost one third (31.1%) of all findings was inferred, and over one quarter (26.7%) of the findings was omitted by one or two pediatricians. A total of 55 findings (3.1%) were conflicting with the paper record.

Table 2. Consensus and differences in representation of findings
In this table the findings have been ordered per category of consensus/difference. Per category the findings are subdivided into patient history and physical examination, and normal and abnormal findings. The percentage behind the numbers corresponds to the percentage of the total number of findings.

Finding	Patient History		Physical Examination		ALL FINDINGS
Category	Normal	Abnormal	Normal	Abnormal	
Identical	90 (5.1%)	79 (4.5%)	198 (11.2%)	21 (1.2%)	**388** (22%)
Structured differently	28 (1.6%)	23 (1.3%)	32 (1.8%)	4 (0.2%)	**87** (4.9%)
Free text	68 (3.9%)	117 (6.6%)	26 (1.5%)	4 (0.2%)	**215** (12.2%)
Inferred	98 (5.6%)	30 (1.7%)	411 (23.3%)	9 (0.5%)	**548** (31.1%)
Omitted	188 (10.7%)	90 (5.1%)	184 (10.4%)	9 (0.5%)	**471** (26.7%)
Conflicting	23 (1.3%)	12 (0.7%)	16 (0.9%)	4 (0.2%)	**55** (3.1%)
ALL FINDINGS	***495*** *(28.1%)*	***351*** *(19.9%)*	***867*** *(49.1%)*	***51*** *(2.9%)*	***1764*** *(100%)*

Discussion

Structured data entry offers the possibility to improve the quality of data [5] and standardize data collection [20]. In this study we investigated the uniformity of recorded data when OpenSDE is used to transcribe data from a common data source. We analyzed 60 transcribed records, which in total covered 1764 findings. Our results show that only 22% of all findings were recorded identically by all three clinicians and in more than three quarters of the findings there was difference in data representation or data content.

Evaluation of data quality in medical records is a topic of ongoing interest in medical informatics. Evaluation meth-

ods and measurement means are not standardized and different studies focus on different aspects as different stakeholders pose different requirements on data quality [14, 21].

Evaluating the quality of the data recorded in OpenSDE is associated with one particular difficulty. Data quality, especially completeness and accuracy, can only be measured as a function of the question that the data set should answer [22]. Winthereik concludes that the goal should not be to produce data, which are accurate in and by themselves, but to produce data, which are pertinent to specific questions [23]. However, the idea behind OpenSDE is that data are recorded during routine care to be available for patient care and clinical research. There is thus no clear question that can be used to evaluate whether a routine data set meets the desired quality. We, therefore, chose to investigate whether the users at least record the same data in the same manner. Although it is difficult to identify criteria against which quality should be judged [24], we feel that uniformity is an important aspect of data quality as it has effects on the ease of data look up and retrieval, which are important incentives for recording data in a structured manner.

The results in Table 2 show that only 22% of the findings were recorded identically by all three clinicians. This number will become progressively lower as the number of clinicians increases. Obviously, when looking at identical recording on a pair wise basis, the percentage will be much higher. However, as long as one or more clinicians have recorded the same piece of medical information differently, a researcher extracting data (or a clinician treating a patient) will have to be aware of it. Whether a patient's penicillin allergy is recorded in a structured manner as allergy or recorded as free text somewhere in the record, it is important that this information is not overlooked. Hence, to obtain insight in the challenges of data extraction, emphasis is on the nature of differences when people record the same information. We will discuss these differences by first focusing on the findings that are represented differently, but where data content is the same (categories: "structured differently" and "free text"). We will subsequently analyze those findings where there is discordance in data content (categories: "inferred", "omitted", and "conflict").

Different representation

In OpenSDE a medical finding can be represented differently either in a structured manner or as free text. In 4.9% of all findings, clinicians recorded the same finding differently in a structured manner. This can occur, for example, if a patient's mother has a heart condition. The heart condition can be recorded as part of the family history or as reason for giving birth to the patient in the hospital instead of at home (which is usual in the Netherlands). Nevertheless, it is still the same heart condition which, depending on the context, can be entered at different places in the tree. The heart condition is relevant both for the family history and the birth conditions of the child. Hence, the same information can be relevant in more than one context. In such scenarios the finding is ideally represented at one place in the domain model with only a reference to this description at the other nodes where the information is relevant.

A finding can also be represented as free text at various nodes in the tree, as was the case with 12.2% of all findings. We encountered data for a patient's father who suffered from hay fever *and* who was allergic to particular types of food. Instead of recording that the father has food intolerance, the clinician chose to record all of the father's allergies at the hay fever node as free text. Such use of free text makes the search for data more complex and less reliable, as data can be recorded anywhere as free text.

Although findings categorized as "structured differently" and "free text" are transcribed in OpenSDE, i.e., the findings are recorded in the patient's electronic record, they are not transcribed by all clinicians in the same manner. The findings are thus not recorded uniformly. Hence, in patient care clinicians may overlook data when they only look at one particular data item in the tree and do not consult all data about the patient. For bulk retrieval, the lack of uniformity can also have consequences. When data are not recorded uniformly, searching for a particular finding requires searching the entire tree of recorded data to ensure that the finding is not overlooked. As it is unpredictable where and how (e.g. abbreviations, codes, spelling or typing errors) findings can be recorded, look up and retrieval can hardly be automated. The potential benefits of SDE are then barely achieved and one should question whether the benefits of structuring the data

still outweigh the efforts.

In their study, Pringle et al. [25] conclude that subjective data are less consistently recorded than objective data. Peat et al. [26] investigated how reliable structured clinical history-taking is and conclude that subjective information leads to higher inter- and intra-observer variability. Our results show that physical examination findings are recorded identically more often than patient history findings. Furthermore, our results show that clinicians use the free text possibility more often for patient history than for physical examination findings. Although data are suitable for more purposes than free text, patient history often requires more use of free text to cover patient-specific detail. Also, a few lines with the essence of the patient's story provide more overview at a glance than when this story is 'scattered' under the various nodes of the tree. Therefore, free text should not be fully replaced with SDE, but rather be combined with SDE [2, 27].

Discordance in data content

The largest of the three categories that cover discordance in data content, is the inferred findings category which constitutes 31.1% of all findings. The majority of these inferred findings constitutes normal physical examination findings; of the 548 inferred findings, 39 (7.1%) were abnormal, meaning that such abnormalities were not recorded in the patient's paper record. Inferred abnormal findings include pain during defecation and a father with hay fever. Regarding the latter example, the paper source revealed that the clinician misread its content: the mother's and father's histories were written directly below each other, where the mother suffered from hay fever and the father from other allergies. Hence, inferred findings, as found in step six of the algorithm, may include such misreadings of the paper record.

The inferred findings are predominantly interpretations of routine expressions and interpretations based on data that are written in the paper record. For a patient suffering from constipation accompanied by blood loss, one clinician recorded that the blood was clear red, whereas the paper record made no reference to the color of the blood. Such

inferred findings, of which this is just one example, lead to believe that if there are no data recorded for an observation, then it is probably normal.

Transcribing data includes interpreting routine expressions such as 'vesicular breath sounds'. When routine expressions are not identically represented in the domain model, some clinicians opt to record such expressions as free text, whereas others opt to "translate" the routine expression into those concepts that approach the meaning of the routine expression (e.g. normal breath sounds). Such translations involve interpreting the routine expression, which can lead to different representations and derivations. This form of discordance can be reduced by adhering to terms frequently used in a particular setting when constructing the domain model.

The inferred findings, especially the inferred normal findings, which constitute the majority of the inferred findings, indicate that OpenSDE induces an effect similar to the checklist effect [28]: clinicians are triggered by the available entry fields to record more data.

Even though OpenSDE triggers clinicians to record particular data, our results also show that 26.7% of all findings were omitted by one or more clinicians. Although the majority of omitted findings (79%) were normal findings, an omission of 26.7% of all findings is dissatisfying when the purpose of SDE is to improve quality, completeness, and consistency of data. Findings are omitted when clinicians overlook findings in the paper record, ignore "irrelevant" findings, or when recording the finding in OpenSDE is not straightforward.

In their study assessing the completeness and accuracy of computer medical records, Pringle et al. [25] suggest that "it was clear that practices were selecting areas that they considered important to record on their computer systems". In line with these results, our results suggest that clinicians are more inclined to omit normal findings than abnormal findings, as normal findings are generally of lesser importance when examining the patient than abnormal findings. Nevertheless, for the purpose of improving data quality, recording observed normal findings is also important.

The last category that we analyzed involved the conflicting findings. Just over 3% of all recorded findings were in direct conflict with the data in the paper record. Conflicting findings include recording previously used medication as current medication, recording a cardiac murmur when the patient does not have a cardiac murmur, and recording incorrect numer-

ical values. In real life, when clinicians directly record findings using OpenSDE instead of transcribing findings from paper records, the percentage of conflicting data will probably be lower; errors in judgment or typographic errors will still be made but transcription errors due to, for example, misreading of the paper record, will no longer apply. SDE may, however, introduce a new category of errors, such as erroneous selection from checklists.

Limitations of the study

Studies such as ours have three particular limitations. The first limitation is that we only analyzed the uniformity in the *transcribed* records. We did not analyze whether the transcription of findings in the paper record using OpenSDE was complete. Such a study would provide insight into the effect of OpenSDE on promoting completeness [29]. However, this was not our goal. Our interest in this particular study is *how* people represent patient data using OpenSDE when they have the same source of data.

The source of data is a second limitation. Transcribing findings from a paper record involves interpreting the recorded interpretations of a colleague. In an ideal situation all clinicians are confronted with the actual patient instead of only with a paper record. However, this is undesirable for patient care, as one cannot ask of patients (especially children) to tell their story to several clinicians. Besides that, the clinicians will approach the patients differently and ask different questions to which the patient may give answers that vary, for example, in level of detail. This introduces bias into the study as the source of data is not identical for all clinicians. In addition, the patient may also become biased by the questions of a previous clinician when elaborating about his complaints to the next clinician. We should also keep in mind our research question: does OpenSDE invite users to record data from the same source in a uniform structure? To answer this question it is essential that all involved clinicians consult the same data source. The results of our analysis give us insight into *how* people can and will represent data and what the differences in representation are.

The third limitation is the number of clinicians included

in the study. Even though there was not one clinician that outperformed the others in terms of the number of structured findings per record, an increased number of clinicians may reveal even more differences in data representation. However, we feel that three clinicians do provide enough insight into how data can be recorded, especially as we encountered situations which we had not considered. There was one clinician, for example, that had four different manners of mapping the same free text term from paper to OpenSDE. Nevertheless, these results do give us insight into how data are represented in a routine clinical situation and which potential pitfalls this creates for data extraction.

Conclusion

Structured data entry is intended to improve the quality and consistency of data by obtaining the data directly from clinicians in a structured format. To analyze the uniformity of data recorded with OpenSDE, we performed a study in which three pediatricians used OpenSDE to transcribe 20 handwritten paper patient records. Our results show that data recorded using SDE are not necessarily represented in the same manner and nearly two-thirds of the recorded data are discordant (i.e., inferred, omitted, or conflicting with data in the paper record).

In line with other studies our results indicate that even though information is more accessible it is not necessarily creditable [30], directly usable [31], or structured uniformly. As a result, data collected with OpenSDE cannot unconditionally be used for subsequent purposes such as clinical research. Mikkelsen and Aasly take the claim even further and say that inconsistencies in information elements used to characterize clinical information represents a potential threat to the safety of using EPRs as source of clinical information [14]. We did not go as far as to evaluate the actual clinical consequences of the differences in data representation but based on our differences and the conclusions of other work, this also requires attention. Studies such as ours increase insight into retrieval pitfalls, independent of the system being used. Standardizing data entry and multiple search strategies are certainly necessary before aggregated data can be

relied upon [25].

Based on the results of this study, we are currently addressing the following two aspects. Firstly, we are focusing on increasing the uniformity in data entry by limiting the number of ways in which the same information can be recorded, without limiting the level of detail in which data can be recorded. Secondly, we are investigating multiple search strategies for data retrieval, to increase the probability that all relevant data are actually retrieved.

This study has pointed out those aspects in the design of OpenSDE where problems arise both during data entry as well as during use of the data. Insight into the difference in data recording is useful because it helps us improve the design of OpenSDE for data entry, with the aim of improving the data set and enhancing the potential clinical use of the database [31]. The results of this study show that recording data using structured data entry does not necessarily lead to uniformly structured data.

In general what can be learned from this study is that for data lookup and retrieval one must be aware of all possible ways in which an item of information may have been recorded.

Acknowledgements

We would like to thank the pediatricians involved in this study for the time and effort spent transcribing data.

The work presented in this paper is funded by a grant from the Netherlands Organization for Health Research and Development (ZonMW).

References

1. Dick RS, Steen EB, Detmer DE, eds. The Computer-Based Patient Record: An Essential Technology for Health Care. Revised Edition ed. Washington: National Academy Press; 1997.

2. van Ginneken AM. The Computerized Patient Record: Balancing Effort and Benefit. Int J Med Inf 2002;65(2):97-119.

3. Powsner SM, Wyatt JC, Wright P. Opportunities for and Challenges of Computerisation. Lancet 1998;352(9140):1617-22.

4. Brown PJ, Sönksen P. Evaluation of the Quality of Information Retrieval of Clinical Findings from a Computerized Patient Database Using a Semantic Terminological Model. J Am Med Inform Assoc 2000;7(4):392-403.

5. Moorman PW, van Ginneken AM, van der Lei J, van Bemmel JH. A Model for Structured Data Entry Based on Explicit Descriptional Knowledge. Methods Inf Med 1994;33(5):454-63.

6. Tange H. How to Approach the Structuring of the Medical Record? Towards a Model for Flexible Access to Free Text Medical Data. Int J Biomed Comput 1996;42(1-2):27-34.

7. Walsh SH. The Clinician's Perspective on Electronic Health Records and How They Can Affect Patient Care. Bmj 2004;328(7449):1184-7.

