Research Issues in Clinical Data Warehousing

Torben Bach Pedersen
Center for Health Informatics
Kommunedata, P. O. Pedersens Vej 2
DK-8200 Århus N, Denmark
email: tbp@kmd.dk

Christian S. Jensen
Department of Computer Science
Aalborg University, Fredrik Bajers Vej 7E,
DK–9220 Aalborg Øst, Denmark
email: csj@cs.auc.dk *

Abstract

Medical informatics has been an important area for the application of computing and database technology for at least four decades. This area may benefit from the functionality offered by data warehousing. However, the special nature of clinical applications poses different and new requirements to data warehousing technologies, over those posed by conventional data warehouse applications. This article presents a number of exciting new research challenges posed by clinical applications, to be met by the database research community. These include the need for complex-data modeling features, advanced temporal support, advanced classification structures, continuously valued data, dimensionally reduced data, and the integration of very complex data. In addition, the support for clinical treatment protocols and medical research are interesting areas for research.

1. Introduction

Modern businesses use a multitude of different computer systems to manage their daily business processes such as sales, production, planning, etc. These systems, commonly referred to as *operational systems*, have been acquired from several vendors over a long period of time and are often based on different technologies. The integration between the operational systems is thus typically poor. However, integration is needed when the business must combine data from several operational systems in order to answer important business questions, e.g., sales and production data must be combined to determine the profitability of a product. The *data warehousing* approach solves the problem by integrating data from the operational systems into one common data store, known as the data warehouse, which is optimized for data analysis purposes [1, 5].

Data warehousing technology has traditionally been used in a business context, in order to answer questions about sales and other important events in the business of concern. The data models employed conceptually provide a multidimensional view of data, whether implemented in relational or dedicated multidimensional DBMS's, and this has proven very successful in the traditional application areas. However, some application areas have a need for more complex data structures. One such area is clinical data warehousing, where clinical data about a large patient population is analyzed to perform clinical quality management and medical research. Clinical data warehousing is a substantial application area in itself, and we focus on describing the requirements of this area. The issues described also apply to other application areas, in science or business, but such areas are beyond the scope of this paper. We will also concentrate on the use of clinical data for analysis purposes. Discussion of the operational use of clinical data, e.g., for cooperative purposes or remote diagnostization, is also not covered here.

The clinical domain requires more powerful data model constructs than conventional multidimensional approaches, and the data model should also provide advanced temporal support, e.g., for bitemporal data. More advanced classification structures are also needed, including means of managing dynamic, non-strict hierarchies, and of handling change. Continuously valued data, e.g., measurements, is very common and has special demands for aggregation and computation compared to conventional business data. The number of dimensions in clinical data is often very large, sparking a need for intelligent ways of dimensionally reducing the data into high-level abstractions.

There should also be a way of integrating very complex data, e.g., X-rays, in the data warehouse for analysis purposes, by more advanced means than just allowing the raw data to be retrieved. Clinical treatment protocols should be tightly integrated with the clinical data warehouse, to allow for follow-up on the corresponding quality of treatment, e.g., outcomes, for the individual protocols. Finally, medical research should be supported directly by the clinical data warehouse, e.g., by integrating data mining capabilities tailored to the specific domain.

The paper is outlined as follows. Section 2 describes the conventional use of data warehousing, as used primarily in business settings. To illustrate the various issues, a case study concerning a small clinical data warehouse is included. Section 3 describes the concept of an Electronic Patient Record (EPR), and lays out a roadmap for a new foundation for Clinical Data Warehouses (CDWs). The primary rationale is that a CDW should be very tightly integrated with the EPR, to support physicians and other clinical users throughout their daily work. We argue that it is attractive to base the CDW on the EPR and introduce EHCRA, the European Standard for EPR's. This standard has some nice features w.r.t. using EPR data for data warehousing. Section 3 also describes the research challenges that CDWs provide, and it compares CDWs with ordinary data warehouse applications. Section 4 summarizes the article and offers suggestions for next steps.

2. Background

This section provides a definition of a data warehouse, describes previous work, and presents a case study of a CDW.

2.1. A Brief Characterization of Data Warehousing

The term "Data Warehouse" (DW) was first used by Barry Devlin [2], but Bill Inmon has won the most acclaim for introducing the concept, defined as follows. "A Data Warehouse is a *subject oriented*, *integrated*, *non-volatile* and *time-variant* collection of data in support of *management's decisions*" [3]. Let us have a closer look at these interesting properties.

- *Subject Oriented:* In operational systems, data is organized to support specific business processes. Thus, the same data might be organized very differently in different operational systems. For example, it is likely that person data in a Human Resource application is organized differently from person data in a Point-of-Sale application. In a DW, data is organized by subject, or topic, e.g., Person, rather than by function.

