Clinical Data Warehousing - A Survey

Torben Bach Pedersen† Christian S. Jensen‡

† Kommunedata, Center for Health Informatics, P.O. Pedersensvej 2,
DK-8200 Århus N, Denmark, email: tbp@kmd.dk

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

Abstract

In this article we present the concept of *data warehousing*, and its use in the clinical area. Clinical data warehousing will become very important in the near future, as healthcare enterprises need to gain more information from their clinical, administrative, and financial data, in order to improve quality and reduce costs. Adoption of data warehousing in health care has been slowed by lack of understanding of the benefits offered by the technology. This paper contributes by providing needed understanding, by introducing the opportunities offered by data warehousing, describing current efforts in the area, and providing criteria for comparing clinical data warehouse systems.

1 Introduction

The concept of data warehousing has taken the computer industry by storm in the recent years, as enterprises have realized the enormous opportunities in extracting useful information from the data "hidden" in their computer systems. The new functionality offered by data warehousing has traditionally been used in business, in areas such as retail and finance, but the technology is now increasingly being used in more "scientific" areas.

One of these areas is clinical, where clinical data about a large patient population is analyzed to perform clinical quality management and medical research [3]. The use of data warehousing in the clinical area will be driven by the *need* to manage the entire care process to stay competitive, as well as the *opportunities* for gaining new insights by actively *using* the enormous amount of patient data available. Data warehousing stands out as the only *viable* technology for realizing the full information potential in operational data. Thus, clinical data warehousing will become very important in clinical enterprises in the not-too-distant future. So far, adoption of the technology in healthcare has been hindered by lack of knowledge on what data warehousing can be used for, and what current products offer. This paper offers the needed knowledge by describing the current state-of-the-art of products and other efforts, providing evaluation criteria for comparing clinical data warehouse systems, and looking at future directions of the area.

This work was done as part of a clinical data warehouse research project, involving university computer scientists, industry, and clinicians engaged in a joint effort to determine how the data warehouse concept should be used and extended to support the needs of clinical users.

In Section 2, we define the concept of data warehousing and provide evaluation criteria for clinical data warehouse systems. In Section 3, we describe the various efforts in the area, and list their conformance to the criteria. Section 4 discusses the relative merits of the efforts, and offers a look at the future, as well as concluding remarks.

2 Data Warehousing

The term "Data Warehouse" (DW) was first used by Barry Devlin [1], but it is Bill Inmon that has won the most acclaim for introducing the concept, see Figure 1 for the basic architecture of a DW.

He defines a Data Warehouse as: "A Data Warehouse is a *subject oriented*, *integrated*, *non-volatile* and *time-variant* collection of data in support of *management's decisions*." [2] Let us have a closer look at these defining properties.

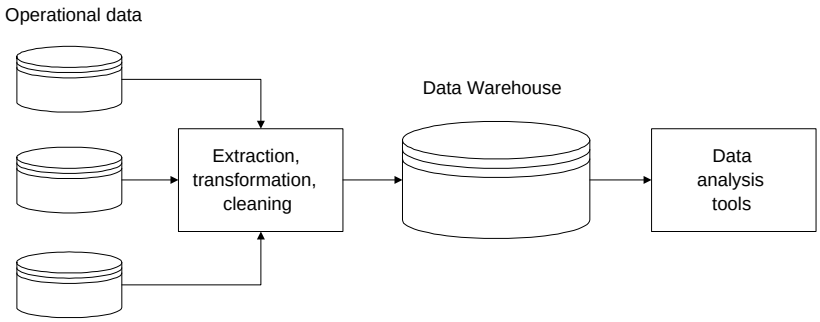

Figure 1: Integration of Operational Data in the Data Warehouse

Data is organized by *subjects*, such as patients, to allow the users to work with terms from their daily life. In operational systems, data is organized to support a particular application, e.g., a laboratory system, often making data incomprehensible to humans. The subject orientation of a DW makes it much easier to understand the data.

Data is *integrated* from multiple operational systems, both by definition and content. As seen in Figure 1, data is extracted from the operational systems, transformed into a format suited for data analysis, and "cleaned" to adjust for invalid, incompatible, or missing data. This makes the previously very hard task of combining data from different systems a lot easier for the end user.