8. Klar R. Selected Impressions on the Beginning of the Electronic Medical Record and Patient Information. Methods Inf Med 2004;43(5):537-42.

9. Tange HJ, Hasman A, de Vries Robbe PF, Schouten HC. Medical Narratives in Electronic Medical Records. Int J Med Inf 1997;46(1):7-29.

10. Los RK, van Ginneken AM, de Wilde M, van der Lei J. Opensde: Row Modeling Applied to Generic Structured Data Entry. J Am Med Inform Assoc 2004;11(2):162-65.

11. OpenSDE. Opensde (Oss). http://webserver.mi.fgg.eur.nl/opensde/. Last accessed: March 31, 2005.

12. Beard CM, Yunginger JW, Reed CE, O'Connell EJ, Silverstein MD. Interobserver Variability in Medical Record Review: An Epidemiological Study of Asthma. J Clin Epidemiol 1992;45(9):1013-20.

13. Williams JG. Measuring the Completeness and Currency of Codified Clinical Information. Methods Inf Med 2003;42(4):482-8.

14. Mikkelsen G, Aasly J. Consequences of Impaired Data Quality on Information Retrieval in Electronic Patient Records. Int J Med Inform 2005;74(5):387-94.

15. Ammenwerth E, Shaw NT. Bad Health Informatics Can Kill—Is Evaluation the Answer? Methods Inf Med 2005;44(1):1-3.

16. Wyatt J. Same Information, Different Decisions: Format Counts. Bmj 1999;318:1501-2.

17. van Ginneken AM, Verkoijen MJ. A Multi-Disciplinary Approach to a User Interface for Structured Data Entry. Medinfo 2001;10(Pt 1):693-7.

18. Los RK, van Ginneken AM, van der Lei J. Opensde: A Strategy for Expressive and Flexible Structured Data Entry. Int J Med Inform 2005;74(6):481-90.

19. Roukema J, van Ginneken AM, Moll HA. The Use of Structured Data Entry in the Outpatient's Clinic for Paediatrics. Health Information Developments in the Netherlands 2003;6:27-30.

20. Kahn CE, Jr., Huynh PN. Knowledge Representation for Platform-Independent Structured Reporting. Proc AMIA Annu Fall Symp 1996:478-82.

21. Stoop AP, Berg M. Integrating Quantitative and Qualitative Methods in Patient Care Information System Evaluation: Guidance for the Organizational Decision Maker. Methods Inf Med 2003;42(4):458-62.

22. Rubenfeld GD. Using Computerized Medical Databases to Measure and to Improve the Quality of Intensive Care. J Crit Care 2004;19(4):248-56.

23. Winthereik BR. "We Fill in Our Working Understanding": On Codes, Classifications and the Production of Accurate Data. Methods Inf Med 2003;42(4):489-96.

24. Thiru K, Hassey A, Sullivan F. Systematic Review of Scope and Quality of Electronic Patient Record Data in Primary Care. Bmj 2003;326(7398):1070.

25. Pringle M, Ward P, Chilvers C. Assessment of the Completeness and Accuracy of Computer Medical Records in Four Practices Committed to Recording Data on Computer. Br J Gen Pract 1995;45(399):537-41.

26. Peat G, Wood L, Wilkie R, Thomas E. How Reliable Is Structured Clinical History-Taking in Older Adults with Knee Problems? Inter- and Intraobserver Variability of the Kne-Sci. J Clin Epidemiol 2003;56(11):1030-7.

27. McDonald CJ. The Barriers to Electronic Medical Record Systems and How to Overcome Them. J Am Med Inform Assoc 1997;4(3):213-21.

28. Wyatt J. Quantitative Evaluation of Clinical Software, Exemplified by Decision Support Systems. Int J Med Inf 1997;47(3):165-73.

29. Roukema J, Los RK, Bleeker SE, van Ginneken AM, van der Lei J, Moll HA. Paper Versus Computer: Feasibility of an Electronic Medical Record in General Pediatrics. Pediatrics 2005;Accepted for publication.

30. Burnum JF. The Misinformation Era: The Fall of the Medical Record. Ann Intern Med 1989;110(6):482-4.

31. Wyatt JC, Wright P. Design Should Help Use of Patients' Data. Lancet 1998;352(9137):1375-8.

5

6 Why are structured data different?

Relating differences in data representation to the rationale of OpenSDE

Renske K. Los, Astrid M. van Ginneken, Jolt Roukema, Henriette A. Moll, Johan van der Lei

Accepted for publication in:
Medical Informatics and the Internet in Medicine

Abstract

Objective: OpenSDE is an application that supports clinicians with structured recording of narrative patient data to enable use of data in both clinical practice and research. OpenSDE is based on a rationale and requirements for structured data entry. In this study we analyze the impact of the rationale and the requirements on data representation using OpenSDE.

Methods: Three pediatricians transcribed 20 paper patient records using OpenSDE. The transcribed records were compared; the findings that were the same in content but differed in representation (e.g. recorded as free text instead of in a structured manner), were categorized in one of three categories of difference in representation.

Results: The transcribed records contained 1764 findings in total. The medical content of 302 of these findings was represented differently by at least one clinician, and was thus included in this study. In OpenSDE clinicians are free to determine the degree of detail at which patient data are described. This flexibility accounts for 87% of the differences in data representation. 13% of the differences are due to clinicians interpreting and translating phrases from the source text and transcribing these to (different) concepts in OpenSDE.

Conclusion: The differences in data representation largely result from initial design decisions for OpenSDE.

'It can readily be seen that all narrative data presently in the medical record can be structured, and […] entered through series of displays, guaranteeing a thoroughness, retrievability, efficiency and economy important to the scientific analysis of a type of datum that has hitherto been handled in a very unrigorous manner.'

-Lawrence L. Weed, 1968 [1].

Introduction

Electronic patient records (EPRs) are associated with many potential benefits such as availability of patient data for decision support, quality assessment, or clinical research [2, 3]. However, to benefit from such advantages, data must be represented in a structured format [4, 5]. Structured Data Entry (SDE) is a method by which clinicians record patient data directly in a structured format. SDE offers predefined fields for data entry. As early as 1968 the potential for SDE, as well as subsequent use of the collected data for analysis purposes, was recognized by Weed [1]. However, to date, SDE remains challenging to apply for medical narratives (especially patient history and physical examination) as these data vary per domain, per patient, and over time [6-9].

Since the early 1990s our philosophy regarding SDE has been that free-text narratives should be minimized in favor of clinically relevant structured data for multiple purposes. Our rationale for SDE is based on data entry by clinicians. The challenge is to approach the expressive power of free text, whilst keeping SDE acceptable for clinicians.

In OpenSDE, our current SDE application, we respected the clinicians' need for flexibility and expressiveness, i.e. data entry with certain degrees of freedom, to describe findings. Freedom in data entry, however, implies that the same data may be recorded differently by different clinicians [10]. For purposes such as research and decision support, on the other hand, a structured, uniform representation of the same data set is essential. The question that thus arises is: what is the impact of the expressiveness and flexibility offered to support SDE, on the uniformity of the data set?

In this paper we discuss differences in data representa-

tion that are a result of the design of OpenSDE, and propose changes in the data entry paradigm to increase the uniformity in data representation.

Research Focus

In view of our rationale for SDE, Moorman et al. formulated requirements that should be met to make structured data entry acceptable for clinicians [11]. Three of these requirements are pivotal for OpenSDE [12, 13]. First of all, SDE should provide sufficient *expressive* power to describe clinically relevant details; this expressiveness must be offered in the form of predefined terms (which by definition limits expressiveness). Secondly, SDE has to be *flexible* to offer the clinician the freedom to determine the order and degree of detail of what he describes; enforcing detail or order in data entry does not enhance acceptability. The third crucial requirement is that data should be presented in a *predictable* order so that, when browsing through the data, clinicians know where to expect specific information.

To analyze the impact of expressiveness, flexibility, and a predictable order on data representation, we performed a study in which three pediatricians transcribed patient data from a common data source. Research has shown that when transcribing findings from the same handwritten paper source into a flexible structured electronic record, differences between the three transcribed records are inherent [14, 15]. Horwitz and Yu report three different data recording errors: conflicting data in the source text, information not transcribed, and transcription errors [14, 16]. In our study, we distinguish two types of differences: firstly, there are differences in data content and secondly, there are differences in data representation. Differences in data content include errors such as those described by Horwitz and Yu. We are, however, particularly interested in patient data that differ in representation, for example, a finding is recorded (partly) as free text instead of in a structured manner. This type of difference implies that the participating clinicians recorded findings representing the same medical content but structured the patient data differently. Even in OpenSDE we cannot guarantee that the same content is represented in the same manner. In

OpenSDE, findings that are transcribed differently can vary both in level of structure and place in the database where the findings were recorded. The object of this study is to identify categories of differences between representations of the same patient data. Based on these categories we can propose changes that limit differences in data representation.

Materials

OpenSDE

OpenSDE supports clinicians with the structured recording of medical narratives [17-19]. The pivot of OpenSDE is the domain model: a tree of hierarchically ordered medical concepts. The tree is domain-specific and holds the concepts necessary to describe findings in a particular domain of medicine. Domain models are created by domain experts using a specifically designed tool [18]. The use of the tool as such is not difficult; the difficulty lies in the actual modeling. There are two main issues that make modeling complex. Firstly, the modelers need to decide to what level of detail concepts should be modeled in the tree [20]. Secondly, one needs to minimalize the number of possibilities in which the same data can be recorded in different manners.

In a domain model, the path from the top of the tree to a particular concept represents the context of that concept. A typical domain model will start with very broad concepts which become more specific as the tree branches. Each concept in the tree is associated with an application generated entry form (which can be customized using an integrated form editor). Using this entry form clinicians can describe a particular finding, such as a new mole, in more detail. Details may include whether the finding applies or not (e.g. a mole is present), the date of the mole's discovery (temporal value), the size (numerical value), and the color (categorical value) of the mole. It is also possible to describe findings more than once in the context of progression over time (e.g. changes in size/color), different circumstances (e.g. before/after sunbathing), or multiple occurrences (more than one mole).

Like many systems designed for recording heterogeneous and evolving data sets [21-24], OpenSDE relies on a generic data model for data storage [12]. Figure 1 presents a screen capture of OpenSDE.

OpenSDE reflects the three essential requirements for SDE. The hierarchical nature of the domain model allows specifying findings at varying levels of granularity to accommodate the desired degree of structured expressiveness (first SDE requirement). The hierarchy presents concepts for data entry in a predictable order (third SDE requirement). Flexibility, the second SDE requirement, is supported in two ways. Firstly, OpenSDE does not enforce a specific order or level of detail at which findings must be described. Secondly, OpenSDE does not enforce structure; recording data as free text is always possible (at every concept in the tree) for particular details not covered by the content of the domain model.

Prior to this study, experienced OpenSDE users recorded data of over 100 pediatric paper records in OpenSDE to evaluate the ordering and coverage of the pediatric domain model. The model was then altered to improve both ordering and coverage, as well as to facilitate data entry[25].

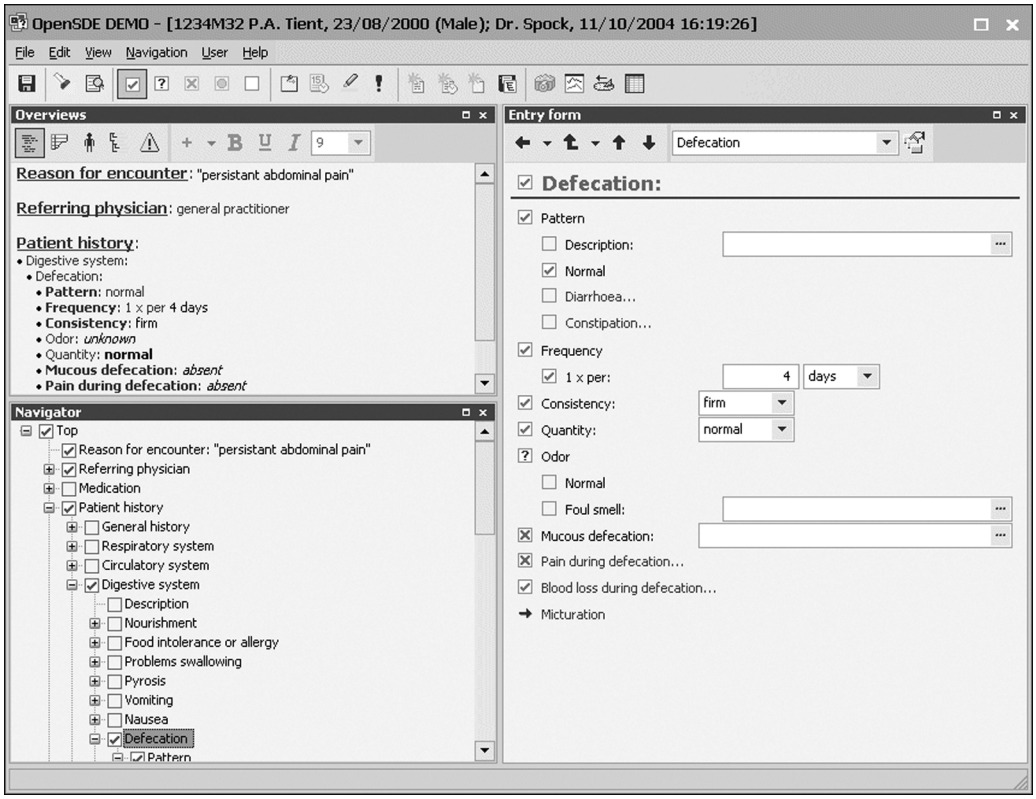

Figure 1. The top left of the screen shows an overview of the data recorded for the patient in the current session. The bottom left shows the domain model tree with medical concepts. On the right is the form on which data are entered. The form is associated with the selected node, in this case 'defecation'. At the bottom of the defecation form, the term 'Micturition' is preceded by an arrow which indicates that micturition is modeled in detail elsewhere. Clicking on the term will present the form used to describe micturition.