- *Integrated:* A business typically employs many different operational systems, each optimized for a special business process, and each with its own data store. In the DW, data from all these systems is integrated, both by *definition*, i.e., the same data has the same type, and by *content*, i.e., the value sets of an attribute are the same, wherever they occur. Integration does not imply data warehousing—an appropriate organization is also required. If all operational data is in one operational system, e.g., the SAP system, it is still necessary to have a DW, where data is organized w.r.t. data analysis instead of data entry [1].

- *Non-volatile:* In the typical operational system, data is often kept only for a short period of time, e.g., 3 to 6 months, as it is only interesting for the daily business during this timespan. In a data analysis situation, however, the need to discover trends in the way business is doing and compare them with those of previous periods sparks a need to keep data for longer periods of time. Most DW's keep data for at least a couple of years, and many intend to keep it much longer.

- *Time-variant:* Operational data does not always have an explicit temporal dimension. It might not be interesting for an application, e.g., an inventory system, to know when a transaction actually took place. Also, operational systems often only store the current state of data. In the DW, time is a whole different matter. When analyzing data for trends, it is almost always important to know "the time of the data," so that all data in a DW can be related to a specific time point or interval. Also, not only the current value of data is stored, but often either snapshots of data at specific points in time, or a complete history of changes of the data.

- *Management's decisions:* Both words in this phrase are interesting in their own right. The word "decisions" indicates the very important fact that data in a DW is optimized for data analysis, not data entry. Thus normal database design principles do not necessarily apply, and managed redundancy of data is usually appropriate in a DW because it simplifies the database schema and improves analysis performance. The word "management's" indicates that DW data is traditionally used at the strategic level, by top management, for setting the course for the entire business. We would like to modify this to "management decisions," to capture the tendency that the DW is now also used at a "lower" level of the organization, by non-management employees, to get to know their part of the business

[1] In fact, the SAP company has done just this, and is now marketing a DW solution as an addition to their operational system.

better, thus providing better "micro" management in the daily work.

2.2. Previous Work

Like the database management area itself, the birth and rise of data warehousing has almost entirely taken place in the business world. Data warehousing was born out of the need of many businesses to view and analyze data from their many different operational systems together, to get a complete understanding of the business. Until recently, academia did not take interest in the area, and thus the field has been driven by the market, rather than by the research community. Research in distributed databases on issues such as global schemas and schema integration address some of the same challenges [4], but data warehousing still differs by employing data scrubbing, data cleaning, non-automatic data mappings, and bulkloading. Among other differences a DW stores more data than the sources and data is aggregated [1, 3, 8].

The focus of DW vendors as well as researchers has been on support for OLAP (OnLine Analytical Processing) functionality with good performance. In database research terms, the work has concentrated on the physical rather than the conceptual level. The data models employed have been of the multidimensional variety, where data is divided into *measurable business facts* and mostly textual *dimensions*, which characterize the facts and have hierarchies in them. In a retail business, *products* are sold to *customers* at certain *times* in certain *amounts* at certain *prices*. A typical fact would be a *purchase*, with the amount and price as the measures, and the customer purchasing the product, the product being purchased, and the time of purchase as the dimensions. A good visualization of the model, is to envisage data as living in an n-dimensional cube, with facts in the cells and the dimensions along the dimension axes [7].

OLAP systems have typically been implemented using two technologies: ROLAP (Relational OLAP) where data is stored in an RDBMS, and MOLAP (Multidimensional OLAP), where a dedicated multidimensional DMBS (MD-DMBS) is used. Reports indicate that traditional database design techniques, i.e., ER modeling [6] and normalized tables, are not well suited for DW applications; as a result, new techniques, e.g., *star schemas* [8], have emerged that better support the DW purpose of data analysis. As mentioned above, most work has concentrated on performance issues; and higher-level issues, such as conceptual modeling, have largely been ignored so far, at least in academia.

Recently, several researchers have pointed to this lack in DW research, and it has been suggested to try to combine the traditional DW virtues of performance with the more advanced data model concepts from the field of *scientific and statistical databases* [9]. This appears to be a very valuable direction, as users of a DW tend to work directly with the data, creating a need to put more semantics directly into the database schema, as opposed to storing the data semantics in application programs, as is the case in operational systems.

2.3. A Case Study

The case study illustrates the special demands of clinical data warehousing. The simplified case is taken from the domain of diabetes treatment [14, 15]. An ER diagram of the case using standard notation [11] is seen in Figure 1.

The most important entity type is the *patient*, as indicated by the placement in the middle of the diagram. A patient is identified by a Social Security Number (SSN) and has additional attributes Name, Birth Date, and Height, all of which we will consider to be static. A patient has many relationships to other entities, whose main purposes are to *characterize* the patient. Thus, these other entities might be viewed as dimensions of the particular patient.

