Proc Int Conf MedInfo, San Francisco CA, AMIA, 2004 (in press)

Data Mart based Research in Heart Surgery: Challenges and Benefit

Bert Arnrich [a], Jörg Walter [a], Alexander Albert [b], Jürgen Ennker [b], Helge Ritter [a]

[a] *Institute of Neuroinformatics, University of Bielefeld, Germany,* mailto:heart@Joerg-Walter.de
[b] *Clinic for Cardiothoracic Surgery, Heart Institute Lahr/Baden, Germany*

Abstract

For many new medical research questions in heart surgery comprehensive and large data bases are essential. We discuss typical challenges for the integration of real-time and legacy data stored in multiple unconnected hospital information systems (HIS). Furthermore the HIS are often operated by autonomous departments whose data base structures are subject to occasional modifications.

We present a solution which integrates and consolidates all research relevant data in a data mart without imposing any considerable operational or maintenance contract liability risk for the existing HIS. The problems of partial consistency and partial redundancy in the data are discussed.

The data mart system serves multiple purposes: beside clinical reporting and quality assessment, the preparation steps for comprehensive studies are enormously simplified.

Keywords:

Heart surgery; integration of hospital information systems; data mart;

Introduction

The increasing availability of information technology (IT) enables the implementation of a fast growing variety of applications and heterogeneous information processing systems also in the medical domain [1]. We find them often disconnected and distributed across several departments.

On the other hand it is more and more recognized that those huge data collection can be very valuable if properly maintained and consolidated. From the scientific viewpoint a prospective, double blind, randomized study is the best method to gain new insights, but it is also the most expensive and time consuming procedure. Since medicine is a rather mature disciple, progress is often made in fields with very subtle interactions and rare constellations. This requires the collection of large data sets.

In this paper we report on a data mart based information system which aims at the support of:

- easy comprehensive medical research;
- quality assurance;
- preoperative risk assessment;
- hospital management (risk adjusted inter- and intra-clinical comparison).

In the following we supply the demand for reports on empirical datamining approaches from the viewpoint of medical data [2]. We describe the construction of our data mart system and discuss challenges and benefits in the particular application domain of heart surgery.

Materials and Methods

Heart operations are the most frequent realized surgeries in Europe and North America. *The Clinic for Cardiothoracic Surgery of the Heart Institute Lahr* is a highly specialized hospital and performs about 2000 open heart operations per year. The majority of cases fall in a rather small spectrum.

The main surgery indication is summarized as cardiovascular disease (CVD) which includes coronary heart disease, high blood pressure, atherosclerosis, and stroke. CVD is an important research topic since it causes nearly half of all deaths in Europe [3]. The major group of CVD is coronary heart disease. Here the coronary arteries that bring blood to the heart muscle are constricted by plaque. If an alternative treatment like medical therapy or balloon angioplasty is not adequate, a coronary arteries bypass grafting (CABG) is indicated. In CABG the surgeon takes a healthy blood vessel from another part of the body and connects multiple bypasses around the blocked parts of the coronary artery.

The second most frequent operation type is concerned with the mechanical reconstruction or replacement of defect heart valves. In Lahr the proportions are 70% isolated CABG, 15% isolated valve surgeries, 10% combined operations, and 5% miscellaneous, even more complex operations.

For historical reasons the Heart Institute Lahr choose to operate independent clinical information systems since its beginning in 1995. At that time these software products were state-of-the-art and they constitute significant investments.