Methods

A. Data entry from a common data source

Three pediatricians working at our hospital's pediatric outpatient department followed a standardized course on the use of OpenSDE in general pediatrics. We randomly selected 20 handwritten paper patient records created for first-

contact patients at the pediatric outpatient department. These patients were not treated by any one of the three pediatricians involved in this study. Each pediatrician transcribed the 20 paper records in OpenSDE, resulting in a total data set of 60 transcribed records. The pediatricians were informed about the goal of the study, and knew that the transcribed records would be analyzed.

B. Non-uniformly transcribed patient data

Our analysis consisted of two steps. The first step involved manually analyzing the medical content and data representation in the transcribed records. Per patient record we explored whether all three clinicians recorded the same medical content identically, differently, or whether there was a difference in data content (e.g. errors, or missing data). If the same medical content was present in all three transcribed records, but represented differently in at least one of the transcribed records, the corresponding findings were included in this study. Patient data represented in the same manner by all three clinicians, or transcribed findings that differed in data content were excluded from the study.

The second step consisted of identifying categories of differences in data representation. We distinguished three categories of differences. Consequently we classified each finding in one of the categories as described below.

If the same medical content was represented (partly) as free text at a different node in the same path, i.e. in more or less detail, the finding was classified as a difference due to flexibility in representation ('Flexibility' in Table 1). Figure 2 presents an example where stomach ache is described in a structured manner by recording details such as onset, localization, and duration at the specific concepts in the tree, and also shows how these details can be described as free text at the concept 'stomach ache'.

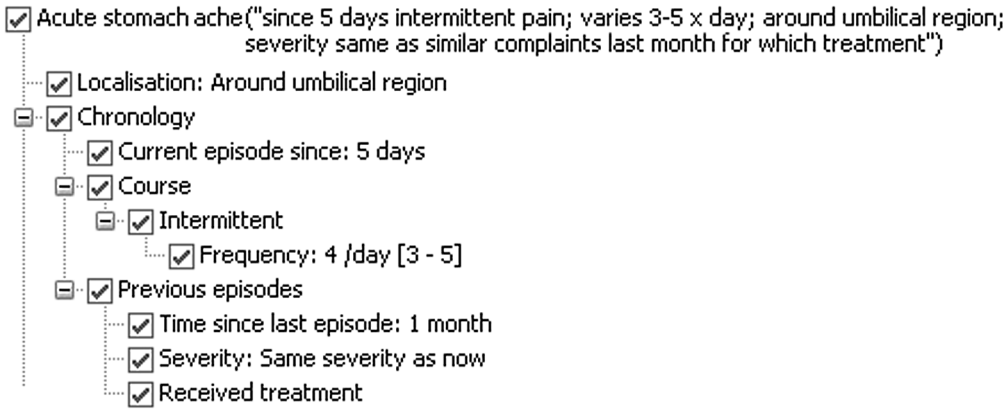

Figure 2. The description of acute stomach ache is represented both in a structured format in the tree, and as free text behind the node 'acute stomach ache'.

If the finding did not belong to the first category, we checked if the patient data was represented in a different path in the tree. In some cases a domain model offers multiple entry options to describe the same medical content. If a finding is represented at such a semantically similar concept, we classified the finding as a difference due to semantic similarity ('Semantic similarity' in Table 1). Nutrition is such an example. Eating habits can be described by normal, increased, or decreased appetite and are relevant in the context of the digestive system. Nutritional intake, on the other hand, was modeled separate from the digestive system. As eating habits and intake are closely related, modeling these concepts apart from each other, at different places in the tree is not practical for data entry. Hence some users record all related patient data at only one of these concepts: e.g. eating habits are represented as free text at the nutritional intake concept.

The last category of difference that we distinguished constitutes findings that involve a judgment or an interpretation as the phrases used in the paper record cannot directly be translated (or mapped) to the same concepts in the domain model ('Mapping' in Table 1). For example, the paper record may contain the phrase 'lively bowel sounds'. If the entry options in OpenSDE only include bowel sounds normal or abnormal, translating 'lively bowel sounds' will require interpret-

ing whether lively bowel sounds are normal or abnormal. Some clinicians will choose to interpret these bowel sounds as 'normal bowel sounds', whereas another may choose to record 'lively bowel sounds' as free text.

Per finding meeting the inclusion criteria we also established the part of the patient record it belonged to (patient history or physical examination) and whether the finding was normal (e.g. 'no cardiac murmur') or abnormal (e.g. 'constipation').

Results

The transcribed records contained a total of 1764 findings; the medical content of 302 of these findings (17% of all findings) was represented differently by the three clinicians. These 302 findings are the findings of interest for this study. In Table 1 we present the findings per category of difference and we divide the findings into patient history or physical examination findings. The table shows that most findings are patient history findings. Patient history findings are predominantly abnormal findings, whereas physical examination findings are mostly normal.

Of all 302 findings, the majority (83%) constitutes 'flexibility' differences, 13% involves 'mapping', and 4% was classified as differences due to 'semantic similarity'.

Table 1. Findings represented differently ordered per category of difference
This table presents the findings structured differently by the three clinicians. A total of 302 findings were represented differently. Per category of difference the findings are split into patient history or physical examination findings, and consequently subdivided into normal or abnormal findings. The percentage behind the numbers corresponds to the percentage of the total number of findings.

Finding	**Patient History**		**Physical Examination**		***ALL FINDINGS***
Category	Normal	Abnormal	Normal	Abnormal	
Flexibility	89 (29.5%)	128 (42.4%)	30 (9.9%)	5 (1.7%)	**252** (83.4%)
Semantic similarity	3 (1%)	9 (3%)			**12** (4%)
Mapping	4 (1.3%)	3 (1%)	28 (9.3%)	3 (1%)	**38** (12.6%)
ALL FINDINGS	**96** *(31.8%)*	**140** *(46.4%)*	**58** *(19.2%)*	**8** *(2.6%)*	**302** *(100%)*

Differences due to flexibility occur most often in the abnormal patient history findings, as opposed to the physical examination where the clinicians chose to represent the normal findings at different places in a path.

As shown in Table 1, differences due to mapping are mainly normal physical examination findings.

Discussion

The goal of OpenSDE is that clinicians can record patient data in a format in which data are usable for multiple purposes. One of the main challenges for OpenSDE was to approach the expressive power and flexibility of free text, whilst keeping SDE acceptable for clinicians. To meet this challenge, Moorman et al. [11] formulated requirements for SDE.

In this study we distinguish three categories of differences in data representation. In hindsight, these categories result from initial design decisions and more specifically from the three requirements of supporting expressiveness, flexibility, and a predictable order for data entry. In this discussion we will first focus on the consequences of our design decisions for SDE on the representation of data in the context of data extraction. Consequently, we propose alterations that aim to reduce the differences in data representation whilst upholding the underlying design philosophy.

A. Effect of requirements for SDE on data representation

A1. 'Flexibility' and flexibility (second requirement for SDE)

Of all findings represented differently in the transcribed records, a majority of 83% was categorized as different due to 'flexibility'.

To support flexibility in SDE, OpenSDE firstly does not enforce a specific detail-level or structure in which findings must be described, and secondly enables use of free text where needed (second SDE requirement). This design has led to a very flexible use of OpenSDE; our results illustrate that clinicians use free text to represent findings that can be structured. Clinicians also record the same data as free text at different places in the tree. Flexibility is an advantage for data entry, but it is a hurdle for data look-up and extraction as data can be recorded at more than one place (making data representation less predictable). The dilemma we now face is: do we uphold our rationale and retain this flexibility in data entry, or do we compromise this flexibility in order to increase uniformity of the data set?

A2. 'Mapping' and expressive power (first requirement for SDE)

Almost 13% of the findings represented differently were categorized as differences due to 'mapping'.

Although predefined terms imply limited expressiveness,

OpenSDE aims to provide the clinician with sufficient expressive power to describe clinically relevant details. A domain model is, however, limited in scope and may not always contain the exact terms that clinicians would like to use, or may not present terms in the exact context in which clinicians would preferably use the terms[1]. Transcribing findings thus involves interpreting the finding in the paper record (for which the clinician uses his own reference [26]) and then translating or mapping the finding to those concepts in OpenSDE that best match the description of the finding in the paper record. Almost 13% of the findings were categorized as such mapping differences.

Analysis of these 38, predominantly physical examination findings revealed that for 30 of these findings, the concept 'normal' or 'abnormal' was modeled in the finding's path. Although the domain model allows for recording of expressiveness (including judgments), the predefined order of the terms has effect on how the data are represented. In cases where judgments are recorded, expressiveness is often incorrectly modeled (observations are often modeled as branching nodes of concepts representing judgments). The domain model forces the user to judge whether particular findings are normal or abnormal. The clinical meaning of terms such as 'normal' is, however, subject to interpretation of the clinician that records the findings and the clinicians or researchers that consult the findings [27]. The example of the lively bowel sounds mentioned previously represents a situation that can be normal in one scenario and abnormal in another. Nevertheless, the frequent use of such subjective terms, both in the pediatric domain model and in the paper records, indicates that clinicians apparently have a need to express that, according to their judgment, particular findings are 'normal'. The question now is: how should OpenSDE support expressive recording of observations and interpretations?

[1] A similar situation can occur when clinicians enter findings directly in OpenSDE instead of transcribing findings from a paper record.

A3. 'Semantic similarity' and predictable order (third requirement for SDE)

The last category of difference involves findings classified in the 'semantic similarity' category. A total of 4% of findings represented differently are recorded at more than one branch in the tree due to duplication of concepts or the presence of semantically very similar concepts in the domain model.

OpenSDE has functionality to handle patient data that are relevant to describe in more than one context. The functionality consists of a reference mechanism within a domain model to accommodate access to concepts via more than one context, while the data are only represented in one, unique way (i.e. in one predictable order: third SDE requirement).

B. Future choices

The results of this study illustrate that providing freedom in data entry, will lead to use of this freedom during data entry. This freedom, therefore, is in conflict with uniform data representation. Our initial goal was that clinicians directly record data in a structured manner suitable for multiple purposes. The question is thus: should we retain focus on facilitating data entry in order to get as many clinicians using the application as possible, and sacrifice the uniformity of the collected data, or should the uniformity of the data set be our priority and should we sacrifice the freedom in data entry? The first option may lead to a more widespread use of OpenSDE, but is less effective in promoting data collection suitable for other purposes such as (retrospective) research or quality assessment. The second option, on the other hand, may be preferable for additional benefits of structured data, but if clinicians refuse to use the application, then there are no data to benefit from. In essence, the clinicians are pivotal in the data collection process, but must we go to great lengths to accommodate needs and preferences for data entry, if this means that potential use of data becomes limited [28]?

In an attempt to reduce the differences in data recording with OpenSDE, we propose four measures that should improve the uniformity of the data set, whilst minimizing the impact on flexibility and expressiveness. The proposed measures include:

1 limiting the use of free text;
2 explicitly separating interpretations from judgments;
3 facilitating uniform data entry by using templates and checklists;
4 developing modeling guidelines.

Our first proposal is to limit the use of free text. When the aim is to structure data for research purposes, free text is ideally limited, as free text complicates research on the data. Nevertheless, for patient care free text cannot be eliminated [29]. Therefore, we propose to limit the use of free text to predefined nodes in the tree. If a clinician chooses to record free text at any node in the path, he will be redirected to the node at which it is allowed. This construction does not sacrifice expressiveness as free text can still be added, and offers the advantage that when consulting the data, free text is limited to predictable places. This may not reduce the use of free text, but it will reduce the chance of overlooking a finding recorded as free text, as free text cannot be scattered anywhere in the tree.

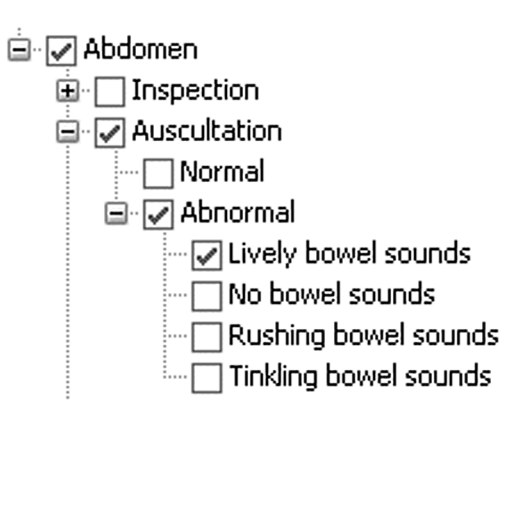

Figure 3a.

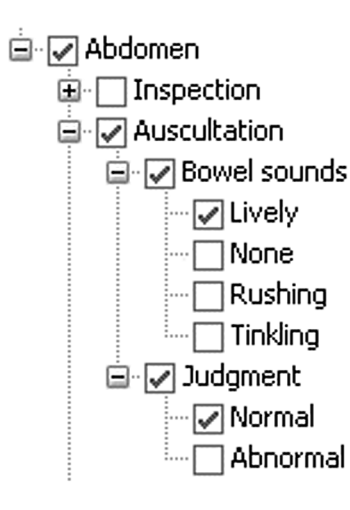

Figure 3b.

Figure 3a shows that modeling findings in the same path as judgmental concepts such as abnormal can lead to erroneous data. **Figure 3b** illustrates a different way to model judgmental concepts, namely not in the same path but parallel to the findings. This allows for structured representation of the (objective) finding, as well as the addition of a judgment.

Our second proposal is to separate specific judgmental concepts about observations from the observations themselves. Judgments such as 'normal' are ideally separated from descriptive concepts such as 'lively bowel sounds' in the domain models. By separation we mean not to place such concepts in the same path. Figure 3 illustrates how not to model judgmental concepts (Figure 3a) and how we propose to model judgmental concepts (Figure 3b).

For those situations in which more than one finding is normal, e.g. the auscultation of the heart is normal, we suggest the use of customized (user-specific) templates for particular sets of findings. A template 'auscultation heart normal' will consist of findings such as first and second heart sounds present, no murmurs. This template will mean the same across all patients seen by a particular clinician. Templates have obvious advantages and disadvantages and their use is not without risk. Templates may lead to documentation of observations that

were not performed. On the other hand, not using templates may lead to findings inadvertently being omitted from the patient record [29]. Using templates has two practical advantages: it can speed up data entry as well as increase the consistency of the entered data.