Data is *non-volatile*, i.e., it is kept for many years and sometimes never deleted. In operational systems, data is often deleted after a few months, when they are no longer needed by the particular application. Retaining the data makes analysis over long time periods possible.

Data is *time-variant*, i.e., it always has a notion of time attached to it, and often a complete history of changes is kept. This makes analysis for trends over time possible. In operational systems, data is often not related to time, and only the newest version of the data is stored.

Data in a DW support *management's decisions*, i.e., it is optimized for data analysis, not data entry. Data is used to *understand and manage* the enterprise, both at a strategic and a tactical level. Examples of this include clinical quality management and medical research.

Compared to an ordinary clinical decision support systems (CDSS), the DW has a much broader scope. Typically CDSS's are specialized for one very specific purpose, with an emphasis on a deep level of functionality. There is a very limited possibility of asking "new" questions. On the other hand, the DW has an emphasis on the data itself, providing a possibility to combine all the kinds of data in the enterprise, and thus more easily answer unanticipated questions. Instead of just providing a set of fixed reports, an explorative way of working with the data is encouraged and supported.

To allow for comparison of the systems to be covered in Section 3, we list seven different criteria, that a full-blown clinical DW should meet in order to provide full business value for the user.

The system should be *open*, i.e., it should allow integration with systems or components from other vendors. This is very important, as systems without this ability will lock users into a proprietary, non-extensible solution, which almost surely will not support all their needs.

The system should have features for importing data from *external systems*, e.g., laboratory systems, into the DW. Without this feature, the system cannot be considered a "true" DW, as users will have to reenter data into the DW system to gain advantage of it. This will mean both extra work and more data errors, which will severely limit the usefulness of the system.

A full-blown clinical DW should support *all the types of data* important to the healthcare enterprise, including financial data (F) such as billing and contracts, demographic data (D) such as age and sex, clinical data (C) such as diagnoses and procedures, numerical data (N) such as lab results, and image data (I) such as x-rays. If all kinds of data are not supported in the DW system, the possibility of combining data to gain new knowledge will be severely limited, thus diminishing the business value of the DW.

Next, a clinical DW system should support data analysis at several levels. The lowest level is the *patient* level, where data about the individual patient can be viewed and analyzed, e.g., to find a pattern in the development of a disease for a particular patient. This level of analysis focuses on giving the particular patient the best possible treatment, and is thus important for the *practice* of care. The next level is the *group* level, where data about a group of patients, e.g., patients having a particular disease, is analyzed. One application of this is clinical quality management, where treatments and outcomes are

analyzed and compared to norms in order to identify how the care process can be enhanced. This level focuses on *medical research* and *care improvement*, and is thus important from a more *scientific* point of view. The top level is that of the healthcare *enterprise*, where clinical, financial, and demographic data are combined to investigate the profitability and overall quality of the services provided. This level of analysis focuses on overall performance of the enterprise and is thus important from a *management* perspective.

We also list if the systems have any particular *advanced features* that makes them stand out from the rest such as support for drug development or predefined disease studies. These features may be very important for some users, e.g., the pharmaceutical industry, while others have different demands.

3 Clinical Data Warehousing Systems

We will now examine the current state-of-the-art of clinical data warehousing by describing the most important efforts we have encountered so far. The first five sections describe commercial data warehouse products, while last three describe specific clinical data warehouse projects. The list is not meant to be exhaustive, but we believe that it is representative for the current state of affairs. For all DW application areas, most of the work has been done in industry, rather than in scientific environments. Compared to using DW for business purposes, clinical data warehousing is still in its infancy, with only a few providers and no wide-spread use. It is, however, recognized as an important and emerging field presenting tough challenges [3].