First, a patient can be given one or more *diagnoses*. These are only valid in specified time intervals, as the patient's condition changes over time. The set of possible diagnoses is given by a classification of diseases, e.g., the World Health Organisation's ICD-10 standard [22]. A diagnosis has an alphanumeric code, a descriptive text, and an associated period of validity. A specific diagnosis might be superseded by another as medical knowledge evolves, thus ending its validity, but for historical reasons it is important to keep it in the classification. Diagnoses are grouped into diagnosis groups, e.g., "Diabetes diseases" or "Pregnancy-related diseases," for overview purposes. One diagnosis can be a part of multiple groups, e.g., "Diabetes during pregnancy[2]" can be a part of both of the just-mentioned groups. The participation in the diagnosis groups of diagnoses can change over time, as the demands for grouping vary. The groups also have an alphanumeric code and a descriptive text.

A patient is treated according to a *protocol*, which is a formal description of how a treatment should progress. Different protocols are used depending on the characteristics, e.g., the age, of the patient. We will not go into the very complex internal structure of a protocol, but will just record a code, a text, and a period of validity. The protocol used for treating a patient may vary over time.

An important indicator for the status of a diabetes patient is the condition of the *feet*, e.g., the blood circulation and presence of wounds. From time to time, the feet are photographed, and the pictures are stored along with the times they were taken.

[2]The reason for having a separate pregnancy related diagnosis is that the diabetes must be monitored and controlled much more intensely during a pregnancy to assure good health of both mother and child.

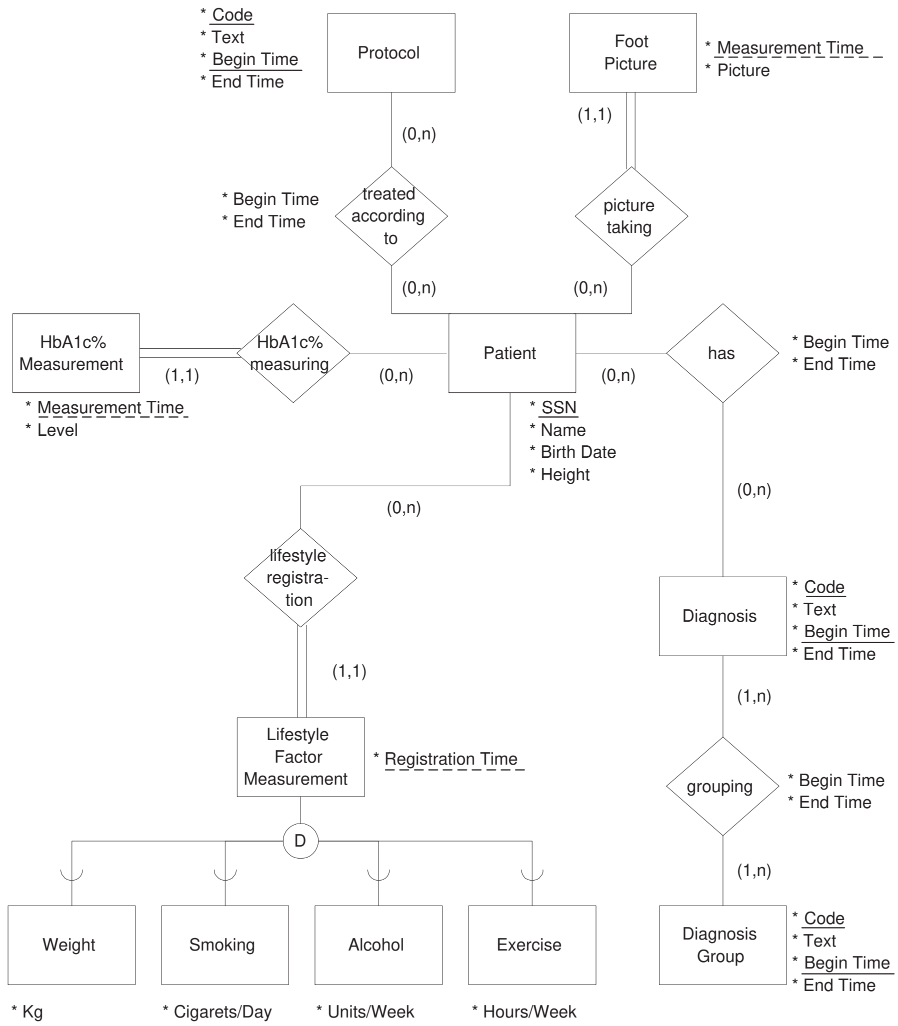

Figure 1. Case Study of a CDW for Diabetes

One of the most important measurements for diabetes patients is *HbA1c%*, which indicates the long-time blood sugar level and provides a good overall indicator of the patient's status during the recent months. This measurement is taken approximately every three months.

For diabetes patients, a healthy lifestyle is even more important than normally, as it can literally make the difference between life and premature death. To monitor the lifestyle, several *lifestyle factors* are measured. These include weight and smoking, alcohol, and exercise habits. These factors are not measured on a regular basis, but rather considered to be valid from the time of registration until a new registration is being made. As the lifestyle factors have a lot in common, they are modeled using subtypes.