In those situations where structured recording of particular findings is essential (e.g. for vital patient characteristics or prospective research purposes) the use of data checklists to remind users to record data about particular findings, is recommended. Checklists are an optional function available in OpenSDE.

A last proposal is to develop modeling guidelines. The guidelines should emphasize recommendations two and three as well as promote the use of referencing to avoid duplication of semantically similar descriptions. The modeling guidelines should pursue a balance between representing those concepts that may be relevant to describe in particular contexts and representing concepts in such a manner that uniformity in data representation is optimized.

We expect that these alterations will improve the uniformity of the data set, whilst impact on flexibility and expressiveness for data entry is minimal [27], thus upholding the rationale and requirements. However, if there is anything that we learned from this study, it is that the impact of decisions is not fully predictable, and repercussion is often unforeseen.

References

1. Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(11):593-600.

2. Dick RS, Steen EB, Detmer DE, eds. The computer-based patient record: an essential technology for health care. Revised Edition ed. Washington: National Academy Press; 1997.

3. van Ginneken AM. The computerized patient record: balancing effort and benefit. Int J Med Inf 2002;65(2):97-119.

4. Powsner SM, Wyatt JC, Wright P. Opportunities for and challenges of computerisation. Lancet 1998;352(9140):1617-22.

5. Brown PJ, Sönksen P. Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. J Am Med Inform Assoc 2000;7(4):392-403.

6. Tange H. How to approach the structuring of the medical record? Towards a model for flexible access to free text medical data. Int J Biomed Comput 1996;42(1-2):27-34.

7. Walsh SH. The clinician's perspective on electronic health records and how they can affect patient care. Bmj 2004;328(7449):1184-7.

8. Tange HJ, Hasman A, de Vries Robbe PF, Schouten HC. Medical narratives in electronic medical records. Int J Med Inf 1997;46(1):7-29.

9. Bernauer J. Conceptual Graphs as an Operational Model for Descriptive Findings. In: Proc Annu Symp Comput Appl Med Care; 1992; 1992. p. 214-18.

10. Lovis C, Baud RH, Planche P. Power of expression in the electronic patient record: structured data or narrative text? Int J Med Inform 2000;58-59:101-10.

11. Moorman PW, van Ginneken AM, van der Lei J, van Bemmel JH. A model for structured data entry based on explicit descriptional knowledge. Methods Inf Med 1994;33(5):454-63.

12. Los RK, van Ginneken AM, de Wilde M, van der Lei J. OpenSDE: row modeling applied to generic structured data entry. J Am Med Inform Assoc 2004;11(2):162-65.

13. OpenSDE. OpenSDE (OSS). http://sourceforge.net/projects/opensde. Last accessed: August 24, 2005.

14. Beard CM, Yunginger JW, Reed CE, O'Connell EJ, Silverstein MD. Interobserver variability in medical record review: an epidemiological study of asthma. J Clin Epidemiol 1992;45(9):1013-20.

15. Stausberg J, Koch D, Ingenerf J, Betzler M. Comparing paper-based with electronic patient records: lessons learned during a study on diagnosis and procedure codes. J Am Med Inform Assoc 2003;10(5):470-7.

16. Horwitz RI, Yu EC. Assessing the reliability of epidemiologic data obtained from medical records. J Chronic Dis 1984;37(11):825-31.

17. Doupi P, van der Lei J. Towards personalized Internet health information: the STEPPS architecture. Med Inform Internet Med 2002;27(3):139-51.

18. Los RK, van Ginneken AM, van der Lei J. OpenSDE: A strategy for expressive and flexible structured data entry. Int J Med Inform 2005;74(6):481-90.

19. Roukema J, Los RK, Bleeker SE, van Ginneken AM, van der Lei J, Moll HA. Paper versus computer: feasibility of an electronic medical record in general pediatrics. Pediatrics 2005;Accepted for publication.

20. Doupi P, van Ginneken AM. Structured physical examination data: a modeling challenge. Medinfo 2001;10(Pt 1):614-8.

21. Duftschmid G, Gall W, Eigenbauer E, Dorda W. Management of data from clinical trials using the ArchiMed system. Med Inform Internet Med 2002;27(2):85-98.

22. Nadkarni PM, Brandt C, Frawley S, Sayward FG, Einbinder R, Zelterman D, et al. Managing attribute—value clinical trials data using the ACT/DB client-server database system. J Am Med Inform Assoc 1998;5(2):139-51.

23. Ganslandt T, Mueller M, Krieglstein CF, Senninger N, Prokosch HU. A flexible repository for clinical trial data based on an entity-attribute-value model. Proc AMIA Symp 1999:1064-67.

24. Miller PL, Nadkarni P, Singer M, Marenco L, Hines M, Shepherd G. Integration of multidisciplinary sensory data: a pilot model of the human brain project approach. J Am Med Inform Assoc 2001;8(1):34-48.

25. Roukema J, van Ginneken AM, Moll HA. The use of structured data entry in the outpatient's clinic for paediatrics. Health Information Developments in the Netherlands 2003;6:27-30.

26. Moorman PW, Siersema PD, de Ridder MA, van Ginneken AM. How often is large smaller than small? Lancet 1995;345(8953):865.

27. Rector AL, Nowlan WA, Kay S. Foundations for an electronic medical record. Methods Inf Med 1991;30(3):179-86.

28. McDonald CJ. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc 1997;4(3):213-21.

29. Gregory J, Mattison JE, Linde C. Naming notes: transitions from free text to structured entry. Methods Inf Med 1995;34(1-2):57-67.

7 Increasing Uniformity in Representation of Structured Data?

Renske K. Los, Jolt Roukema, Astrid M. van Ginneken, Marinka Twilt, Christiaan de Bruin, Chantal van het Zandt, Henriette A. Moll, Johan van der Lei

Submitted

7

Abstract

Objective: OpenSDE is an open source application that supports structured recording of narrative patient data to enable use of these data in both clinical practice and clinical research. Clinical research requires uniform data representation. A previous study on uniformity of data led us to modify OpenSDE. Modifications included: reducing the ease of recording data as free text, separating observations from value judgments, and expanding the domain model with terms that correspond more to the terms used in the paper records. In this study we investigate whether these modifications increase uniformity in data representation.

Methods: Three clinicians transcribed 20 pediatric paper patient records using OpenSDE. The transcribed records were compared and all corresponding findings were classified into one of six categories of difference.

Results: Almost 28% of all findings were recorded identically; in the previous study (which led to the modifications) only 22% of all findings were recorded identically. Findings denoting the same medical content, but represented in different ways, compose 11% of all findings (17% in the previous study). More than half of these findings are recorded as free text instead of structured. In the first study, almost three quarters of these findings were recorded as free text.

Conclusion: Reducing the ease of recording data as free text increases the percentage of findings recorded identically but also increases the number of omitted findings. The largest challenges in structuring findings for multiple purposes are reducing semantic equivalents for data entry and finding the right balance between free text and structured data.

Introduction

The medical narrative in patient history, physical examination, and progress notes forms a substantial part of a patient's medical record. Although the patient record is becoming increasingly computerized and data are progressively represented in coded or structured manners, the medical narrative often remains represented as free text [1]. The content of the narrative varies across different medical domains and even varies across patients seen by the same clinician for the same complaint. This potential diversity makes the narratives challenging to structure in a generic manner [2-5]. Nevertheless, to benefit from potential advantages of electronic patient records, such as decision support, clinical research, and management of patient care [6, 7], the medical narrative needs to be represented in a structured format [8, 9].

Structured Data Entry (SDE) is a method to obtain data in a structured format by offering predefined options for data entry. Our aim of SDE was to capture structured, coded data without a priori limitation on expressiveness [10]. Over the last decade, research and development has led to OpenSDE: an (open source) application to support clinicians with structured recording of data [11, 12]. OpenSDE has the potential to lead to a data repository that can be used for (retrospective) research. However, reliability and accuracy of collected data are pivotal if data will be collected over long periods of time and by different users [13, 14]. Insight into differences in data representation is particularly important for data retrieval and subsequent data use [15, 16].

In a previous study we concluded that when OpenSDE is used to transcribe data from the same source (i.e., a paper patient record), differences occur both in the content of the transcribed data and in the representation of the data [17]. In this study we, therefore, propose a number of modifications to improve uniformity in data representation. Subsequently three clinicians transcribe the same data set in the modified OpenSDE and we analyze the consensus and differences in the representation of structured patient data. We then assess whether the proposed modifications indeed lead to improved uniformity in data representation by comparing the results of this study with the previous study and discussing the differences in light of the induced changes.

Background

OpenSDE

OpenSDE is an application that supports clinicians with structured recording of narrative patient data [11]. The principle of OpenSDE is that clinicians can traverse a tree of predefined medical concepts and select those concepts that correspond with the relevant medical observations. The context of a concept in the tree is represented by the path leading to the concept. Per medical domain, a tree (or domain model) is created to accommodate the specific level of detail needed to describe the narratives for that domain.

When a clinician selects a node in the tree the application will display a form associated with this node alongside the tree, as shown in Figure 1. Each form presents the selected concept and the corresponding descriptors (branching nodes) of the concept [18]. Clinicians can record the relevant information at the subsequent nodes. Symptoms can be described more than once in the context of progression over time, different circumstances, or multiple occurrences. OpenSDE also supports the use of free text for particular details not covered by the content of the domain model. Furthermore, users can create custom entry forms (using an integrated form editor) to suit their individual data entry preferences.

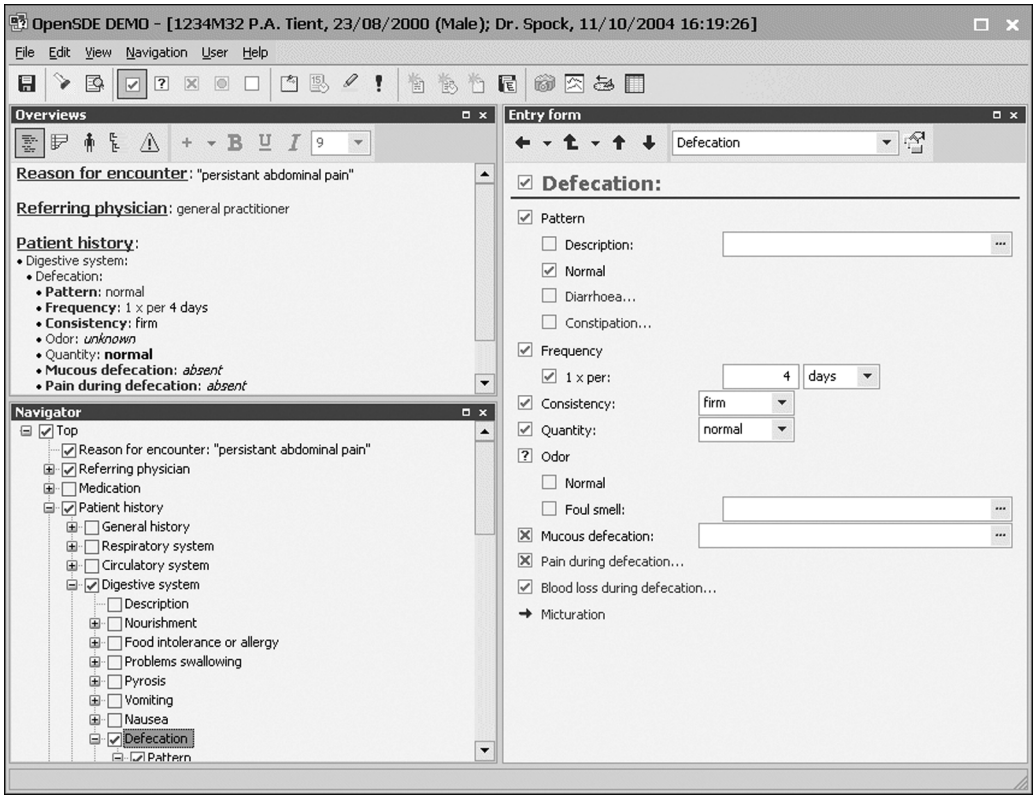

Figure 1. Screen capture of the OpenSDE data entry application. The top left of the screen shows an overview of the data recorded for the patient in the current session. The bottom left shows the domain model tree with medical concepts. On the right is the form on which data are entered. The form is associated with the selected node, in this case 'defecation'. At the bottom of the defecation form, the term 'Micturition' is preceded by an arrow which indicates that micturition is modeled in detail elsewhere. Clicking on the term will present the form used to describe micturition.

Investigating uniformity in data representation

When the purpose of SDE is to collect data suitable for both patient care and research, uniformity in data representation has a high priority. In a previous study we analyzed the uniformity of recorded data when three clinicians transcribed the same 20 paper patient records in OpenSDE [17]. In that study we found that for 17% of all findings the clinicians had chosen to represent the findings as free text instead of in a structured manner, or they had structured the same medical information at different places in the domain model. The three main reasons for these differences in representation are: a low threshold for recording data as free text, representation of value judgments, and terms missing in the domain model. Below we describe the modifications made to OpenSDE in order to reduce the differences in data representation.

Modifications

Although the pediatric domain model was tested prior to use in this study [19], the use of the model by different clinicians to describe the same findings, revealed that the domain model itself did not yet lead to uniform representation of the same findings. We made three modifications to increase uniformity in data entry.

The first modification was to reduce the ease of recording data as free text. Whereas previously it was possible to record free text at every node in a path, we reduced this to a maximum of two nodes per path[1].

To record free text, the clinicians now have to navigate to these specific nodes. By limiting the number of places where free text comments can be added we aim to reduce scattering of free text throughout the domain model, and hence increase the predictability of the places where free text can be expected.

[1]For data required as free text, or not required in a structured manner (e.g. names of siblings) we still support predefined free text fields at the last (deepest) node in the path.