	Open	Ext. Data	Datatypes	Patient	Group	Enterprise	Advanced Features
CC	No	Yes	C,N	Yes	Yes	No	Collaborative info
OC	Yes	No	C,N	Yes	Yes	No	Drug development
SAS	Yes	Yes	C,N	Yes	Yes	No	Drug development
Ma	Yes	Yes	F,C,N	Yes	Yes	No	Disease studies
IAI	Yes	Yes	F,D,C,N	Yes	Yes	Yes	Longitudinal studies
SMS	Yes	Yes	F,D,C,N	Yes	Yes	Yes	Rules Engine
QI	No	No	C,N	No	Yes	No	Very large database
TUCH	No	No	C	No	Yes	No	None
SMI	No	Yes	C,N	No	No	No	Temporal, protocols

CC: Clinical Computing; OC: Oracle Clinical; SAS: SAS Institute; Ma: MEDai; IAI: Information Architects Inc.; SMS: Shared Medical Systems; QI: Quest Informatics; TUCH: Turku University Central Hospital; SMI: Stanford Medical Informatics

Table 1: Comparison of Clinical DW Systems

3.1 Oracle and Partners

As one of the major DW players, with an extensive consulting business and a great number of partners, the Oracle corporation has been involved in a lot of DW projects, including some of the clinical variety.

Clinical Computing [4], an Oracle partner, provides the di-Proton/Clinical Data Warehouse solution that is aimed at renal, i.e., kidney function, information management. The management of information such as treatment assessments, dialysis equipment checks, interdialytic vital signs, complications, procedures, medications, and infection history is supported. This enables analysis of core clinical indicators and outcomes. Interfaces to laboratory systems and dialysis machines facilitates the data collection process. An interesting feature is that collaborative care information such as nutritional recommendations, psychosocial assessments, etc., is also recorded, providing a larger picture of the patient's status. Table 1 summarizes the evaluation of each DW effort covered in this section according to the criteria given in Section 2.

Recently, Oracle Inc. itself ventured into the clinical world with the release of Oracle Clinical [5], the first in the Oracle Pharmaceutical suite of applications. The product is meant to address the needs of large, pharmaceutical companies for managing information about the extensive clinical trials that are needed to test a new drug. This involves meeting strict government regulations as well as facing hard competition. Previous, the companies had to build their own applications for each series of trials,

resulting in very high costs and longer development cycles. The product is already in use at several major companies, including Boehringer IngelHeim, Genentech, and Hoffmann-La Roche. The core product will later be supplemented by products supporting adverse event handling, remote data entry, and data analysis, the objective being to provide a complete turn-key solution for the industry.

3.2 SAS Institute

Another major player in the DW market, SAS Institute, a long-time champion in the data analysis marketplace, has recently introduced the SAS Pharmatechnology Process [6], a clinical data warehouse framework for the specific needs of clinical research. The cornerstone in the framework is SAS/PH-Clinical, a software system for assimilating and reviewing data from clinical trials.

The product already supports the review process, and is being further developed to support all the processes involved in developing a new drug, such as laboratory analysis and pharmacokinetics. The product allows the clinical trials data to be viewed in spreadsheet or graphical form. The patient population may be subsetted based on specific attributes, test results or the treatment protocol used, and the subset may be compared to other subsets or the complete patient population. The data is usually displayed in summary form, but the user may drill down to watch the complete patient history, e.g., when an anomaly occurs. Both ad-hoc data exploration and standardized reporting is supported.

The system is not a complete clinical data warehouse in itself, but it may be used with SAS Institute's Data Warehouse Administrator to build such a solution. Indeed, a clinical data warehouse is a very central part in SAS Institute's visions for the use of information in the health care industry [7]. These include data warehousing, continuous quality improvement, outcomes management, health plan management, and utilization analysis. Especially interesting on the clinical side is the focus on using data mining and OLAP techniques to identify key clinical indicators, thus improving the quality of care.

SAS Institute has several partners in this area, including Xerox for document integration, and Oracle for integration with the Oracle Clinical software [8]. It is possible to map views defined in Oracle Clinical into the study definitions that are required by SAS/PH-Clinical in a reasonably straightforward way, transferring both data and metadata from Oracle Clinical. This allows the rich data analysis features offered by the SAS System to be used for clinical data analysis.

3.3 MEDai

MEDai is a company focused on utilizing Artificial Intelligence (AI) techniques for health care data analysis. They provide the Clinical Decision Support System (CDSS) [9], which is a data warehouse/clinical data repository with powerful analytical capabilities. The CDSS system can extract data from existing hospital systems to provide both clinical and financial data. It allows for comparison of performance to norms, both at the facility and physician level. It also provides outcome analysis and has facilities for the development of treatment protocols. AI techniques are used for severity/risk adjustment for the patients, and data mining and drill down capabilities allow for data exploration. A distinguishing feature is the more than 20 predefined "disease studies" for data analysis, covering more than 80 percent of normal admissions. These include pneumonia, cesarean sections, chest pain/coronary artery disease, HIV, asthma, diabetes mellitus, and migraine. This makes it possible to immediately perform data analysis of the most common diseases.