3. Clinical Data Warehousing Requirements

In this section we characterize the special requirements that surface in a CDW. We will start by introducing the concept of the Electronic Patient Record (EPR), explaining how it can serve as a solid foundation for a CDW.

3.1. The Electronic Patient Record

It is important to define exactly what is meant by an EPR. The Medical Records Institute (MRI), an independent, customer-owned non-profit organization, provides a six-page definition of electronic patient records [16], organized into five levels of computerization of patient information. We will use the definition from Level 3 "The Electronic Medical Record," as this is the lowest level where all patient information originating from one healthcare enterprise, e.g., all patient data kept by a single hospital, resides in the EPR in a structured format, i.e., as separate data items rather than simply as scanned documents.

To paraphrase the standard, an Electronic Medical Record (EMR) shall be able to uniquely identify the person that the information concerns, e.g., through the use of an enterprise-wide patient index. It is the *complete* collection of information; thus data from other clinical systems should be integrated in the EMR and harmonized accordingly. The EMR shall be used directly by all healthcare staff to record information. It shall have legal validity as any other document. This places severe demands on the security system for access control, electronic signatures, auditing, and data integrity, i.e., data can only be corrected by amendments.

The EPR is thus the central component in the IT-infrastructure of a modern healthcare enterprise. It is the common tool used by all healthcare professionals working in the enterprise. It is the point of entry for most patient information, and it provides access to the data born in other systems, e.g., laboratory or financial systems. In spite of these characteristics, the EPR cannot be considered a data warehouse in itself. Data in the EPR is used and organized according to operational purposes, where many kinds of data about one patient is presented to get an overview of the health status of the patient. Thus, data in the EPR is used and organized in a *by-patient* fashion.

In a DW, specific aspects of properties for a large population of patients are analyzed for trends, thus data is used and organized in a *by-property* fashion. The EPR is more akin to what Inmon defines as an *operational data store* (ODS) [18]: the integrated data store used as the basis for building the DW. The most important obstacle in using the EPR as a basis for a CDW is the multitude of different EPR systems on the market; the task of integrating data from several EPR systems is a hard one. This creates the need for a common standard for EPR data.

3.2. The EHCRA Standard for EPRs

The EHCRA standard [17] is the result of a European EPR standardization effort. It describes how to structure an EPR and lists demands that an EPR should meet. The ideas in the EHCRA standard originated from the Norwegian NORA project [19], which has led to the development of the DocuLive EPR system by Siemens Nixdorf Norway. DocuLive EPR is a tool for implementing EPR's based on the EHCRA standard.

EHCRA defines the EPR by means of a *document metaphor*: The EPR for a patient should be thought of as consisting of a number of documents containing information about the patient. The documents are structured, i.e., are not in free-form text. The structure of a document is *hierarchical*, with a document made up of *record items* or *record item complexes*, see Figure 2.

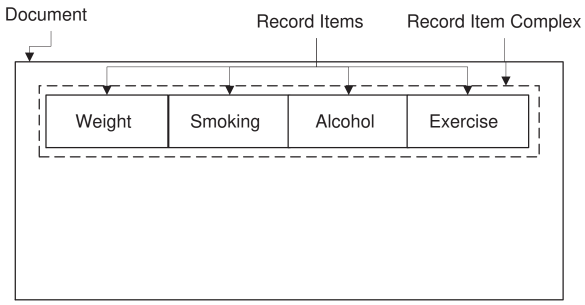

Figure 2. An EHCRA Document

The lowest level of a document is the record item, which can be thought of as a basic element of information, e.g., the patient's name. Record item complexes can be made up of record items or record item complexes, leading to a tree-structured document. A typical record item complex would consist of lifestyle information such as weight, smoking, alcohol, and exercise, grouped together. A record item can also point to the record item or record item complex in another document, where this piece of information was originally entered, thus providing multiple views of the data. This gives the EPR the structure of a *directed, acyclic graph*. All record items have some common properties such as a reference to the patient and a reference to the original context of the data (in case of a pointer). The original context is important when information is exchanged, as new users of the information can then get a complete view of the patient's situation when, e.g., a diagnosis, was chosen. The legal requirements are met by recording the time the data is included in the EPR, the status of the data, i.e., valid or invalid, and the unique id of the healthcare person entering the data. An update of the patient's weight is thus made by marking the old weight as invalid, and inserting a new valid weight, thereby keeping the full history of changes.

3.3. New Challenges

As mentioned above, clinical data warehousing introduces several new challenges to DW technology, compared to conventional data warehousing. We will illustrate these by referring to the example from the previous section.

3.3.1 Utilizing EPR Features

One very important issue is to utilize the features of the EPR optimally for building a clinical data warehouse. EPR systems in general, and EHCRA-based systems, in particular, have features that make them a very good foundation for clinical data warehouses.