The second modification involved representation of value judgments. In the paper records, the clinicians often use value judgments such as normal or abnormal. Although these terms were predefined in the pediatric domain model, the judgments are intermingled with objective data. Judgments such as 'normal' are ideally separated from descriptive concepts such as 'lively bowel sounds' in the domain models. Separation means not placing such concepts in the same path. Figure 2 illustrates how not to model judgmental concepts (Figure 2a) and how we propose to model judgmental concepts (Figure 2b).

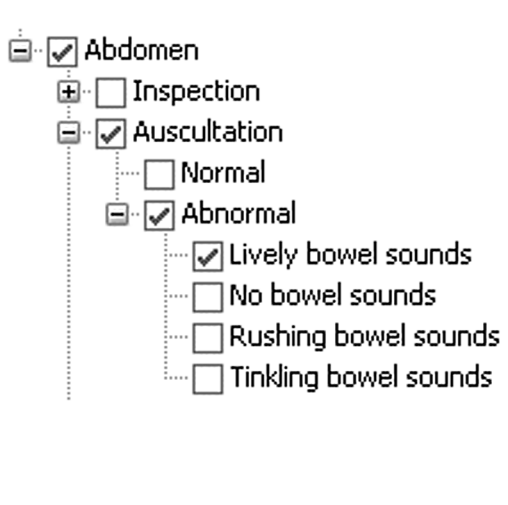

Figure 2a.

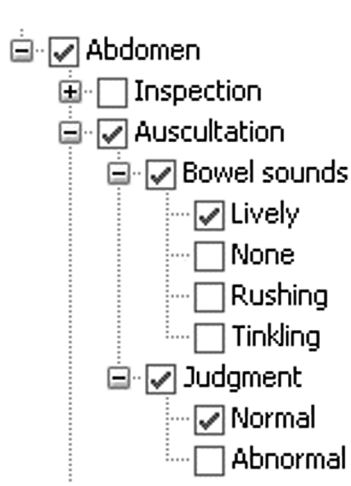

Figure 2b.

Figure 2 represents an extract of the domain model used for pediatrics. **Figure 2a** shows that modeling findings in the same path as judgmental concepts such as abnormal can lead to erroneous data. **Figure 2b** illustrates a different way to model judgmental concepts, namely not in the same path but parallel to the findings. This allows for structured representation of the (objective) finding, as well as the addition of a judgment.

The third modification was an expansion of the domain model with additional terms. A number of terms had not been modeled in the domain model whilst the terms frequently appeared in the paper record. As a result, during the transcription process, clinicians represented the finding (as free text) at other nodes in the tree or did not record the finding at all.

For example, a patient's breathing could be described by terms such as 'breath sounds normal', but the term 'vesicular breath sounds', which is frequently used on paper, had not been modeled in the domain model.

Methods

A. Data entry from a common data source

This study, assessing the impact of the modifications, took place 18 months after the initial study. We approached the same three clinicians that participated in the first study and they volunteered to participate in this second study. The clinicians again received training in the use of (the changed) OpenSDE prior to transcribing the same 20 paper patient records, again creating a data set of 60 transcribed records.

We randomly selected the 20 handwritten paper patient records created for first-contact patients at the pediatric outpatient department. We chose first-contact patients as intake and physical examination data are written on semi-structured forms. The paper records are thus comparable in format, degree of detail, and in amount of data content. The clinicians were informed about the goal of the study, and knew that the transcribed records would be analyzed.

B. Consensus and differences

In line with the first study, our main interest in this study is the consensus and differences in the representation of structured patient data. We identified six categories of consensus and differences between the transcribed findings. To classify the findings into one of the six categories we developed the algorithm described below and presented in Figure 3. This is the same algorithm used to classify the findings in the first study.

For every finding we analyzed how it was represented by each of the three clinicians. If the finding was represented in exactly the same manner in all three transcribed records, the finding was classified as *identical*. If there was a difference in at least one of the transcribed records, we searched through that entire transcribed record to establish whether the finding was recorded elsewhere. If the finding was represented in a structured manner elsewhere, the finding was classified as *structured differently*. If the finding was recorded as free text, at the same place or elsewhere in the tree, the finding was classified as *free text*.

For those cases where the finding was not represented in all three transcribed records we consulted the paper record. If the finding was present in the paper record, we classified the finding as *omitted*. If, however, the description of the finding in the paper record conflicted with the description of the finding in the transcribed record, we classified the finding as *conflicting*. A last possibility is that (one of) the transcribed records contained a finding not present in the paper record. In this scenario the finding was classified as *inferred*. The classification process was repeated for all findings.

The findings were subsequently classified as normal findings (e.g. 'no cardiac murmur') or abnormal findings (e.g. 'constipation'), and were split into patient history and physical examination findings. In the discussion we compare the results of this study with the results of the previous study.

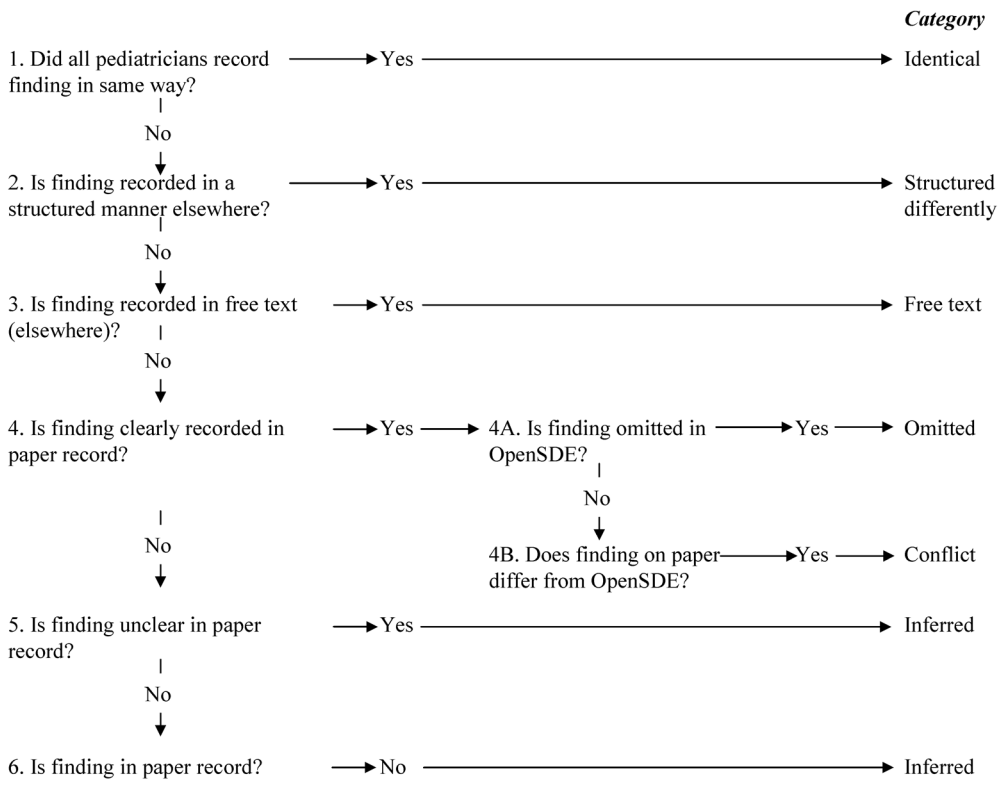

Figure 3. Algorithm used to categorize findings.

Results

Table 1 shows that a total of 1834 findings were recorded. A total of 1044 patient history and 790 physical examination findings were recorded. Of all findings, 22.5% were abnormal and 77.5% were normal findings.

Table 1. (Ab)normal findings in the patient record
The results presented in this table represent the number of findings per part of the patient record (patient history or physical examination) and per type or finding (normal or abnormal). The percentage corresponds to the percentage of the total number of findings.

Part of patient record / Type of finding	Patient History	Physical Examination	Entire Record
Normal Findings	**687** (37.5%)	**743** (40.0%)	**1421** (77.5%)
Abnormal Findings	**357** (19.5%)	**56** (3.1%)	**423** (22.5%)
All Findings	**1044** (56.9%)	**790** (43.1%)	**1834** (100%)

In Table 2 we present the findings as classified (according to the algorithm) per category of consensus/difference. The last column holds the total number of findings per category. The largest category is the omitted category which contains findings represented by at least one clinician but omitted by at least one other clinician, followed by the identical category which comprises findings represented identically by all three clinicians.

Table 2: Consensus and differences in representation of findings
In this table the findings have been ordered per category of consensus/difference. Per category the findings are subdivided into patient history and physical examination, and normal and abnormal findings. The percentage behind the numbers corresponds to the percentage of findings in that row.

Finding	Patient History		Physical Examination		ALL FINDINGS
Category	Normal	Abnormal	Normal	Abnormal	
Identical	129 (25.3%)	69 (13.6%)	300 (58.9%)	11 (2.2%)	**509** (27.8%)
Structured differently	23 (24.2%)	17 (17.9%)	51 (53.7%)	4 (4.2%)	**95** (5.2%)
Free text	38 (35.2%)	42 (38.9%)	13 (12%)	15 (13.9%)	**108** (5.9%)
Inferred	138 (33.7%)	58 (14.1%)	205 (50%)	9 (2.2%)	**410** (22.4%)
Omitted	314 (51.6%)	148 (24.3%)	133 (21.8%)	14 (2.3%)	**609** (33.2%)
Conflicting	45 (43.7%)	23 (22.3%)	32 (31.1%)	3 (2.9%)	**103** (5.6%)
ALL FINDINGS	***687*** *(37.5%)*	***357*** *(19.5%)*	***734*** *(40.0%)*	***56*** *(3.1%)*	***1834*** *(100%)*

In the last column of Table 2 we see that 509 findings were recorded identically. The majority of the findings recorded identically are normal physical examination findings (n=300), followed by normal patient history findings which make up a quarter of all identically recorded findings. The abnormal findings constitute just over 15% of all identical findings, of which the majority are patient history findings.

Appendix 1 shows the results of the initial study; the findings are also classified per category of consensus/difference [17].

Discussion

OpenSDE is intended to support structured recording of narrative patient data for use in both routine care and clinical research. Structured data entry offers the possibility to improve the quality of data [20] and standardize data collection [21]. Structured data are only beneficial if the same data are predominantly recorded and structured in the same manner; unpredictable data representation makes data retrieval, and subsequent use of data, complex.

The results of this study show that the largest category of findings is the omitted category, which contains findings not recorded by one or two clinician(s). Normal patient history findings (such as normal eating or drinking pattern) are the largest omitted group; only a small percentage of the omitted findings are abnormal physical examination findings.

The second largest category is the identical findings category, i.e. findings recorded identically by all three clinicians. Circa 60% of the identical findings are normal physical examination findings.

Inferred findings (findings represented in OpenSDE but not present in the paper record) compose almost one quarter of all findings, but are primarily normal findings. Inferred normal findings include findings such as 'normal urinary pattern', 'healthy nutritional intake', and 'normal aspect of the tonsils'.

Almost 6% of all findings are recorded as free text instead of structured. Nearly three quarters of these findings are patient history findings; the findings vary from descriptions of convulsions, to abnormal growth curves.

More than half of the findings structured differently are normal physical examination findings, such as auscultation of the lungs.

The last and smallest category of difference is the conflicting findings category. Two thirds of the conflicting findings are patient history findings, such as throat ache with/without swallowing problems and previous illnesses.

When comparing the first study with this second study we observed an increase in findings structured identically as well as a slight increase in the number of findings that were structured differently. These increases can both be explained

in light of the expansion of the domain model with terms that better correspond to the terms used in the paper records. The pediatric domain model was expanded with terms frequently used in the paper record, such as 'vesicular breath sounds'. Adding terms that appear in the paper record to the domain model makes mapping from paper to OpenSDE easier, thus increasing the number of findings recorded identically. In this study we also observed a decrease in the number of inferred findings, as a direct mapping can now be made from the paper record to the corresponding terms in OpenSDE. Adding terms, however, also means that there are more concepts that can be used to describe the same medical finding, which may have caused the slight increase in findings structured differently.

In this study we also observed a decrease in findings recorded as free text elsewhere. By reducing the number of places where free text can be recorded we have reduced the scattering of free text throughout the domain model. However, the reduction in free text comes paired with an increase in the number of omitted findings. Limiting where free text can be recorded thus leads to an increased uniformity but reduces the information content in the medical record. Therefore, free text should not be fully replaced with SDE, but rather be combined with SDE [7, 22].

The reduction in the number of inferred findings can also be explained in view of the new approach of modeling value judgments. In the pediatric domain model, value judgments such as 'normal' or 'abnormal' preceded observations in the same path. The domain model thus forces the user to judge whether particular findings are normal or abnormal. By separating the value judgments from the actual observations, we have reduced the number of inferred findings.

Conclusion

In this study we investigated whether modifications made to OpenSDE increased uniformity in data representation. The modifications were targeted particularly at those findings that were represented in OpenSDE but that were represented differently by the clinicians. The results of this study show that the percentage of findings that has been structured identically increased, and the number of findings recorded

as free text decreased. This study also pointed out the following.

Patient history data are generally more subjective than physical examination data. In their study assessing completeness and accuracy of computer medical records [23], Pringle et al. conclude that subjective data are less consistently recorded. Peat et al. [24] reach a similar conclusion namely that subjective data leads to higher inter- and intra-observer variability. The results of our uniformity studies imply that our SDE solution is more suitable for physical examination than for patient history. The need for prose to narrate a patient's story is higher for patient history than for the physical examination.

For purposes such as clinical research, increasing uniformity of patient history data is important. As general pediatrics is a very broad domain in which many aspects can be described, the domain model quickly becomes large which can make data entry a more complex task. Besides that, the more options you offer for data entry, the higher the chance that findings will be described in different manners. Given only the small increase in uniformity as a result of our modifications, the question that remains is: what is the best way to record patient history data?

An important lesson that this study taught us, is that supporting data entry in terms of completeness remains a challenge. Clinicians may only be prepared to record a limited amount of information in a structured manner, which must be taken into consideration when creating domain models.