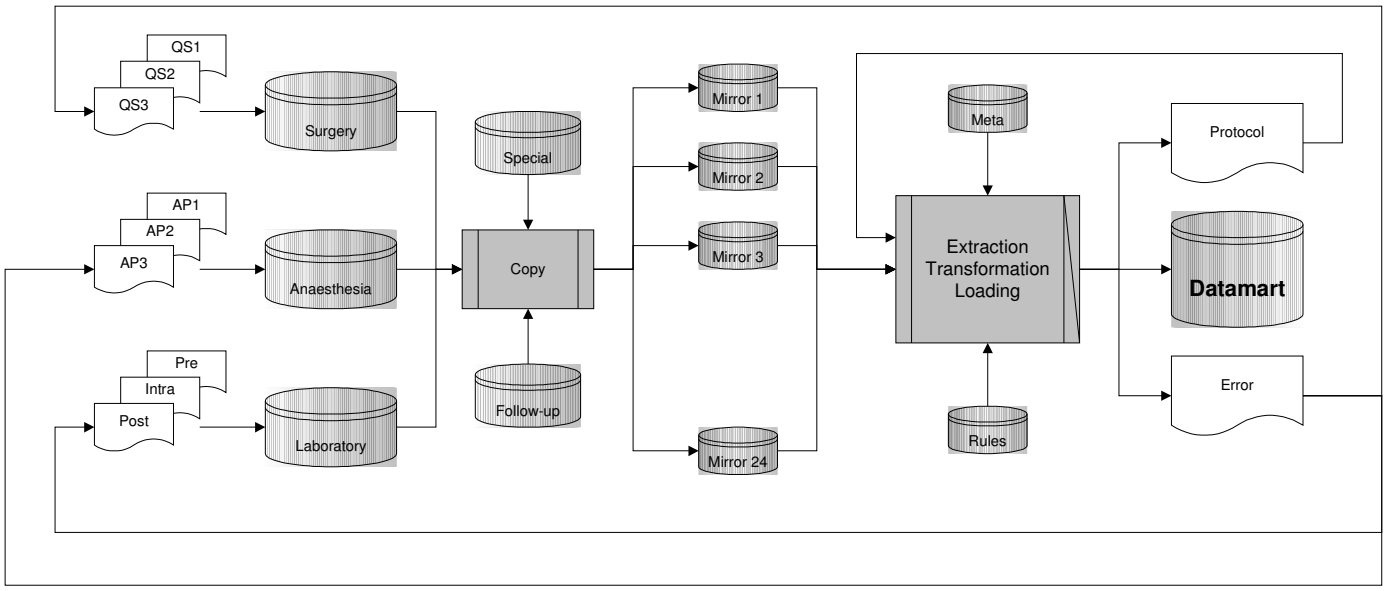

Figure 1- The data mart system mirrors, extracts, and consolidates all research relevant data from the existing, but disconnected HIS with the additional use of patient data from several special studies. Up to now 28 source tables adding up to 389 attributes are used in the data mart system. Inconsistencies of the source data can be reported with correction suggestions. Currently the data mart database contains 277 pre-, intra- or postoperative attributes.

The following circumstances are not untypical: during the time a patient stays in the hospital, several departments collect case specific data with (partly) different objectives in "there" HIS:

1. *surgical data base*: medical history, planned surgery, operative strategy, postoperative events, medication, etc.
2. *anaesthesiological data base*: preoperative medication, renal function, neurology, blood circulation parameters, postoperative medication, breathing, etc.
3. *clinical chemistry data base*: more than 60 pre-, intra- and postoperative laboratory parameters, such as white blood cell count, cholesterol, etc. (with time stamp);
4. *administrative data base*: accounting information etc.

Additionally data collections from special medical studies are stored in various file formats (e.g.: a collection of more than 40 new parameters for all patients suffering a stroke during or after cardiac surgery).

Typical Problems and Challenges

Some of the following issues for building a comprehensive research oriented medical database are ubiquitous, some are institution dependent:

1. *isolated data sources,* mainly in *disconnected HIS* operated by *autonomous departments* (the HIS are not built with the intention to support easy interoperability);
2. data with *partial redundancy* and *partial consistency*;
3. *departments prefer to retain autonomy, minimize work flow risk and protect previous investment* (due to liability and maintenance contract regulation any changes to the proprietary HIS are rather difficult);
4. *privacy protection regulations* must be obeyed;
5. *legacy data* can be very valuable. The integration of all potentially useful data requires conformation of all relevant changes in the data base structures in the history of all HIS (due to software updates, report form changes, etc.; fortunately the HIS data bases usually retain old data sets) and special file formats.

These challenges are met with the following concept.

A Solution: the Data Mart System

Figure 1 illustrates the chosen concept. All departmental HIS are left unchanged, thus the risk of operational effects are minimal (see 1+3). Only read access to the three major data bases is granted. Mirror processes copy the relevant relational data base tables. Any patient related personal data is not affected, only pseudonym identifiers are employed in the data mart system. For security reasons the target computer resides in an isolated area of the intranet (DMZ).

The first step in the construction of the data mart system is the selection and detailed documentation of the suitable source databases and tables. Here a availability of a domain expert is essential for effective progression of this crucial and also time consuming part. He also leads the quality assessment and the diagnostic precision rating of various attributes. The results are the basis for data understanding, and the following two sub stages: design of the inconsistency detection and transformation rules.

