Medical Infobahn for Europe
A. Hasman et al. (Eds.)
IOS Press, 2000

Data Warehouse and Data Mining in a Surgical Clinic

Guenter TUSCH, Margarete MÜLLER, Katrin ROHWER-MENSCHING,
Karlheinz HEIRINGHOFF and Jürgen KLEMPNAUER
Medical School Hanover, 30623 Hannover, Germany

Abstract. Hospitals and clinics have taken advantage of information systems to streamline many clinical and administrative processes. However, the potential of health care information technology as a source of data for clinical and administrative decision support has not been fully explored. In response to pressure for timely information, many hospitals are developing clinical data warehouses. This paper attempts to identify problem areas in the process of developing a data warehouse to support data mining in surgery. Based on the experience from a data warehouse in surgery several solutions are discussed.

1. Introduction

A data warehouse can potentially provide enormous benefit to a hospital or a clinical department in clinical research, quality improvement, and decision support by enabling quick and efficient access to information from legacy systems and linkage to departmental databases. Although the same database can be used for real-time clinical data transactions and for aggregate data analysis and research, an increasing number of institutions have a separate data warehouse for the analysis functions. For hospitals with a strong research community, and for any hospital's management reporting needs, a separate warehouse may be worth the investment. The data warehouse platform being optimised for user-driven queries (in SQL or similar form), does not need to provide real-time performance, and performs its data-intensive work without impairing the performance of the computer-based patient record (CPR) or hospital information system (HIS). The data warehouse gets some of its data directly from the CPR or HIS through batch transfers during off-hours; it is also fed by other data sources not related to the CPR or HIS. The usability of the warehouse depends on how well the data from various sites are matched. This paper attempts to identify problem areas in the process of developing on-line analytical processing (OLAP) capacity from data generated in an on-line transaction processing (OLTP) system (the CPR or HIS) based on experiences from a surgical clinic.

2. Fundamentals of Data Warehouses in Hospitals

Data warehouse development is a continuous evolution and refinement process [1] driven by user needs. This requires rapid application development and prototyping techniques. Four types of users can be identified for a data warehouse [2,3]: 1. The *novice* or the *casual user* will in almost all cases be satisfied by canned structured reports. 2. The *analyst* needs some guidance to use the system, but also wants to generate ad hoc queries using parameters. The result may be a database or a spreadsheet that he/she uses to generate his/her own reports or graphics. He/she might also use the data as input to a statistical analysis program like SPSS (SPSS Inc., Chicago, IL) to perform simple statistical analyses. 3. The technically sophisticated analyst ('*power user*') uses SQL queries and processes the results in SPSS or SAS (SAS Institute, Cary, NC). Finally, 4. the *system planner* and *application developer* performs very sophisticated queries and statistical analyses in a similar way as the power user to test, validate, tune and observe the behaviour of the data warehouse.

The focus of the users in the medical domain when querying the data warehouse can be manifold: 1. Clinical research, publications or grant applications; 2. outcomes evaluations and assessment; 3. treatment patterns and protocols; 4. analysis of utilisation of resources; 5. medical management and cost containment studies, and 6. educational purposes. See also for usage the University of Virginia Health System [4].

Because requirements cannot be given for a data warehouse, a process model is not feasible [1]. This reinforces the importance of the data model. W.H. Inmon, regarded as the founder of the data warehouse concept, has characterised the data in a data warehouse as being subject oriented, time variant, non-volatile, and integrated [1]. 'Subject oriented': Data is organised by subject to facilitate easier reporting. For example, information about patients can be stored in one table, information about operations in another, making queries much easier because data is stored in one place. 'Time variant': Time is implicit in the data warehouse and can be used for trend analysis, while an operational system always represents a 'snapshot' of hospital activity. A data warehouse represent a regular sequence of these 'snapshots'. 'non-volatile': The information repository in the data warehouse is only to be read, and not to be modified. The only modification is adding a new 'snapshot'. 'Integrated': When loading data into the warehouse, it must be integrated into a consistent structure that meets the querying needs of the hospital or clinic. Inconsistencies between operational systems have to be eliminated. Information is also structured to varying levels of details (granularity) to support user queries. This is best achieved by a star schema. (See e.g. [2].) Often the data warehouse is split up into several *data marts* for serving specific user needs, e.g. data from a specific clinic or section of a clinic.