Another important lesson is that it is difficult to propose which data must be structured and which terms must be offered for data entry, prior to knowing the purpose for which data will be used. Clinical research will often require data that have a high granularity and that are recorded uniformly, which will not always correspond to the format in which data are recorded for patient care. Completeness, accuracy, and required uniformity of data, therefore, remain functions of the use of the data [25, 26].

Acknowledgements

The work presented in this paper is funded by a grant from the Netherlands Organization for Health Research and Development (ZonMW).

Appendix 1

Table 3. Consensus and differences in representation of findings (INITIAL STUDY).
In this table the findings have been ordered per category of consensus/difference. Per category the findings are subdivided into patient history and physical examination, and normal and abnormal findings. The percentage behind the numbers corresponds to the percentage of the total number of findings.

Finding	**Patient History**		**Physical Examination**		**ALL FINDINGS**
Category	Normal	Abnormal	Normal	Abnormal	
Identical	90 (5.1%)	79 (4.5%)	198 (11.2%)	21 (1.2%)	**388** (22%)
Structured differently	28 (1.6%)	23 (1.3%)	32 (1.8%)	4 (0.2%)	**87** (4.9%)
Free text	68 (3.9%)	117 (6.6%)	26 (1.5%)	4 (0.2%)	**215** (12.2%)
Inferred	98 (5.6%)	30 (1.7%)	411 (23.3%)	9 (0.5%)	**548** (31.1%)
Omitted	188 (10.7%)	90 (5.1%)	184 (10.4%)	9 (0.5%)	**471** (26.7%)
Conflicting	23 (1.3%)	12 (0.7%)	16 (0.9%)	4 (0.2%)	**55** (3.1%)
ALL FINDINGS	***495*** *(28.1%)*	***351*** *(19.9%)*	***867*** *(49.1%)*	***51*** *(2.9%)*	***1764*** *(100%)*

References

1. Klar R. Selected impressions on the beginning of the electronic medical record and patient information. Methods Inf Med 2004;43(5):537-42.

2. Tange H. How to approach the structuring of the medical record? Towards a model for flexible access to free text medical data. Int J Biomed Comput 1996;42(1-2):27-34.

3. Walsh SH. The clinician's perspective on electronic health records and how they can affect patient care. BMJ 2004;328(7449):1184-7.

4. Tange HJ, Hasman A, de Vries Robbe PF, Schouten HC. Medical narratives in electronic medical records. Int J Med Inf 1997;46(1):7-29.

5. Bernauer J. Conceptual Graphs as an Operational Model for Descriptive Findings. In: Proc Annu Symp Comput Appl Med Care; 1992; 1992. p. 214-18.

6. Dick RS, Steen EB, Detmer DE, eds. The computer-based patient record: an essential technology for health care. Revised Edition ed. Washington: National Academy Press; 1997.

7. van Ginneken AM. The computerized patient record: balancing effort and benefit. Int J Med Inf 2002;65(2):97-119.

8. Powsner SM, Wyatt JC, Wright P. Opportunities for and challenges of computerisation. Lancet 1998;352(9140):1617-22.

9. Brown PJ, Sönksen P. Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. J Am Med Inform Assoc 2000;7(4):392-403.

10. Moorman PW. Towards formal medical reporting. An evaluation in endoscopy. Rotterdam: Erasmus University; 1995.

11. Los RK, van Ginneken AM, van der Lei J. OpenSDE: A strategy for expressive and flexible structured data entry. Int J Med Inform 2005;74(6):481-90.

12. OpenSDE. OpenSDE (OSS). http://sourceforge.net/projects/opensde. Last accessed: August 24, 2005

13. Beard CM, Yunginger JW, Reed CE, O'Connell EJ, Silverstein MD. Interobserver variability in medical record review: an epidemiological study of asthma. J Clin Epidemiol 1992;45(9):1013-20.

14. Williams JG. Measuring the completeness and currency of codified clinical information. Methods Inf Med 2003;42(4):482-8.

15. Mikkelsen G, Aasly J. Consequences of impaired data quality on information retrieval in electronic patient records. Int J Med Inform 2005;74(5):387-94.

16. Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository. J Am Med Inform Assoc 2000;7(1):42-54.

17. Los RK, Roukema J, van Ginneken AM, de Wilde M, van der Lei J. Are structured data structured identically? Investigating the uniformity of pediatric patient data recorded using OpenSDE. Methods Inf Med 2005.

18. van Ginneken AM, Verkoijen MJ. A multi-disciplinary approach to a user interface for structured data entry. Medinfo 2001;10(Pt 1):693-7.

19. Roukema J, van Ginneken AM, Moll HA. The use of structured data entry in the outpatient's clinic for paediatrics. Health Information Developments in the Netherlands 2003;6:27-30.

20. Moorman PW, van Ginneken AM, van der Lei J, van Bemmel JH. A model for structured data entry based on explicit descriptional knowledge. Methods Inf Med 1994;33(5):454-63.

21. Kahn CE, Jr., Huynh PN. Knowledge representation for platform-independent structured reporting. Proc AMIA Annu Fall Symp 1996:478-82.

22. McDonald CJ. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc 1997;4(3):213-21.

23. Pringle M, Ward P, Chilvers C. Assessment of the completeness and accuracy of computer medical records in four practices committed to recording data on computer. Br J Gen Pract 1995;45(399):537-41.

24. Peat G, Wood L, Wilkie R, Thomas E. How reliable is structured clinical history-taking in older adults with knee problems? Inter- and intraobserver variability of the KNE-SCI. J Clin Epidemiol 2003;56(11):1030-7.

25. Winthereik BR. "We fill in our working understanding": on codes, classifications and the production of accurate data. Methods Inf Med 2003;42(4):489-96.

26. Rubenfeld GD. Using computerized medical databases to measure and to improve the quality of intensive care. J Crit Care 2004;19(4):248-56.

7

Summary

'The statement of present illness and the progress notes, usually related in an unstructured manner, are the portions of the medical record that present the greatest difficulty in computerization.'

Lawrence L. Weed, 1968[1].

Introduction

The aim of this research is to investigate structured uniform representation of medical narratives to enable use of these data in both patient care and clinical research. The underlying assumption is that data represented in a structured format can be used for multiple purposes. In *Chapter one* we introduce our research by focusing on medical narrative data and our approach for structuring these data. We briefly describe our Structured Data Entry (SDE) application, called OpenSDE, which supports structured recording of medical narrative data. Subsequently, we introduce the main research which consists of two parts. In the first part we focus on supporting data entry, storage, and retrieval. In the second part we concentrate on uniformity of data representation in OpenSDE applied to the domain of general pediatrics.

Part 1: Supporting Data Entry, Storage, and Retrieval

Data Entry

In *Chapter two* we describe the possibilities for data entry in OpenSDE. OpenSDE is our open source application for structured recording of narrative sections of the patient record. The challenge is to approach the expressive power of free text, whilst keeping SDE acceptable for clinicians.

[1]Weed LL. Medical Records, Medical Education, and Patient Care. Cleveland, Ohio: The Press of Case Western Reserve University; 1969. Page 112

The principle of OpenSDE is that clinicians traverse a tree of predefined medical concepts ('domain model') and select those concepts that correspond with the relevant medical observations. The content of the tree is domain specific; trees are modeled by medical experts using a specific editor.

Clinicians can select a node in the tree, and the application will display a data entry form, associated with this node. Each form presents the selected concept and the corresponding descriptors (branching nodes) of the concept. For the concepts presented on the entry form, users may indicate whether or not the concept applies (present, absent, or unknown) or, when relevant, record a specific value (numerical, temporal, or free text). Symptoms can be described more than once in the context of progression over time, different circumstances, or multiple occurrences. OpenSDE also supports the use of free text for particular details not covered by the content of the domain model.

Data Storage

The medical narrative consists primarily of descriptive information including temporal, numerical, and free text values. The content of the medical narrative varies strongly per medical domain, yet the data are preferably shareable for multi-disciplinary care. These characteristics of the narrative require a data structure that can accommodate data that are similar in type but different in content.

The data recorded with OpenSDE are stored, as is described in *Chapter three*, using an extended row-modeled approach. Row modeling in essence involves a column-to-row transformation: the attributes (or column headings) of the conventional column-modeled table are stored as data in the row-modeled table. A column-modeled table contains a column for every attribute. A row-modeled table contains *one* column that holds all attributes and *one* column that holds the values of the attributes.

To enable the recording of expressive medical narrative data, we chose to extend the number of value columns in the row-modeled table. The extensions we made to the row model fall in two categories. The first category deals with data types. Whilst other researchers introduce different tables to

deal with different data types such as numerical or free text values, OpenSDE extends the row model with additional columns to reflect the data type. The second category deals with the complexity of the medical narrative (e.g., repeated descriptions over time of multiple lesions). OpenSDE extends the row model to represent descriptions of (multiple) occurrences of findings or symptoms that may progress over time.

The main advantage of row modeling is that the metadata needed for data interpretation and interface generation are separated from the actual physical data. In our case the metadata are represented in the form of the domain models. The advantage of this separation is that domain models can be altered or updated with new nodes without having to change the physical structure of the table in which the data are stored. The entry forms are generated runtime so changes to the domain model do not require adaptation of standard entry forms.

Row modeling does have disadvantages. The separation of metadata from the actual data, combined with the transformed data representation, makes data extraction and subsequent use less straightforward. The first step in investigating the feasibility of using the data recorded with OpenSDE for research purposes is, therefore, to investigate whether it is technically possible to extract and represent the data in a format suitable for clinical research.

Data Retrieval

In *Chapter four* we present our solution for facilitating data extraction and research. The use of row-modeling to store the data implies that both data and metadata are not stored in the conventional approach. To support research on the data we chose to export and convert the row-modeled data to a conventional relational format suitable for querying with conventional analysis software. The exported data have as advantages that analysis can be performed with available tools often known by researchers, and does not require development of functionality that already exists. However, this does require reintegrating the semantic information with the actual data. The main challenge is to develop a method that permits selection of data in the

same context as data entry, and that transforms conceptually hierarchical, row-modeled data to a conventional format without losing the important contextual information held in the hierarchy.

The method that we developed to select and transform the conceptually hierarchical data to a conventional format is an application called Entity Export. Entity Export supports selection of concepts from a domain model and converts the corresponding data to the appropriate columns in one or more newly created conventional tables suitable for analysis purposes. The original row-modeled table remains intact; Entity Export duplicates the data for output in conventional relational tables.

Entity Export has been tested with a straightforward and well-defined data set in the combined domain of radiology and neurology. So far, the results of Entity Export are promising: data extraction and representation in a conventional format is *technically* possible.

Having technically enabled data extraction we continue in *Chapter four* to describe the semantic barriers faced when extracting and interpreting data.

There is one feature of OpenSDE that needs consideration when extracting and interpreting data: OpenSDE does not infer data that are not explicitly recorded. In Chapter four we provide two examples of scenarios in which this potentially causes problems for subsequent use of data. In short, complete querying requires querying both explicitly recorded data as well as implicit data, that is, data that are implied by recording other data. For example, when a node in a tree is marked absent, all children of this node are absent as well, although this is not explicitly recorded (OpenSDE does not allow recording data after a node that has been marked absent).

In part two of this research we focus on the challenges that remain on a semantic level.

Part 2: Uniformity in Data Representation

Reliability and accuracy of collected data are pivotal if data will be collected over long periods of time and by different users. Although OpenSDE supports structured data entry, the actual concordance in data representation has not yet been explored. In the second part of this research we, therefore, *focus on pitfalls for data extraction for research purposes, and aim to formulate strategies to improve uniformity in data entry*

to enhance the reliability of data retrieval.

In collaboration with our hospital's pediatric department, we analyzed the uniformity of recorded data when OpenSDE is used to transcribe data from the same data source. In *Chapter five* we describe the results of our analysis. Of interest in this qualitative analysis is whether recording data using OpenSDE by definition leads to uniformly structured data.

Three clinicians transcribed 20 first-contact paper pediatric patient records in OpenSDE, creating a data set of 60 transcribed records. The paper records consist of semi-structured handwritten forms and are comparable in content and level of detail.

The results of this study show that recording data using structured data entry does not necessarily lead to uniformly structured data. Analysis of the recorded findings shows that only 22% of all findings were recorded identically by all three clinicians; in more than three quarters of the findings there were differences in data representation or data content. Data that vary in representation can be recorded in a different structured manner, which can occur when two semantic variations are offered to describe the same medical finding, or the data can be recorded as free text. Data that vary in content are either omitted by one or two clinicians, inferred (i.e. not in the paper record), or conflicting with the paper record.

Our results show that clinicians use the free text possibility more frequently for patient history than for physical examination findings. Furthermore, physical examination findings are recorded identically more often than patient history findings. Our results suggest that clinicians are more inclined to omit normal findings than abnormal findings.

This study emphasizes that for data lookup and retrieval one must be aware of all possible ways in which an item of information may have been recorded.

Differences in recording practices between clinicians are inevitable. Hence, achieving complete uniformity in data representation was not our aim. However, we did expect that the data that were transcribed in OpenSDE by all three clinicians would have a higher degree of uniformity. We had not expected to see 17% of all findings represented differently, either structured differently or recorded as free text. Therefore, in *Chapter six* we analyze those findings that were represented differently to establish whether we could modify

OpenSDE to improve uniformity.

In hindsight, the differences in representation can be explained in light of initial design decisions of OpenSDE. One of the main challenges for OpenSDE was to approach the expressive power and flexibility of free text, whilst keeping SDE acceptable for clinicians. To support this challenge we made several decisions with respect to the design of the application. The two design decisions that have the largest impact on uniformity of data representation concern supporting flexibility and expressiveness.

To support *flexibility* in SDE, OpenSDE firstly does not enforce a specific detail-level or structure in which findings must be described, and secondly enables use of free text where needed. This design has led to a very flexible use of OpenSDE; our results illustrate that clinicians use free text to represent findings that can be structured. Clinicians also record the same data as free text at different places in the tree. Flexibility is an advantage for data entry, but it is a hurdle for data look-up and extraction as data can be recorded at more than one place (making data representation less predictable).