Inconsistency detection

The first stage in the extraction/transformation process is devotes to securing the data quality. Inconsistencies are detected with numerous plausibility checks. As displayed in Figure 1, inconstancy reports and, if applicable, also correction suggestion are generated. Conflicts must be solved by the health professionals who also correct the primary data bases. All modifications are automatically transferred to the data mart system. By this means data consistency is achieved throughout the system in a persistent manner.

Rule-based Transformation

For the next transformation stage the data integration rules are defined. These rules describe in the simplest case a copy process or an univariate transformation from mirror to target data base. The results of the attribute quality assessment are used to define responsibility chains in case of attribute redundancies.

Sometimes the combination of various data is logically unique (e.g. for all laboratory values the operation time is needed for classification of the time sampled measurements into pre-, intra- and postoperative values; see also middle part of Figure 2). In other cases the translation from several source values to construct the semantic meaning of a target value is required. For example the attribute *critical preoperative state* is gained from 15 different source values (see also top of Figure 2: the data mart attribute *circulatory disorder* is derived from surgical and anaesthesiological data). These rules are carefully designed in tight cooperation with domain experts (see next section).

Furthermore, due to the historical changes in the data base structures, entire transformation rules can be dependent on the time of original data recording.

Tool support for Verification

The use of various sources with historical changes in the data base structure requires an effective tool for the inspection of the whole integration process. During the data mart assembling all relevant source values for each attribute and for each case are stored in a verification database. We developed a web-based inspection tool for the inspection of various aspects of the extraction/transformation process on the basis of an individual case, a group, or a specific rule. Figure 2 presents a screen shot of a case-based verification for three data mart parameters.

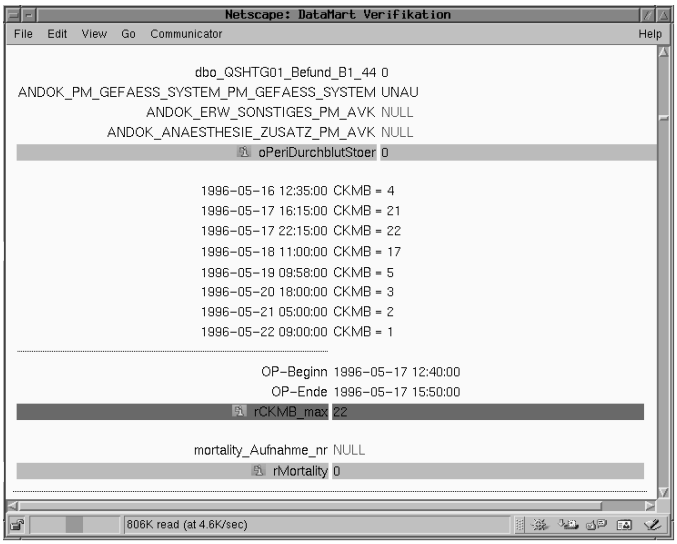

Figure 2 - Example of a case-based verification for three data mart parameters. [Top:] circulatory disorder (oPeriDurchblutStoer) is build from the surgical (first item) and anaesthesiological (remaining three items because of three HIS software updates) database; the values "0", "UNAU" and the last two missing marker "NULL" (case was recorded with the first version of the HIS software) result in 0 which denote no circulatory disorder. [Middle:] determination of the maximum value of the blood enzyme CKMB (rCKMB_max) after end of operation at "1996-05-17 15:50:00". [Bottom:] mortality status (rMortality) is 0 since the case does not appear in the mortality list.

Tool support for Missing Value Analysis

Missing data are a familiar problem on the realization of medical studies based on patient data. The ad hoc methods for analyzing incomplete data focused on ignoring subjects with incomplete items or substituting plausible values[7]. Correlation between the existence of missing values at one attribute and the characteristic of another may produce a significant bias in the analysis results.

The usage of common statistic tools for the bi-variate missing data analyzes of already a moderate number of attributes produces an overwhelming amount of output. Here we need a compact presentation of the observed relationships. We developed a web-based tool that allowed us to inspect the distributions of a set of attributes if the others are missing (see Figure 3). The results of all available statistical procedures (e.g. Chi-Square, Student-t, entropy based measurements, etc.) are displayed in a condensed form - partly with inline or hyperlinked graphics.