The EHCRA standard is the European standard for the structure of the EPR. All EPR systems must thus be able to at least deliver data in an EHCRA-compliant format, even if they do not structure data internally according to EHCRA. Thus, utilizing the features of EHCRA in the integration of EPR data with the CDW allows for a very attractive and open solution that will work with many different EPR systems.

All versions of data in the EPR are stored along with their times of update. This gives full *transaction-time* support in EHCRA-based systems, thereby making it much easier to provide this support in the clinical DW. Often *valid time* is also attributed to the data in the EPR, providing for full *bitemporal* support [20].

The EPR is supposed to be the only (or at least the primary) tool that the clinical user is using in the daily work, so there is a great need to have access to all data, also lab results, etc., through the EPR. Thus *integration of operational data is already achieved* in the EPR, making the integration process in the DW very easy in comparison to conventional data warehousing. At a higher level in the MRI standard mentioned in Section 3.1, data from several healthcare enterprises, potentially the whole world, is integrated in the EPR. This makes the integration in the DW of data from different locations much easier.

The EU-sponsored project Synapses [21] concerns the building of a *federated healthcare record server* that integrates a wide variety of EPR systems. This should provide access to clinical data in an EHCRA-compliant format, no matter how the actual EPR systems structure the data internally. Initially, the goal of the Synapses project is to facilitate the exchange of electronic patient records between different healthcare units, possibly using different EPR systems. However, another exciting application would be to use the Synapses server to transfer data to the CDW in a uniform way, no matter what the underlying EPR systems are, thus making the task of integrating data from different EPR systems very easy. A lot of effort could be saved, compared to accessing the proprietary EPR systems directly.

In normal operational systems, data is entered "post mortem" either automatically or by a clerk. The data is almost never used again by the person registering or entering the data, thus giving little incentive for carefully registering all the data, the correct data, and nothing but the correct data. This means that extensive data cleaning procedures must be established when the data is to be transferred to the DW [3]. In the EPR, the physician entering the data is also the primary user of the data, so entering dirty data will directly translate into problems in the daily treatment of the patients. Thus, data is quality assured continuously by the primary users. This means that the operational data is more likely to be of *high quality*, thus requiring less cleaning when being moved to the CDW. This will give the results obtained from the CDW a high level of credibility.

There are many unresolved issues in how to optimally exploit EHCRA-compliant systems as the basis for CDW's.

3.3.2 Complex Data Modeling Features

One of the most prominent demands is a data model for the CDW that includes more complex modeling constructs than typical multidimensional models, while not losing their obvious strengths in the area of decision support, i.e., we should not return to the full generality of the ER model.

In the multidimensional model [7], facts are in a n-1 relationship to the base elements of the dimensions, which in turn encode strict hierarchies, i.e., lower levels have n-1 relationships to upper levels. An example of this is a purchase. Exactly one product can be purchased, the product can belong to exactly one product category, etc.

But consider the case study where a patient has multiple diagnoses at the same time. The relation between patient and diagnosis is most naturally modeled as an n-n relationship, as the *same* patient may have multiple diagnoses. For instance, if we ask the question "How many patients have diagnoses A or B" we only want patients with both to be counted once. We should be able to capture this intended behavior in the schema. This is not easily possible using a conventional multidimensional model.

In multidimensional modeling [8], we have three alternatives for encoding n-n relationships: traditional dimensions, mini-dimensions, and snowflaking. Using traditional dimensions, we would enumerate all the possible combinations of diagnoses. Having 10.000 diagnoses, this would amount to $2^{10.000}$ dimension records, making this solution practically unusable. Enumerating only the combinations actually used would still yield a very large number of dimension records. Furthermore, the dimension tables would be very wide and incomprehensible to the users. Using mini-dimensions with one dimension for each possible diagnosis would yield 10.000 dimensions, making the solution bad-performing as well as incomprehensible. Using snowflaking will not give any advantages over traditional dimensions, as the basic elements of the dimensions would the same, i.e., the possible combinations of diagnoses.

Another characteristic of clinical data is that we have many "loosely coupled" facts, e.g., the weight and smoking measurements from the case study. The values of these two measurements can change independently of each other, and a value is not always present at a given point in time, i.e., if the patient has not reported smoking habits. The measurements can be viewed in two ways, as time-variant attributes of the patient, or as separate entities that can be manipulated independently. The data model should be able to handle both treating the facts together, as if they belonged to the same entity, e.g., patient, or treating them separately.

In addition the data model should provide integrated semantic support for the demands listed in the following sections, e.g., temporal support, so that the solutions do not appear as poorly integrated "add-ons."

3.3.3 Advanced Temporal Support

One very important property of clinical data is the importance of temporal aspects. The same test, e.g., the HbA1c% measurement, can be made hundreds of times, so it is important to know both when the data is considered to be *valid in the real world*, and when it is *stored and changed in the database*. These temporal aspects of the data, known as *valid time* and *transaction time*, must both be supported to provide *bitemporal* support [12]. This support is for instance needed in order to "couple" different facts, e.g., smoking and weight, thereby computing "snapshots" of measurement values at specific intervals or points in time[3]. These snapshots are used to observe temporal trends in the evolution of one type of data values or in the relation between different types. In order to conduct these and other types of time studies, it is necessary to have available a strong support for time-series data, including a rich set of temporal analysis tools. It should be possible to use the above-mentioned advanced temporal concepts wherever meaningful. These advanced temporal concepts are not supported by current models.