To support *expressiveness*, OpenSDE offers predefined terms that can be used to describe medical findings. A domain model is, however, limited in scope and may not always contain the exact terms that clinicians would like to use, or may not present terms in the exact context in which clinicians would preferably use the terms. Transcribing findings thus involves interpreting the finding in the paper record and then translating or mapping the finding to those concepts in OpenSDE that best match the description of the finding in the paper record.

The analysis described in *Chapter six* illustrates that providing freedom in data entry, will lead to use of this freedom during data entry. This freedom, therefore, is in conflict with uniform data representation. In this chapter we ask ourselves the question whether we should retain focus on facilitating data entry in order to get as many clinicians using the application as possible (hence sacrificing the uniformity of the collected data) or whether the uniformity of the data set should be our priority (thus sacrificing freedom in data entry).

Based on the analysis of the differences in data representation, we proposed modifications to increase uniformity in data entry whilst maintaining flexibility in data entry. The mod-

ifications include: reducing the ease of recording data as free text, separating observations from value judgments, and expanding the domain model with additional terms to increase the coverage of the domain model and to correspond more to the terms used in the paper records. In *Chapter seven* we describe these modifications in more detail. Subsequently, we describe the study performed to evaluate whether these modifications increase uniformity.

To evaluate whether the modifications increase uniformity in data representation we approached the same three clinicians that participated in the initial study. These clinicians transcribed the same 20 paper pediatric patient records in the modified OpenSDE, creating a data set of 60 transcribed records. We analyzed the results in the same manner as in the first study described in Chapter five.

When comparing the first study with this second study we observed an increase in findings structured identically as well as a slight increase in the number of findings that were structured differently. These increases can both be explained in light of the expansion of the domain model with terms that better correspond to the terms used in the paper records. In this study we also observed a decrease in the number of inferred findings, as a direct mapping can now be made from the paper record to the corresponding terms in OpenSDE.

In this second study we also observed a decrease in findings recorded as free text elsewhere. The reduction in free text has, however, come paired with an increase in the number of omitted findings. Limiting free text thus leads to an increased uniformity but reduces the information content in the medical record.

An important lesson that this study taught us, is that supporting data entry in terms of completeness remains a challenge. Clinicians may only be prepared to record a limited amount of information in a structured manner, which must be taken into consideration when creating domain models.

Another important lesson learned is that it is difficult to propose which data must be structured and which terms must be offered for data entry, prior to knowing the purpose for which data will be used. Clinical research will often require data that have a high granularity and that are recorded uniformly, which will not always correspond to the format in which data are recorded for patient care. Completeness,

accuracy, and required uniformity of data, therefore, remain functions of the use of the data.

This study has provided us with insight into those aspects that can be influenced with structured data entry and those aspects which cannot be influenced. What cannot be influenced are interpretations made by clinicians. What we can influence, on the other hand, is the number of ways in which the same medical information can be represented in a structured format.

The need for comparable patient information in hospitals, to serve purposes as managed care and outcome research was already articulated by Florence Nightingale over a century ago[2]. Since then, so many others have expressed similar needs. But as Larry Weed already concluded in 1968, medical narratives present the greatest difficulty in computerization. However, Weed also stated that "*all narrative data presently in the medical record can be structured, […] guaranteeing a thoroughness, retrievability, efficiency and economy important to the scientific analysis*" but that this has *"hitherto been handled in a very unrigorous manner"*[3]. In this research, we have approached the structuring of medical narrative data in a rigorous manner. Nevertheless, we must conclude that structuring narrative data does not per se guarantee a thoroughness and retrievability of routinely recorded clinical data for subsequent use in clinical research.

[2]Tang PC, LaRosa MP, Gorden SM. Use of computer-based records, completeness of documentation, and appropriateness of documented clinical decisions. J Am Med Inform Assoc 1999;6(3):245-51.

[3]Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(11):593-600.

Dutch Summary (Samenvatting)

'The statement of present illness and the progress notes, usually related in an unstructured manner, are the portions of the medical record that present the greatest difficulty in computerization.'

Lawrence L. Weed, 1968[1].

Inleiding

De nadruk van dit onderzoek ligt op het uniform structureren van medische verslaglegging voor gebruik in zowel patiëntenzorg als wetenschappelijk onderzoek. De onderliggende aanname is dat gegevens die op een gestructureerde manier gerepresenteerd worden, gebruikt kunnen worden voor meerdere doeleinden. In *Hoofdstuk één* introduceren we ons onderzoek met een beschrijving van het soort gegevens waar het om draait, namelijk verslaglegging van anamnese en lichamelijk onderzoek, en onze aanpak om deze gegevens te structureren. We beschrijven kort onze toepassing voor gestructureerde gegevensinvoer (Structured Data Entry SDE), genaamd OpenSDE. OpenSDE ondersteunt het gestructureerd vastleggen van medische verslaglegging. Het daadwerkelijke onderzoek bestaat uit twee delen. In het eerste deel concentreren we ons op het ondersteunen van gegevensinvoer, gegevensopslag en gegevensontsluiting. In het tweede deel ligt de nadruk op uniformiteit in gegevensrepresentatie in OpenSDE, toegepast in de algemene kindergeneeskunde.

[1]Weed LL. Medical Records, Medical Education, and Patient Care. Cleveland, Ohio: The Press of Case Western Reserve University; 1969. Pagina 112.

Deel 1: Ondersteuning van Gegevensinvoer, Gegevensopslag en Gegevensontsluiting

Gegevensinvoer

In *Hoofdstuk twee* beschrijven we de mogelijkheden voor gegevensinvoer in OpenSDE. OpenSDE is de door ons ontwikkelde open source applicatie voor het gestructureerd invoeren van de vrije tekst delen van het medisch dossier. De uitdaging is om op gestructureerde wijze de uitdrukkingskracht van vrije tekst te benaderen terwijl gegevensinvoer wel acceptabel blijft voor clinici. Het principe van OpenSDE is dat clinici een boom, bestaande uit medische begrippen ('domein model'), doorlopen en díe begrippen selecteren die overeenkomen met de relevante medische bevindingen.

De inhoud van de boom is specialisme specifiek; medische experts modelleren deze bomen met behulp van een specifieke applicatie.

Clinici kunnen een knoop in de boom selecteren waarna OpenSDE een invulformulier genereert dat geassocieerd is met deze knoop. Elk invulformulier presenteert het geselecteerde begrip en de bijbehorende descriptoren (takken) van het begrip. De begrippen die op de invulformulieren gepresenteerd worden, kunnen door de gebruiker als aanwezig, afwezig of onbekend worden aangevinkt, of, indien van toepassing, van een waarde (datum/tijd, numeriek of vrije tekst) worden voorzien. Symptomen kunnen meermaals beschreven worden vanwege mogelijke verandering in de tijd, andere omstandigheden of meervoudig voorkomen. OpenSDE ondersteunt ook het gebruik van vrije tekst om díe specifieke details vast te kunnen leggen waarvan de lading niet gedekt wordt door het domein model.

Gegevensopslag

Medische verslaglegging bestaat voornamelijk uit beschrijvende informatie, inclusief temporele, numerieke en vrije tekst gegevens. De inhoud van een verslag varieert sterk per medisch specialisme, terwijl de gegevens idealiter uitwissel-

baar zijn voor multi-disciplinaire samenwerking. De uitdaging voor gegevensopslag is een gegevensstructuur waarin gegevens die vergelijkbaar zijn in soort, onafhankelijk van hun inhoud weergegeven kunnen worden.

Zoals wordt omschreven in *Hoofdstuk drie* maakt OpenSDE gebruik van een uitgebreide rij modellering aanpak om gegevens op te slaan. Rij modellering behelst in essentie een kolom-naar-rij transformatie: de attributen (of kolomnamen) van een conventionele kolom-georiënteerde tabel worden opgeslagen als data in de rij-georiënteerde tabel. Een kolom-georiënteerde tabel bevat één kolom voor ieder attribuut. Een rij-georiënteerde tabel bevat één kolom waarin alle attributen worden opgeslagen en één kolom waarin de corresponderende waarden worden opgeslagen.

Om het opslaan van de gedetailleerde verslagleggінggegevens te ondersteunen, hebben wij het aantal kolommen in de rij-georiënteerde tabel uitgebreid om zo op uniforme wijze meer detail vast te kunnen leggen. Feitelijk zijn er twee soorten uitbreidingen. De eerste soort uitbreiding behelst data types. Terwijl andere onderzoekers nieuwe tabellen toevoegen om verschillende data types, zoals numerieke of vrije tekst gegevens, op te slaan, is voor OpenSDE het rij model uitgebreid met additionele kolommen om het data type weer te geven. De tweede soort uitbreiding is het gevolg van de complexiteit van de verslaglegging (e.g., meerdere beschrijvingen in de tijd van meerdere laesies). Voor OpenSDE is het rij model uitgebreid om meervoudige beschrijvingen van (meerdere) bevindingen, of het verloop van symptomen in de tijd, te kunnen beschrijven.

De voornaamste eigenschap van rij modellering is dat de metadata die nodig zijn voor data interpretatie en het genereren van de interface, gescheiden zijn van de fysieke data. In het geval van OpenSDE wordt de metadata weergegeven in de vorm van de domein modellen. Het voordeel van deze scheiding is dat domein modellen veranderd of uitgebreid kunnen worden met nieuwe knopen zonder dat dit aanpassing vereist van de fysieke data structuur van de tabellen waarin de gegevens worden opgeslagen. De standaard invoerformulieren worden ad-hoc gegenereerd waardoor veranderingen aan de domein modellen geen aanpassingen vereisen aan deze invoerformulieren.

Rij modellering kent ook nadelen. Het scheiden van de metadata van de fysieke data, in combinatie met de getransformeerde data opslagstructuur, maakt gegevensontsluiting en daaropvolgend gebruik van de gegevens, minder

eenvoudig. De eerste stap in het analyseren van de haalbaarheid van het gebruiken van de met OpenSDE verzamelde gegevens voor wetenschappelijk onderzoek is, daarom, het onderzoeken of het technisch mogelijk is om de gegevens te ontsluiten en weer te geven op een manier geschikt voor analyse bij medisch wetenschappelijk onderzoek.

Gegevensontsluiting

In *Hoofdstuk vier* presenteren we onze oplossing voor het faciliteren van gegevensontsluiting voor onderzoeksdoeleinden. Het gebruik van rij modellering om de gegevens op te slaan impliceert dat de gegevens op onconventionele wijze opgeslagen worden. Om onderzoek op de gegevens te ondersteunen hebben we gekozen om de gegevens te exporteren en converteren naar een relationeel data formaat geschikt voor analyse met conventionele analyse software. Het exporteren van gegevens heeft als voordeel dat de analyses uitgevoerd kunnen worden met reeds beschikbare programma's die bekend zijn bij de onderzoekers, en geen ontwikkeling van nieuwe analyse tools vereist. Echter, dit houdt wel in dat de semantische informatie die in de metadata zit weer geïntegreerd moet worden met de abstracte opgeslagen gegevens. De voornaamste uitdaging is het ontwikkelen van een methode, waarbij gegevens geselecteerd worden in dezelfde context als waarbinnen ze ingevoerd zijn, en die vervolgens de rij-georiënteerde gegevens transformeert naar een conventioneel formaat zonder de belangrijke contextuele informatie die in de hiërarchie verscholen zit, te verliezen.

Het door ons ontwikkelde Entity Export is een methode die gegevensselectie en transformatie van conceptueel hiërarchische gegevens naar een conventioneel formaat ondersteunt. Entity Export ondersteunt selectie van begrippen uit domein modellen en zorgt voor conversie van corresponderende gegevens naar passende kolommen in één of meerdere nieuw aangemaakte tabellen geschikt voor analyse doeleinden. De originele rij-georiënteerde tabel blijft intact; Entity Export dupliceert de gegevens voor ontsluiting naar conventionele relationele tabellen.

Entity Export is getest met een eenvoudige en goed afgebakende gegevensset verzameld voor een studie van de afdelingen radiologie en neurologie. Tot op heden zijn de met Entity Export behaalde resultaten hoopvol: gegevensontsluiting en representatie in een conventioneel formaat is technisch mogelijk.

Nu gegevensontsluiting technisch mogelijk is, gaan we in Hoofdstuk vier verder in op de semantische barrières waar men tegen aanloopt bij het ontsluiten en vervolgens interpreteren van de gegevens.

Er is één eigenschap van OpenSDE waarmee rekening gehouden moet worden tijdens gegevensextractie en interpretatie: OpenSDE infereert geen gegevens die niet expliciet zijn vastgelegd. In Hoofdstuk vier laten we aan de hand van twee voorbeelden zien hoe deze eigenschap van OpenSDE potentieel tot problemen zou kunnen leiden. In het kort komt het erop neer dat het volledig ontsluiten van gegevens vereist dat zowel de expliciet vastgelegde gegevens als de niet-vastgelegde impliciete gegevens ontsloten worden. Impliciete gegevens zijn gegevens die geïmpliceerd worden door het vastleggen van andere gegevens. Bijvoorbeeld: als een knoop in een boom als afwezig is aangekruist, impliceert dit dat alle kinderen van deze knoop ook afwezig zijn, ook al is dit niet expliciet vastgelegd (het is in OpenSDE niet mogelijk om gegevens vast te leggen na een knoop die als afwezig is aangekruist).

In deel twee van dit onderzoek richten we de aandacht op de uitdagingen die er nog liggen op het semantische vlak.

Deel 2: Uniformiteit in Gegevensrepresentatie

Betrouwbaarheid en nauwkeurigheid van verzamelde gegevens zijn essentieel als gegevens gedurende lange tijd en door verschillende gebruikers verzameld worden. Ook al ondersteunt OpenSDE gestructureerde gegevensinvoer, de daadwerkelijke vergelijkbaarheid in gegevensrepresentatie is nog niet onderzocht. In het tweede deel van dit onderzoek ligt, daarom, de nadruk op de *valkuilen voor gegevensontsluiting en proberen we strategieën te ontwikkelen om uniformiteit in*

gegevensinvoer te verhogen om zo de betrouwbaarheid van gegevensontsluiting te verbeteren.