An important aspect of the data warehouse environment is that of *metadata*. Metadata allows the end user to navigate through the possibilities. Typically, metadata include the structure of data as known to the programmer and/or to the analyst, the source data feeding the data warehouse, the transformations of data as it gets into the data warehouse, the data model, the relation between the data model and the data warehouse, and the history of 'snapshots' that are put into the data warehouse.

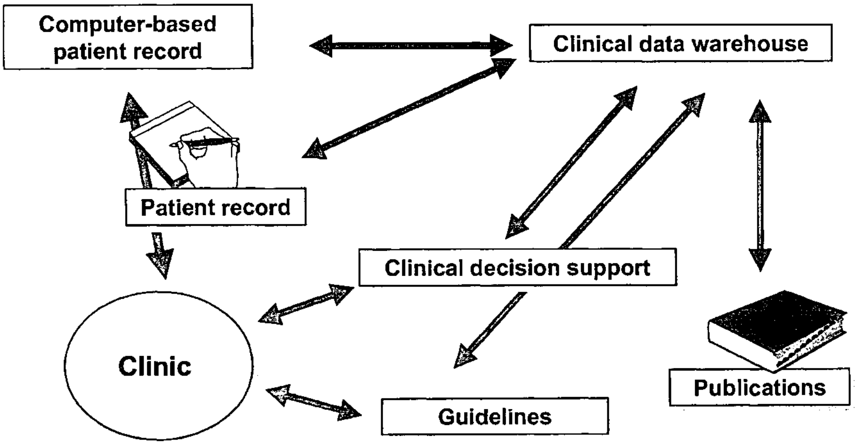

Figure 1: The mutual dependent communication structure in a clinic.

Using the data warehouse for data mining and/or statistical analysis creates new medical knowledge, e.g. in form of publications, clinical decision support or guidelines, which itself results in a feedback to new queries, different patient management or clinical procedures. This forms a mutual dependent communication structure in a clinic and requires a continuous reengineering of the data warehouse. (see figure 1.)

3. An Example of a Surgical Clinic

In the Clinic for Visceral and Transplantation Surgery at the Medical School Hanover more than a decade ago a data warehouse project was started [5]. The migration path of the data warehouse project is from MS FoxPro and MS ACCESS to ORACLE in the near future. Performance problems for larger databases using MS ACCESS are reported by several authors (e.g. [6]). Administrative data and admission diagnoses come from the ADT legacy system, surgery data from the operation theatre information system, laboratory data from the clinical chemistry information system, outpatient data from the outpatient computer-based patient record and data from a departmental clinical database (including survival and other information that is not available in the hospital information system). Query results can be stored as a database table or in a spreadsheet, and can further processed for data mining/ statistical analysis in SPSS or SAS. Today, the system predominantly supports casual and power users, for the analyst user a Web access is under construction. The data warehouse has been built as a bottom-up approach from several data marts including data from liver transplantation (approx. 1,500 transplantations - approx. 78,000 patient days), liver resection (> 4,000 resections), kidney and pancreas transplantation (> 3,000 transplantations). Regular reports are mailed to Eurotransplant, Leiden or ELTR, Paris. From this experience we will identify problem areas for data warehouses and data mining in surgery and discuss several solutions.

4. Problem Areas for Data Warehouses and Data Mining in Surgery

4.1 User Groups – User Interfaces

There is a widespread agreement among authors that the analyst user is most challenging for the data warehouse developer (e.g. [1,2,3]). This is because he/she needs guided access and want a free choice of focus. Here a WWW query interface is most promising to get a relatively soft- and hardware independent access. A flexible approach has been established and evaluated at the University of Virginia Health Sciences Center [7,8,4]. However, the evaluation shows that the problem of the user interface has to be accessed from different perspectives. For difficult tasks it will be necessary - also in the near future - to provide human assistance serving as an information mediator to the querying clinician.