3.3.4 Advanced Classification Structures

A data type of extreme importance in the clinical sector is "coded," or classified, data. One example is a diagnosis, which at the lowest level is a very precise indication of one specific medical condition. Diagnoses are then grouped repeatedly into larger, more general classes. A diagnosis is a good example of a typical OLAP "dimension," as it characterizes the condition of the patient, it is attached to; but unlike the typical dimension, the diagnosis hierarchy is non-strict. Take for instance the diagnosis "Diabetes during pregnancy." This is in the group "Other pregnancy related diseases," but also in the group "Diabetes." This leads to an interesting requirement. If we ask for the number of patients, grouped by diagnosis at the lowest level, we naturally only want each pregnant-diabetes patient to be counted once. Then, if we "roll up" to the next level of diagnoses, we want the patients to be counted both in the pregnancy-related and the diabetes diseases groups. If we then roll up again, not considering the diagnosis dimension at all, we should again only count the patients once. Clearly the user of the DW should be able to work with the data and get the correct results, without having to worry about double-counting, etc. In the case of strict hierarchies, this feature is referred to as *summarizability* [10].

[3] This is possible because although the same property may be measured many times for the same patient, only one measurement value is considered to be "valid" at any given point in time.

Current data models do not specifically address this issue of correct aggregation in the case of non-strict hierarchies.

Another requirement related to classification structures is that they should be able to handle temporal change. Classifications change and new diagnoses and new groups come and go at a steady rate. The CDW should support this in an intelligent way, so that analysis of data across changes is handled smoothly and preferably transparently to the user. This requirement is also not handled well by current techniques.

3.3.5 Continuously Valued Data

Measurements and lab results, e.g., the HbA1c% measurement from the case study, are the key facts in the CDW. Unlike typical DW facts, these types of data clearly do not yield any meaning when summed. Other standard aggregation operators, such as MIN, MAX and AVG do apply, but the real demands are for more complex operations, such as standard deviation and other statistical functions. These operators are mainly used during follow-up on treatment in relation to clinical protocols, see below, or in medical research. The CDW should be able to support these advanced operations very efficiently, to supply the performance necessary to analyze large amounts of data accumulated over long periods of time. To do so, it must be investigated how pre-stored and pre-aggregated data can be used to achieve high performance. Current techniques for maintaining pre-aggregated data support only simple aggregation operators such as SUM.

3.3.6 Dimensionally Reduced Data

In a clinical DW, average patients might have hundreds of different facts describing their current situation, in diabetes treatment about 200 facts are recorded. There is an urgent need to be able to aggregate this massive amount of information in a useful way. In the case study, a patient has four independent indicators describing the lifestyle, i.e., levels of smoking, exercise, alcohol, and weight. These could be combined into one aggregate measure indicating the overall lifestyle of the patient. In multidimensional terms, we have reduced the previous four dimensions to just one. In a traditional OLAP world, the only way to reduce dimensionality is by projection, thereby *ignoring* all information about the omitted dimensions. The dimension reduction approach clearly has advantages over this, as the complexity of the data is reduced, while the essence is maintained. The clinical DW should be able to support the definition of such combination functions, and it should provide good performance for reducing/increasing the number of dimensions. The issue of pre-aggregation in connection with dimensionality reduction is also very interesting.

3.3.7 Integration of Very Complex Data

The clinical world is also characterized by very complex types of data. One example is the 2048 by 2048 pixel foot picture from the case study, which in multidimensional terms could be viewed as being 4.194.304 dimensional, by considering all pixels to be independent dimensions. While this clearly is an overly complex way of looking at it, it should be possible to incorporate this type of data in the CDW for data analysis purposes. The functionality should be more advanced than just allowing the raw data to be stored and retrieved, i.e., the support often associated with "blobs." Rather, it should be possible to define *feature extractors* on the raw data, e.g., pattern recognition functions for wounds, and to perform analyses on the extracted features. The extractors should be tightly integrated with the DW, allowing for addition of new and modification of existing extractors incrementally, i.e., without having to recompute every feature from scratch. Existing data warehouse techniques accommodate only with simple, structured data such as text and numbers.

3.3.8 Support for Clinical Protocols

The introduction of *managed care* is a very prominent current trend in the clinical world. Instead of relying solely on the judgment and knowledge of one doctor, the treatment of specific diseases is conducted according to well-defined protocols that specify the conditions and actions for using specific treatments. The protocol can be viewed as a "best practice" or an advanced set of business rules. A patient can be treated according to different protocols at different times, as shown in the case study.