In samenwerking met de afdeling algemene kindergeneeskunde van het Sophia Kinderziekenhuis hebben wij een onderzoek uitgevoerd waarin we de uniformiteit van gegevens analyseren wanneer verschillende gebruikers met OpenSDE gegevens uit eenzelfde bron vastleggen. In *Hoofdstuk vijf* beschrijven we de resultaten van deze analyse. In deze kwalitatieve analyse kijken we vooral of het vastleggen van gegevens met OpenSDE per definitie leidt tot uniform gestructureerde gegevens.

Voor dit onderzoek vertaalden drie artsen 20 papieren dossiers van eerste consulten in de kindergeneeskunde naar OpenSDE, resulterend in 60 vertaalde dossiers. De papieren dossiers bestaan uit semi-gestructureerde handgeschreven formulieren die vergelijkbaar zijn qua inhoud en detailniveau.

De resultaten van deze studie laten zien dat het vastleggen van gegevens met behulp van gestructureerde gegevensinvoer niet per se leidt tot uniform gestructureerde gegevens. Analyse van de vastgelegde bevindingen wijst uit dat slechts 22% van alle bevindingen identiek vastgelegd zijn door alle drie de artsen; bij meer dan driekwart van de vastgelegde bevindingen was er een verschil in gegevensrepresentatie of gegevensinhoud. Gegevens die verschillen in representatie kunnen door één of twee artsen als vrije tekst vastgelegd zijn of kunnen op verschillende gestructureerde manieren vastgelegd zijn, hetgeen voor kan komen als verschillende begrippen aangeboden worden om dezelfde medische bevinding te beschrijven. Gegevens die verschillen in inhoud zijn ofwel weggelaten door één of twee artsen, afgeleid (dat wil zeggen: niet aanwezig in het papieren dossier), of conflicteren met het papieren dossier.

Onze resultaten suggereren bovendien dat artsen de mogelijkheid voor het vastleggen van bevindingen in vrije tekst vaker gebruiken voor de anamnese dan voor het lichamelijk onderzoek. Bovendien zijn de bevindingen uit het lichamelijk onderzoek vaker identiek vastgelegd dan de anamnese gegevens. Onze resultaten suggereren dat artsen geneigd zijn om normale bevindingen vaker weg te laten dan abnormale bevindingen.

Dit onderzoek onderschrijft dat het, voor het terugkijken van gegevens en het ontsluiten van gegevens, noodzakelijk is om op de hoogte te zijn van alle mogelijke manieren waarop

een bevinding vastgelegd zou kunnen worden.

Verschillen in verslaglegging tussen artsen zijn onvermijdbaar. Het verkrijgen van een volledige gegevensverzameling was derhalve ook niet ons doel. Desalniettemin hadden wij verwacht dat de gegevens die vertaald zijn naar OpenSDE een hogere mate van uniformiteit zouden hebben (uniformer waren geweest). Wij hadden niet verwacht dat 17% van de bevindingen verschillend gerepresenteerd zouden zijn: of verschillend gestructureerd of als vrije tekst weergegeven. Hieruit volgt dat we in *Hoofdstuk zes* díe bevindingen analyseren die verschillend gerepresenteerd zijn om te beoordelen of we OpenSDE zo aan kunnen passen dat de applicatie uitnodigt om gegevens uniformer vast te leggen.

Achteraf beschouwd kunnen de verschillen in representatie verklaard worden aan de hand van oorspronkelijke aannames over OpenSDE. Eén van de voornaamste uitdagingen voor OpenSDE was het benaderen van de uitdrukkingskracht en flexibiliteit van vrije tekst, terwijl SDE bruikbaar blijft voor de artsen die het moeten gebruiken. Om deze uitdaging te realiseren hebben we een aantal keuzes gemaakt met betrekking tot het ontwerp van de applicatie. De twee keuzes die de grootste invloed hebben op de uniformiteit van gegevensrepresentatie betreffen het ondersteunen van flexibiliteit en uitdrukkingskracht.

Om *flexibiliteit* in SDE te ondersteunen, dwingt OpenSDE allereerst geen specifiek detailniveau af waarop bevindingen beschreven moeten worden. Ten tweede is het mogelijk om waar nodig vrije tekst te gebruiken. Dit ontwerp heeft geleid tot een variabel gebruik van OpenSDE; onze resultaten illustreren dat artsen vrije tekst gebruiken om bevindingen te beschrijven die ook gestructureerd beschreven hadden kunnen worden. Artsen leggen ook dezelfde bevindingen vast in vrije tekst op verschillende plaatsen in de boom. Flexibiliteit is een voordeel voor gegevensinvoer, maar het is een nadeel voor het terugkijken en ontsluiten van gegevens omdat gegevens op verschillende plekken vastgelegd kunnen worden (hetgeen gegevensrepresentatie minder voorspelbaar maakt).

Om *uitdrukkingkracht* te ondersteunen maakt OpenSDE het mogelijk om zeer gedetailleerde domein modellen te maken. De strekking van een domein model is echter beperkt en bevat niet per se de exacte begrippen die artsen graag zouden gebruiken, of de begrippen worden in een andere context aangeboden dan waarin een arts het begrip nor-

maal zou gebruiken. Het vertalen van de bevindingen van papier naar OpenSDE vereist daarom het interpreteren van de bevindingen in het papieren dossier en het vertalen of mappen van de bevinding naar díe begrippen in OpenSDE die het beste overeenkomen met de bevinding in het papieren dossier.

De analyse beschreven in *Hoofdstuk zes* illustreert dat het aanbieden van vrijheid in gegevensinvoer leidt tot het gebruik van deze vrijheid bij gegevensinvoer. Deze vrijheid conflicteert met uniforme gegevensrepresentatie. In dit hoofdstuk vragen wij ons af of de nadruk moet liggen op het faciliteren van gegevensinvoer om zo zoveel mogelijk artsen achter de applicatie te krijgen (waarbij de uniformiteit van de gegevens lager uitvalt) of dat de nadruk moet liggen op het waarborgen van de uniformiteit van de gegevensset (waarbij de vrijheid in invoer aangetast wordt).

Gebaseerd op de analyse van de verschillen in gegevensrepresentatie stellen wij aanpassingen voor om de uniformiteit te verhogen terwijl de flexibiliteit in invoer gewaarborgd blijft. Deze aanpassingen omvatten: het reduceren van het gemak waarmee gegevens als vrije tekst kunnen worden vastgelegd, het scheiden van observaties en waardeoordelen, en het uitbreiden van het domein model met additionele begrippen om zo de dekking van het domein model te vergroten en meer overeen te laten komen met de termen die gebruikt worden in het papieren dossier. In *Hoofdstuk zeven* beschrijven we deze aanpassingen in meer detail. Daaropvolgend gaan we in op de studie die we uitgevoerd hebben om te evalueren of de doorgevoerde aanpassingen daadwerkelijk de uniformiteit verbeteren.

Om te evalueren of de aanpassingen ook echt de uniformiteit verhogen hebben we dezelfde drie artsen uit het eerste onderzoek gevraagd om aan deze tweede evaluatie mee te werken. Deze artsen hebben dezelfde 20 papieren dossiers met de aangepaste OpenSDE vastgelegd, zodat er wederom een set van 60 vertaalde dossiers ontstond. Vervolgens hebben we de resultaten op dezelfde manier geanalyseerd als in het eerste onderzoek beschreven in Hoofdstuk vijf.

Bij het vergelijken van de eerste en de tweede studie zien we een toename in het aantal bevindingen dat identiek gestructureerd is, alsmede een beperkte toename in het aantal bevindingen dat anders gestructureerd is. Deze toenames kunnen beide verklaard worden aan de hand van de uitbreiding van het domein model zodat dit beter overeenkomt

met de termen die gebruikt worden in het papieren dossier. Deze studie laat tevens een afname zien van het aantal afgeleide bevindingen nu er een rechtstreekse mapping gemaakt kan worden van de termen in het papieren dossier naar de corresponderende begrippen in het OpenSDE domein model.

In de tweede studie zagen we een afname van het aantal bevindingen elders vastgelegd als vrije tekst. Deze afname in vrije tekst gaat echter gepaard met een toename in het aantal weggelaten bevindingen. Het beperken van vrije tekst invoer mogelijkheden heeft dus geleid tot een toegenomen uniformiteit maar het verlaagt het informatiegehalte in het medisch dossier.

Een belangrijke les die wij uit deze studie leren, is dat het ondersteunen van gegevensinvoer in termen van volledigheid een uitdaging blijft. Artsen zijn bereid of slechts in de gelegenheid om een beperkte hoeveelheid informatie op een gestructureerde wijze vast te leggen, hetgeen overwogen moet worden bij het modelleren van domein modellen.

Een andere belangrijke les die voortkomt uit dit onderzoek is dat het moeilijk is om op voorhand te bepalen welke gegevens gestructureerd moeten worden en welke begrippen daarvoor aan moeten worden geboden voor gegevensinvoer, voordat men weet voor welke doeleinden de gegevens gebruikt zullen worden. Klinisch onderzoek vereist vaak gegevens met een hoge mate van granulariteit en uniformiteit, hetgeen niet altijd overeenkomt met de manier waarop routinematig verzamelde zorggegevens vastgelegd zijn. Eisen voor volledigheid, nauwkeurigheid, en uniformiteit van gegevens blijven gebonden aan het gebruik van de gegevens.

Deze studie heeft ons inzichten geboden in díe aspecten die beïnvloed kunnen worden met gestructureerde gegevensinvoer en díe aspecten die niet beïnvloed kunnen worden. Wat niet beïnvloed kan worden zijn de interpretaties die de arts zelf maakt. Wat men wel kan beïnvloeden zijn het aantal manieren waarop dezelfde medische informatie op gestructureerde wijze vastgelegd kan worden.

De behoefte aan vergelijkbare patiënt informatie in ziekenhuizen, voor doeleinden zoals “managed care” en “outcome research” is reeds meer dan een eeuw geleden uitgesproken door Florence Nightingale[2].

[2]Tang PC, LaRosa MP, Gorden SM. Use of computer-based records, completeness of documentation, and appropriateness of documented clinical decisions. J Am Med Inform Assoc 1999;6(3):245-51.

Sindsdien hebben vele anderen vergelijkbare behoeften kenbaar gemaakt. Maar, zoals Larry Weed reeds concludeerde in 1968 is de medische verslaglegging het moeilijkst te automatiseren. Echter, Weed schreef tevens dat *"all narrative data presently in the medical record can be structured, […] guaranteeing a thoroughness, retrievability, efficiency and economy important to the scientific analysis"* , maar dat dit tot op heden op een onzorgvuldige manier is aangepakt[3]. In dit onderzoek hebben wij het structureren van medische verslaglegging rigoureus aangepakt. Desalniettemin moeten wij concluderen dat het structureren van de medische verslaglegging niet per se garandeert (al dan niet door tijdsdruk) dat routinematig verzamelde gegevens qua volledigheid en extraheerbaarheid geschikt zijn voor gebruik bij klinisch wetenschappelijk onderzoek voor het gebruik van voor toepassing.

[3]Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(11):593-600.

Acknowledgements

The research presented in this thesis could not have been completed without the following people to whom I am, therefore, greatly in debt.

Prof.dr.J. van der Lei. Johan, considering the fact that when we started this research four years ago, we really had to start at the beginning… redefining my definition of science… I think we've come quite a long way. You have a gift for putting results in a simple, but strong perspective. You've played an important role in creating a thesis of which, in hindsight, I can say I'm proud.

Dr. A.M. van Ginneken. Astrid, I know I've said it before, but I'm going to say it again: your dedication is truly inspiring. It's great to see that your dedication is paying-off now; whenever my commitment is fading, I'll think of you! Thanks for all the time and effort you spent working with me to get this thesis finished!

Dr. H.A. Moll. Henriette, thanks to committed clinicians like you, we medical informaticians might just see our dream of the Electronic Medical Record come true in our working lives. Your effort and enthusiasm for OpenSDE really sped up the progression of this research!

Marcel de Wilde. Marcel, with you, the impossible always ends up being possible in some way. Whenever you said that a particular functionality was difficult to program, a new version of the software (including the functionality) was available first thing the next morning. Thanks to your endless hard work and our lively discussions OpenSDE, Entity Export, and this thesis are where they are now.

Jolt Roukema. Thanks to your first 'ALKG' dataset the second part of this thesis ended up being what it is now. Your 'clinician perspective' was not only really helpful to the progression of OpenSDE, but also to the progression of my research. Please keep up your eagerness for computers in healthcare and good luck finishing your thesis!

Colleagues from the Department of Medical Informatics. Thanks for making the five years that I spent with you memorable.

My 'paranimfen' Ferry van Oeveren and Cobus van Wyk. Guys, thanks for standing by me on the day of the defense. Let's make it a blast!

Mam en Daddy. Dit proefschrift was er niet gekomen zonder het van jullie afgekeken doorzettingsvermogen. Jullie hebben intens meegeleefd met de totstandkoming van dit proefschrift, daarom draag ik dit proefschrift aan jullie op.

And finally, an immeasurable thank you…

… to all of you involved in whatever way with the research presented in this thesis, or with the distraction you provided ;-) It was so much easier with your help!

Curriculum Vitae

Renske Kirsten Los was born February 9th, 1978 in Roosendaal, the Netherlands. After having lived in England, Switzerland, the Netherlands, and Spain she obtained her International Baccalaureate degree in 1996 at the Rijnlands Lyceum in Oegstgeest, the Netherlands. From 1997 onwards Renske studied Medical Information Sciences at the University of Amsterdam. The obligatory internship during the last year of the study brought her to the department of Medical Informatics at the Erasmus University in Rotterdam (now Erasmus MC University Medical Center). After her graduation in 2001 Renske stayed with the department and continued working on her research on structured data extraction.

In 2004 Renske received a Master's Degree in Healthcare Management at the Erasmus University. In future she hopes to combine her Medical Informatics background with management skills to manage information and knowledge in healthcare.