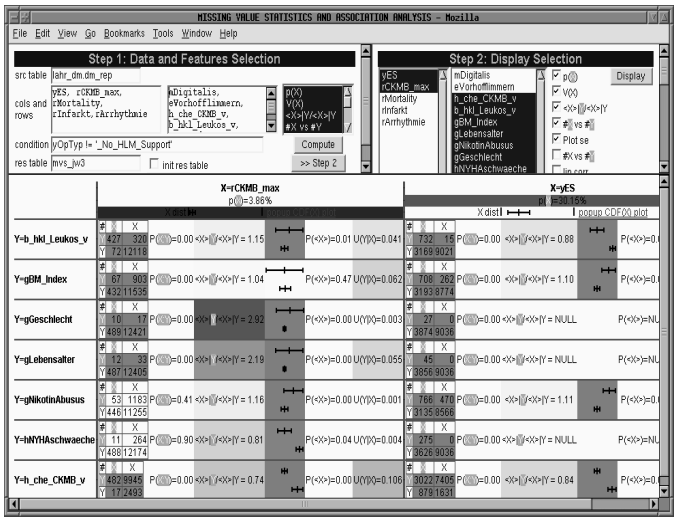

Figure 3 - Missing-Value-Statistic tool for the inspection of distribution parameters of attribute X (column) if attribute Y (row) is missing. In the upper left and right frame attributes, methods and display content can be selected. Each table-cell in the upper frame contains in this example three items: contingency table with p-value, average ratio with visualization and p-value and conditional entropy measure.

Results

Up to now, data from more than 13.000 heart operations with 277 pre-, intra- or postoperative attributes per case are available for multiple purposes.

Web based Information portal

Registered user can access the data mart system via a web-based information portal in the intranet. Three main categories are available: data export for medical research, online reports and performance visualization for clinical reporting.

1. Subsets of the consolidated data set can be selected and are exported after authorization from the data mart administration. Here the patient pseudonym is replaced with a cryptographic one-way hash code in order to fully anonymise the data set. Hence the time and effort consuming collection, preparation and consolidation steps for retrospective, more comprehensive studies are no longer hindering medical research.
2. For performance monitoring various online reports can be generated. More frequent selections (e.g. on time ranges, operation types, surgeons) are directly linked for authorized users (clinical management, director of department, individual surgeons only for their own and aggregated data).
3. Risk adjusted temporal performance graphs in selected subgroups are dynamically generated. and can be visualized.

In the following three examples of medical results with emphasis on the benefit of the data mart system are reported.

Prevention of stroke

Medical background: Stroke is the second most important cause of mortality and morbidity in the western world. During and after cardiac surgery the risk of stroke is increased mainly due to manipulation of the heart and the brain supplying arteries. Well known risk factors for stroke are known, however in recent years investigators focus on hemorheological factors, which may contribute to the development of stroke.

Clinical situation: Stroke in cardiac surgery is still a devastating complication.

Data mart benefit: The occurrence of stroke in the heart institution Lahr is fortunately limited at 1,5% of all cases. For these rare events only a large number of cases allow the detection of significant relationships. With the integration of the historical data and all measured laboratory values of a patient during the whole hospital stay in the data mart system, we can analyze blood cell alterations before the onset of stroke. This is particularly valuable since blood measurement data are rarely available shortly before suffering a stroke. By this means we can contribute to the understanding of stroke in general. In [4] we could demonstrate for the first time a significant correlation between high preoperative white blood cell count (WBC) and stroke during or after cardiac surgery.

Surgical quality assessment

Medical background: Assessing the quality of cardiac surgical care through inter-hospital and inter-surgeons comparison of mortality rates after cardiac surgery is of increasing importance. For a fair comparison the differences in patient case-mix between different institutions must be taken into consideration in the relevant statistical analyses. One of the established risk score systems for postoperative mortality in Europe is the European System for Cardiac Operative Risk Evaluation (EuroSCORE) [5]. It includes 17 different parameters to assess the individual risk.

Clinical Situation: Current methods used to adjust mortality rates by preoperative status cannot adequately explain whether different results from different institutions are due to differences in patient severity or quality of care. Fair comparison requires accurate scoring models which predict mortality and morbidity outcomes from preoperative, objective risk factors.