There is a need to analyze the actual treatments, to investigate conformance to the protocols, outcomes, etc. As an example, we could ask what protocol provides the best treatment in terms of keeping the HbA1c% close to normal.

Ideally, the protocols would be specified formally, to allow for "automatic" follow-up on treatments. That is, queries against the CDW could be generated directly from the protocols, and the results of these be used to test conformance, to adjust the protocols, etc. The CDW should have integrated support for clinical protocols, to accommodate this important part of clinical practice. Today, data warehouse systems do not have any support for advanced business rules like these.

3.3.9 Support for Medical Research

Medical research can take several different forms; one form that the clinical DW enables is the so-called *qualitative* research where large amounts of data is analyzed to confirm known or discover unknown trends and correlations in the data. For example, a correlation between the weight and the HbA1c% of a patient could be discovered. The discovery process in medical research would benefit enormously from having data mining facilities integrated into the CDW. There should be a conceptually simple, fast-performing, and yet flexible way to produce the "flat" sets of data that are normally fed into data mining algorithms. The results could be used as inspiration for hypotheses that could then be tested in controlled, formal clinical studies. The integration of data mining and data warehousing and the use of a DW for research purposes are both in their infancy in today's DW products.

3.4. Comparison of Conventional and Clinical DW

The differences between clinical and conventional data warehouses extend into their corresponding operational systems. For conventional systems, the operational systems often consist of a wide range of poorly integrated legacy systems. In a modern clinical setup, however, almost all data is already accessible in an EPR, thus providing integration at the operational level. The EPR also has other characteristics that differ from most typical operational systems (see Table 1). Both types of systems have small granularity of data, but in the EPR, data is never deleted, and a full trace of all updates are maintained for legal reasons.

	Conventional	*Clinical (EPR)*
Integration	No	Yes
Granularity	Small	Small
Volatility	High	Zero
History	No	Yes

Table 1. Comparing Conventional and Clinical Operational Systems

If we consider the same characteristics for conventional versus clinical data warehouses, we also see some interesting trends (see Table 2).

Integration of data is achieved for both types, but in the typical conventional DW, integration is difficult to achieve because data is scattered in many legacy systems. In contrast, integration of data is already achieved in the EPR, thus making it easier to build the DW. Granularity varies from small to large in a conventional DW, but in a CDW, we always need to have the operational granularity of data. This

	Conventional	*Clinical*
Data Model	Simple	Complex
Temporal Support	Medium	Advanced
Classifications	Simple	Advanced
Continuously Valued Data	No	Yes
Dimensionally Reduced Data	No	Yes
Very Complex Data	No	Yes
Advanced Business Rules	Maybe	Yes (Protocols)
Data Mining	Maybe	Yes (Medical Research)

Table 3. Characteristics of Conventional versus Clinical Data Warehouses

	Conventional	*Clinical*
Integration	Yes (hard)	Yes (easy)
Granularity	Medium	Small (drill back)
Volatility	Low	Zero
History	Sometimes	Always

Table 2. Comparing Conventional and Clinical Data Warehouses

is caused by the need to "drill back" to the EPR, e.g., when encountering an interesting anomaly in the data. The physician then needs access to the full patient record to determine the exact cause. In a conventional DW only 6-10 years of data is kept, but in the clinical world it is important to see the full disease history, which might span 50 years or more, e.g., in the case of diabetes patients. It is also very important to have the full update history of the EPR, to facilitate trend analysis. This level of temporal support is not always present in business data warehouses.

3.5. Standardization Efforts

An area of high importance to clinical information systems is the various standardization efforts in the field of healthcare informatics.

First, the Health Level 7 (HL7) organization's work on standardization of electronic data interchange in healthcare environments [25] specifies the content of electronic messages transmitting healthcare information, by referring to a common model that specifies domain concepts and legal data values. The HL7 standard is widely used in the industry for interfacing different systems.

Second, the Object Management Group (OMG) has launched the CORBAmed initiative [23] that is aimed at providing standard interfaces to healthcare information systems based on OMG's Common Request Broker Architecture (CORBA). Several CORBAmed workgroups focus on specific areas such as clinical decision support, clinical observations, patient identification, and HL7 integration.

Third, the most recent player is Microsoft, with the "ActiveX for Healthcare" initiative [24] that is a direct competitor to CORBAmed, but is based on Microsoft's Distributed Component Object Model (DCOM) instead of CORBA.

From a clinical data warehousing perspective, these initiatives deal almost entirely with integration of clinical data in the CDW and can thus be seen as enablers of integration, whether the CDW is based on an EPR or not. These initiatives do not address the other challenges presented in the paper.

4. Summary

Table 3 summarizes the challenging needs covered in the paper and compares them with the characteristics of a conventional DW, i.e., a DW as it is often used in a business context. This does not imply that business or other domains do not need the advanced features presented in the paper. Rather, the comparison shows why conventional DW techniques fail to meet the requirements of clinical data warehousing. The investigation of these requirements are the focus of this paper. From the comparison it is clear that the needs of a clinical DW poses some very interesting challenges for researchers and developers alike.