4.2 The Data Model

There is a trade-off concerning the data model. On the one hand all information that might be of interest for the clinician or administrator should be in the data warehouse. On the other hand information not available in the OLTP system in a directly accessible form (e.g. hidden in text fields as a result of bad database design) will bind human capacity on every update of the data warehouse. This is true both for systems bought from a vendor or systems made in the own institution. A reasonable approach has to be an optimum in achieving a maximum coverage of the user's information need within an established cost frame. For surgical queries not only (time) granularity is important but also the subject level: Patient, operation or day level.

4.3 Data Mapping and Transformations (Metadata)

In the medical domain, a crucial point is integration and cleansing of the data. Data from two different patients must not be merged in any case. Data may be missing, incomplete or otherwise inadequate [9], and current medical vocabularies are rather limited in expressing medical concepts [10]. Furthermore, often medical data are context-sensitive. These problems have to be taken into account for clinical research and therapy based on results from a data warehouse. Therefore, it is crucial for safety (and ethical) reasons that automatic procedures are complemented by human inspection of critical data. In many cases, the selection of critical data itself can be controlled by methods of data mining saving human resources.

Thus, data transformations include automatic, semi-automatic and manual transformations. Automatic transformation are feasible when the content of fields is identical but the field names or codes vary in different applications and can be mapped one-to-one. Semi-automatic often means a lot of manual intervention, e.g. when information has to be extracted from text fields. Sometimes a transformation is hardly possible, e.g. when the coding system for admission diagnosis was switched from ICD 9 to ICD 10 in Germany this year.

A medical data warehouse incorporates at least two concepts of time: Calendar time given by different 'snapshots' of the OLTP system, and individual patient time e.g. survival time after surgery. A promising approach for temporal clinical queries based on patient time is to define temporal concepts [11]. Also a clear identification of missing information is necessary to achieve high quality data. Data quality means adressability, domain, entity and data integrity, data consistency etc. To check data quality also statistical techniques are used [12].

4.4 Data Update

After the initial load has been performed, there are two possible approaches to updating the data warehouse: 1. Complete data refreshment and 2. changed data capture. Complete data refreshment is feasible only for smaller data marts. Changed data capture makes a sophisticated data management necessary to make sure that changes in data already loaded to the data warehouse will be noticed. This might affect especially lab data, e.g. from microbiology.

4.5 Data Mining and Statistical Analysis

Classical data mining technique in a data warehouse or Executive Information System (EIS) is the drill-down analysis. This technique surely is adequate for management queries and queries e.g. to explore trends in medical supply. Data warehouses can be viewed at as a statistical database and the general limitations including biases in the retrospective analyses

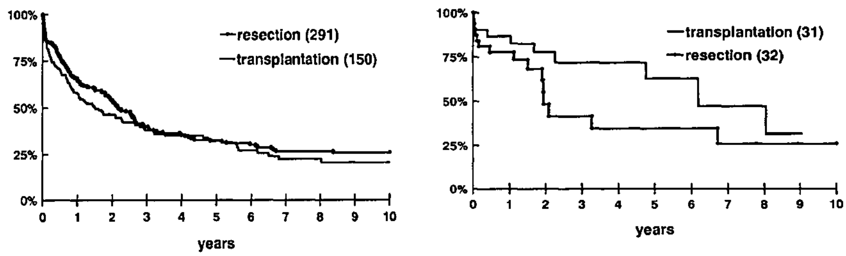

Figure 2: Typical subgroup analysis in liver surgery (HCC patients - from [15]).
Left survival curves: survival after liver resection shows an advantage over survival after transplantation; right survival curves: in the subgroup UICC stage I-II with underlying cirrhoses this relationship is inverted.

apply (e.g. [13,14]). In surgery the classical investigation is survival analysis complemented by subgroup analysis (see figure 2 for an example). Other statistical methods are used as well. This can be accomplished by the statistical packages SPSS for Windows and SAS.