Data mart benefit: The data mart database has enabled us to apply the EuroSCORE retrospectively in the heart-center in 75% of all cases. After calibration and adaptation we can perform risk adjusted inter-hospital and surgeon comparison with higher accuracy as current methods. Furthermore we have studied the role of age as a determinant of mortality in cardiac surgery in our institutional patient population [6].

Renal impairment and Creatinine Clearance (CC)

Medical background: The renal filtration capacity is a vital factor especially in heart surgery.

Clinical situation: Renal impairment is not well captured in the standard EuroSCORE risk evaluation system where usually a binarized value of serum creatinine is used to assess the renal function.

Data mart benefit: The derived parameter CC can be estimated from serum creatinine, gender, age and body weight. We could show that the CC estimation can advantageously replace the serum creatinine in the EuroSCORE preoperative risk assessment. Variable rank comparison identified CC as the best single variable predictor [8].

Discussion

The presented data mart architecture proved useful and effective. It integrates the current and historical data from all relevant data sources without imposing any considerable operational or liability contract risk for the existing HIS's. By this means the potential resistance of involved persons in charge can be minimized and the project specific goal effectively met.

This approach allows to turn redundancies into value. On the one side, redundancies are used to detect inconsistencies within and across departmental data bases. On the other hand they allow to a certain extend to derive attributes from data sources which originally do not contain the desired semantic definition. With increasing number of available partly redundant attributes the freedom for formulation of transformation rules increases. Good verification tool support helps to effectively refine the required rules while the missing value analysis assists in evaluating potentially helpful attribute combinations.

In the past the possibilities to perform retrospective comprehensive studies in the heart center was extremely time consuming and therefore limited. Attempts were already made to extract and combine data from the different HIS. Dependent on the desired scientific task, the queries to extract and connect the data were often rebuilt and modified. Consequently the semantics and definitions of the jointed data changed from one study to the other. Additionally, due to the temporal changes of the data base structures it was very difficult to maintain an overview over all versions of performed queries and derived data sets. With the implementation of the presented data mart system the time and effort consuming correction process could be replaced and the research basis remains stable and leads to reproducible results.

Conclusion

In this paper we presented typical problems and challenges for building large research oriented data bases found in the domain of heart surgery: patient based information is stored in several disconnected HIS operated by autonomous departments. We describe how a data mart can effectively meet the goals without any modifications to the existing HIS and any associated risk for the work flow and existing software maintenance contracts.

Partial consistency of the data within and across departmental data bases is tackled by plausibility checks and by taking advantage of partial redundancy. The latter can be also useful to derive attributes with semantic definitions which are not present in the original data sources.

Furthermore we introduced three examples of medical results we were able to achieve due to the large number of consolidated and extensive data records.

Acknowledgments

We would like to thank all colleagues for their contribution in medical research and Holger Hussy for the network infrastructure support in the heart institute Lahr.

References

[1] Beuscart-Zephir MC, Brender J, Beuscart R, and Menager-Depriester I. Cognitive evaluation: How to assess the usability of information technology in healthcare. Comput Methods Programs Biomed. 1997 Sep;54(1-2):19-28.

[2] Tsumoto S. Clinical Knowledge Discovery in Hospital Information Systems: Two Case Studies. Principles of Data Mining and Knowledge Discovery. 2000: 652-656

[3] Rayner M, and Petersen S. European cardiovascular disease statistics 2000 edition. British Heart Foundation. 2000

[4] Albert AA, Beller CJ, Walter JA, Arnrich B, Rosendahl UP, Priss H, Ennker J. Preoperative high leukocyte count: a novel risk factor for stroke after cardiac surgery. Ann Thorac Surg. 2003 May;75(5):1550-7.

[5] Roques F, Nashef SA, Michel P, Gauducheau E, de Vincentiis C, Baudet E, Cortina J, David M, Faichney A, Gabrielle F, Gams E, Harjula A, Jones MT, Pintor PP, Salamon R, Thulin L. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg. 1999 Jun;15(6):816-22

[6] Mortasawi A, Arnrich B, Rosendahl U, Frerichs I, Albert A, Walter J, Ennker J. Is age an independent determinant of mortality in cardiac surgery as suggested by the EuroSCORE? BMC Surg. 2002 Oct 7;2(1):8.

[7] Schafer J, and Olsen M. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behavioural Research. 1997; 33: 545-571

[8] Walter J, Mortasawi A, Arnrich B, Albert A, Frerichs I, Rosendahl U, Ennker J. BMC Surg. 2003 Jun 17;3(1):4.