3.4 Information Architects Inc.

In a data warehouse, different types of data are integrated to get a complete view of an enterprise. This is what Information Architects Inc. (IAI) makes possible with it's "healthcare information warehouse" product [12]. In this system, administrative, financial, and clinical data are integrated to provide a foundation for measuring both cost and value of the services delivered in the healthcare delivery process. The data warehouse consists of more than 200 tables, with integrated desktop reporting and server-side tools. It supports quality-of-care reporting, provider comparisons, outcome analysis using advanced statistics, and longitudinal health studies. Categories of care and treatment groups are supported for summary information, with associated norms for comparison of performance. The system is highly scalable, ranging from PC servers to Massively Parallel Processor (MPP) machines.

3.5 Shared Medical Systems

Shared Medical Systems (SMS) have a long history in the healthcare informatics business. Their product line Novius.ihn [13] is aimed at the Integrated Health Network (IHN) market, and has several interesting features. It has an integrated DW that standardizes, stores, and manages demographic, financial, and clinical data from across the IHN. A common vocabulary engine allows for the definition of terms and relationships, which can then be used for the definition of clinical protocols. The rules engine allows for transformation and abstraction of data. Relational data structures supporting analysis over time and aggregation are provided. The product has an integrated management solutions component for strategic and tactical analysis. A quality management component providing study definition, indicator derivation, and statistical analysis is available as an option. The DW is well integrated with the wide range of clinical operational system that SMS offers and has support for receiving data in many formats, including EDI. This provides for easy acquisition of quality DW data.

3.6 Quest Informatics

The clinical data warehouse run by Quest Informatics [10] may well be the currently largest clinical data repository. Each week, 20 million new test results are loaded, and the system is predicted to break the terabyte (1000 GB) barrier in the near future, making it a very large DW by any standard. The system is used by Quest Informatics to turn the vast amount of lab results received from Quest Diagnostics Incorporated into valuable knowledge. This information is then employed by the users of Quest to improve their healthcare delivery. This is done by offering summary data, including comparisons between the users' patients and a standard patient population, thus identifying the broad areas of improvement. The user can then drill down to more detailed levels, getting the specific information about where improvements may be possible. The demands to the system are very tough; it must be very flexible to meet changing customer requirements, while still being able to perform effectively on the vast amounts of data.

3.7 Turku University Central Hospital

One of the only European efforts in the area is reported by Turku University Central Hospital in Finland [11]. The system integrates data from several hospital and laboratory information systems to provide a broad view of clinical data suitable for research. The system is comprised of a data transportation tool, a data warehouse database, and a proprietary front-end query tool. The DW has primarily been used in two studies, one on drug-laboratory interference and one on drug-drug interactions. The first study is concerned with the interference of drug prescription with the thyrotropin (TSH) test and determines how often drugs affecting the TSH test are prescribed. In fact, 11.6% of drug prescriptions turned out to interfere with the TSH test. The second study concerned drug interactions for nephrological patients, and the results were used in developing a drug interaction reminder system. An interesting issue in clinical data warehousing is noted, namely that data protection is very important due to the sensitive nature of clinical data, as opposed to typical business data warehouses. The project concludes that the presence of a clinical data warehouse is very facilitating for clinical research, as it provides an easy way to get precise answers to questions that otherwise often would not even have been asked due to the difficulties in getting to the data.

3.8 Stanford Medical Informatics

The efforts done at the Department of Medical Informatics at Stanford University are highly relevant to clinical data warehousing, although their approach to clinical data analysis is not strictly in the DW category.

An important aspect in any DW is integration of heterogeneous source data. This problem has been treated by the TransFER [14] project, which has developed a method for a query on a single, integrated, global schema to be translated to queries against the relevant local schemas, combining the results in the end. However, unlike the data warehouse approach, data in the global schema is not materialized, but only kept in the local databases.