4.6 Confidentiality Issues

When providing web access the data access is to be restricted by logon id and password. Some institutions disguise the identifying information of the patients to the average user to avoid additional encryption (e.g. [8]). This is reasonable for users with administrative queries in mind. For clinical queries project specific restrictions to patient groups (as far as they can be identified) should apply. Here confidentiality regulations have to be obeyed.

5. Conclusion

Data warehouses allow users to access clinical information in a cleaned environment, free from impediments of on-line transaction processing (OLTP) systems. Unfortunately, in OLTP systems often clinically important data is stored in a format that makes human intervention necessary before transferring the data to the data warehouse. This leads to a significant increase of costs for these systems in the hospital environment. The aim has to be to put more structure to the data in the computer-based patient record to facilitate the data entry to the data warehouse. Thus data warehouses can help to improve both medical knowledge and the quality of medical documentation.

Acknowledgement

We would like to thank the many people who helped to make the effort of our data marts possible, especially Prof. J. Bernauer, Prof. A. Weimann, R. Weiß, P. Wübbelt.

References

[1] W.H. Inmon, Building the Data Warehouse. 2. ed. ISBN: 0-471-14161-5. Wiley, New York, NY, 1996.

[2] V. Poe *et al*, Building a Data Warehouse for Decision Support. 2. ed. ISBN: 0-13-769639-6. Prentice Hall PTR, Upper Saddle River, NJ, 1998.

[3] R.C. Barquin and H.A. Edelstein (eds.), Building, Using, and Managing the Data Warehouse. ISBN: 0-13-534355-0. Prentice Hall PTR, Upper Saddle River, NJ, 1997.

[4] J.R. Schubart *et al.*, Evaluation of a Data Warehouse in an Academic Health Science Center. ISBN: 1-56053-371-4. Proc. AMIA Annual Symp. 1999, Hanley & Belfus, Inc., Philadelphia, 1999, pp. 614-618.

[5] G. Tusch *et al.*, Ein departmentelles Informationssystem zur Unterstützung der Lebertransplantation. In: O. Rienhoff *et al.* (eds.), Expert Systems and Decision Support in Medicine. ISBN: 0-387-50317-X. Springer-Verlag, Berlin, 1988, pp. 509-515.

[6] A. Ebidia *et al.*, Getting Data Out of the Electronic Patient Record: Critical Steps in Building a Data Warehouse for Decision Support. ISBN: 1-56053-371-4. Proc. AMIA Annual Symp. 1999, Hanley & Belfus, Inc., Philadelphia, 1999, pp. 745-749

[7] K.W. Scully *et al.*, Evaluation of a Data Warehouse in an Academic Health Science Center. Proc. AMIA Annual Symp. 1997, Hanley & Belfus, Inc., Philadelphia, 1997, pp. 32-36.

[8] K.W. Scully, A Flexible WWW Query Interface for a Patient Data Warehouse. Proc. AMIA Annual Symp. 1998, Hanley & Belfus, Inc., Philadelphia, 1998, p. 1078.

[9] G. Hripcsak *et al.*, Unlocking Clinical Data from Narrative Reports, *Ann. Intern. Med.* **122** (1995) 681-8.

[10] J. Cimino, Desiderata for Controlled Medical Vocabularies in the 21st Century, *Methods Inf. Med.* **37** (1998) 394-403.

[11] G. Tusch *et al.*, A Knowledge-Based Decision Support Tool for Liver Transplanted Patients. In: B. Barber *et al.* (eds): Medinfo 89. ISBN: 0-4444-88138-7. North-Holland Publ., Amsterdam, 1989, pp. 131-135.

[12] H.D. Stein *et al.*, Exploring the Degree of Concordance of Coded and Textual Data in Answering Clinical Queries from a Clinical Data Repository, *JAMIA* 7 (2000) 42-54.

[13] J.M. Dambrosia and J.H. Ellenberg, Statistical Considerations for a Medical Database, *Biometrics* **36** (1980) 323-332.

[14] D.P Byar, Problems with Observational Databases to Compare Treatments, *Stat. Med.* **10** (1991) 663-666.

[15] A. Weimann *et al.*, Is liver transplantation superior to resection in early stage hepatocellular carcinoma? *Transpl. Proc.* **31** (1999) 500-501.