The data model must support advanced constructs such as many-to-many relationships between facts and dimensions. Full support for bitemporal data and analysis over time is also needed. Advanced classification structures must be provided that integrate support for non-strict hierarchies with means of handling change and time, while maintaining support for correct aggregations.

Continuously valued data must be efficiently supported, including how to perform advanced operations on them. Dimensional reduction of data in the CDW is important in order to make sense of the high-dimensional data. Very complex data such as pictures or x-rays must be available in the CDW, with facilities for doing analysis on them.

The integration of clinical protocols in the CDW is important to allow for follow-up on the treatment of patients.

Support for medical research, e.g., via data mining facilities, will enable the clinical community to perform their research much more efficiently than is possible today.

We have shown that DW technology faces exciting new challenges from the area of clinical data warehousing. Clinical data warehousing provides excellent opportunities for first-class DW research that will also have applications in areas beyond the clinical world.

The challenges that are especially important to the general database research community include the following: advanced data models including temporal support, advanced classification structures, continuously valued data support, and dimensional reduction of data.

We will work on these issues in clinical applications in order to support the successful application of data warehousing in the clinical world.

Acknowledgements

This research was supported in part by the Danish Technical Research Council through grant 9700780, by the Danish Academy of Technical Sciences, contract no. EF 661, and by the CHOROCHRONOS project, funded by the European Commission, contract no. FMRX-CT96-0056.

References

[1] J. Widom. Research Problems in Data Warehousing. In *Proceedings of CIKM 1995*, pp. 25–30.

[2] B. A. Devlin and P. T. Murphy. An Architecture for a Business and Information System. *IBM Systems Journal*, 27(1), 1988.

[3] W. H. Inmon. *Building the Data Warehouse*, 2nd edition. Wiley Computer Publishing, 1996.

[4] C. Batini, M. Lenzerini, S. B. Navathe. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys* 18(4):323–364, 1986.

[5] J. Widom (editor). Special Issue on Materialized Views and Data Warehousing. *IEEE Data Engineering Bulletin*, 18(2), 1995.

[6] P. P-S. Chen. The Entity-Relationship Model — Toward a Unified View of Data. *ACM Transaction on Database Systems*, 1(1):9–36, March 1976.

[7] J. Gray et al. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab and Sub-Totals. *Data Mining and Knowledge Discovery* 1(1):29–53, 1997.

[8] R. Kimball. *The Data Warehouse Toolkit*. Wiley Computer Publishing, 1996.

[9] A. Shoshani. OLAP and Statistical Databases: Similarities and Differences. In *Proceedings of ACM PODS 1997*, pp. 185–196.

[10] M. Rafanelli and A. Shoshani. STORM: A Statistical Object Representation Model. In *Proceedings of SSDB 1990*, pp. 14–29.

[11] R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*, 2nd edition. Benjamin/Cummings Publishing Company, 1994.

[12] C. S. Jensen and R. T. Snodgrass. Semantics of Time-Varying Information. *Information Systems*, 21(4):311–352, March 1996.

[13] E. B. Baatz. Return on Investment - What's It Worth? *CIO Magazine*, October 1, 1996.

[14] Kommunedata. *GS-Open Design Specification* (in Danish). Internal document, 1996.

[15] L. H. Nielsen. Clinical Aspects of the Concept Electronic Record in the Danish Health Service (in Danish). Project report, Health Informatics Education, Aalborg University, Denmark, 1996.

[16] Medical Records Institute WWW page. `<http://www.medrecinst.com/oldsite/MRImain/levels.html>`. Current as of April 20th, 1998.

[17] European Committee for Standardization (CEN). *European Prestandard - prENV 12265. Medical Informatics - Electronic Healthcare Record Architecture.*, 1995.

[18] W. H. Inmon. *Building the Operational Data Store.* John Wiley & Sons, 1995.

[19] K. Skifjeld. From concept to product - experiences from the development of the DocuLive EPR system. *The British Journal of Healthcare Computing & Information Management*, March 1997.

[20] C. S. Jensen et al. A Consensus Glossary of Temporal Database Concepts. *ACM SIGMOD Record*, 23(1):52–65, March 1994.

[21] Commission of the European Communities, Directorate General XII. *Synapses: Federated Healthcare Record Server*. Contract no. HC 1046 (HC).

[22] World Health Organization (WHO). *International Classification of Diseases (ICD-10)*. Tenth Revision, 1992.

[23] CORBAmed WWW page. <http://www.omg.org/corbamed>. Current as of April 13th, 1998.

[24] ActiveX for Healthcare WWW page. <http://www.mshug.org/activex/>. Current as of April 8th,1998.

[25] Health Level 7 (HL7) WWW page. <http://www.mcis.duke.edu/standards/HL7/hl7.htm>. Current as of April 8th, 1998.