Support for clinical treatment protocols is very important in clinical information systems. The EON project [15] has studied how to formally represent and reason about clinical protocols. A domain model of clinical concepts, suitable for protocol-based care has been developed, along with appropriate problem-solving methods. Representation and reasoning about *time* plays a very central role in clinical systems, and as part of the EON project, the Tzolkin Temporal Data Management System [16] has been developed

to allow temporal abstractions of raw clinical data, i.e., identify the time periods when certain clinical generalizations are true.

4 Discussion and Summary

In this section, we discuss the relative merits of the products and projects covered in Section 3, look at future prospects, and offer concluding remarks.

The DW at Turku University solves a specific problem, but does not have general applicability, and is thus not interesting for the general customer. Clinical Computing offers a good system, but it is limited to the renal domain. Quest Informatics has an impressive system in terms of data volume, but it is still a proprietary system designed specially for the company. Oracle Clinical and SAS/PH-Clinical both are feature-rich products targeted at an entire industry. Especially the SAS product seems to offer a high level of functionality. Their drawback is that they are not optimized for ordinary clinical functions, but rather for the highly specialized process of pharmaceutical drug development. The work at Stanford Medical Informatics is very interesting and touches on many of the core issues in clinical data warehousing. However, data warehousing as such is not treated, and no products are offered. It is the systems from MEDai, IAI, and SMS that are the most interesting. MEDai employs advanced AI techniques for data analysis, including predefined disease studies. SMS offers the widest range of clinical operational systems, and thus achieves very good integration with operational data, along with advanced functionality such as support for protocols and rules. IAI perhaps has the most "true DW" offering, with their integration of administrative, financial, and clinical information, supported by integrated analysis tools. Selecting one system out of these three must be based on the particular needs of the customer.

The future of clinical data warehousing looks very bright indeed, provided that the systems of tomorrow can fulfill the real needs of clinicians for data analysis, most of which are not well supported by current commercial products. The needs include a richer data model for capturing more of the semantics of the data, advanced temporal support to allow for analysis of data that change over time, support for advanced queries on continuously valued data, e.g., advanced statistics, and intelligent integration of very complex data, e.g., x-rays, in the DW for analysis purposes. Support for reducing the complexity of the data while still maintaining the essence is also very much needed, as well as means of handling the advanced classification structures employed in medicine. The concepts of clinical treatment protocols and medical research should also be tightly integrated into the clinical DW.

To summarize, we have introduced the concept of a data warehouse and described how it is used in a clinical setting. We have described the efforts in the area of clinical data warehousing and seen what products might be interesting for the general clinical customer. We think that the use of data warehouses in clinical settings will explode in the coming years, as systems mature and the clinicians realize the potential of using their data for quality improvement and research. We are currently working to meet some of the important challenges to data warehousing provided by clinical applications.

References

[1] B. A. Devlin and P. T. Murphy. An Architecture for a Business and Information System. *IBM Systems Journal*. Vol. 27, No. 1, 1988

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

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

[4] http://alliance.oracle.com/cat-doc/html/p18409.htm current as of July 18, 1997

[5] http://www.oracle.com/corporate/press/html/PR011397.110413.html current as of August 21, 1997.

[6] http://www.sas.com/new/preleases/050797/news1.html current as of July 21, 1997

[7] http://www.sas.com/software/ind/health.html current as of July 21, 1997

[8] http://www.sas.com/software/app/oracle.html current as of July 21, 1997

[9] http://www.medai.com/cdss.htm current as of July 18, 1997

[10] http://www.sybase.com/inc/success/quest.html current as of June 27, 1997

[11] Niinimäki J, et al. Medical Data Warehouse, an Investment for Better Medical Care. *In Proceedings of Medical Informatics Europe 1996*, IOS Press 1996.

[12] http://www.info-arch.com/IAI2-HVIW.html current as of August 11, 1997

[13] http://www.smed.com/products/product2/nov-ihn.htm current as of September 30, 1997

[14] http://www-smi.stanford.edu/projects/helix/hetero.html current as of June 25, 1997

[15] http://www-smi.stanford.edu/projects/eon/index.html current as of June 25, 1997

[16] http://www-smi.stanford.edu/projects/eon/tzolken.html current as of June 21, 1997