

B. Arnrich
J. Walter
A. Albert
J. Ennker

Herausforderungen und Nutzen eines Datamart-Systems in der Herzchirurgie

Eingegangen: 23. Oktober 2003
Akzeptiert: 28. November 2003

Prof. Dr. med. Dr. h.c. mult. Roland Hetzer
zum 60. Geburtstag gewidmet

Challenges and benefit of a datamart system in heart surgery

■ **Zusammenfassung** Der hohe Entwicklungsstand in der Herzchirurgie erfordert große Datenbanken für die Untersuchung neuer medizinischer Fragestellungen. In diesem Artikel diskutieren wir typische Probleme und Herausforderungen bei der Integration von aktuellen und historischen Daten, die in den unverbundenen und von autonomen Fachabteilungen betriebenen klinischen Informationssystemen (KIS) gespeichert werden.

Wir stellen ein Datamart-System vor, das alle für medizinische Forschung relevanten Daten integriert und konsolidiert. Dabei werden die Modifikationen in den Strukturen der Quelldatenbanken berücksichtigt und die Risiken im Arbeitsablauf der KIS minimiert. Der partiellen Konsistenz und Redundanz der Daten in und zwischen den Abteilungen wird durch Plausibilitätsregeln begegnet. Weiterhin können damit Attribute, die nicht in den originalen Datenquellen vorhanden sind, abgeleitet werden.

Der Datamart hat das klinische Berichtswesen, das Qualitätsmanagement und die für umfassende Studien nötige Datenaufbereitung beträchtlich vereinfacht. Wir beschreiben die Ergebnisse von drei medizinischen Studien deren Durchführung durch den umfangreichen und konsolidierten Datenbestand ermöglicht wurde.

■ **Summary** Since heart surgery is a rather mature discipline, large data bases are essential for many new medical research questions. In this paper we discuss typical problems and challenges which appear at the integration of real-time and legacy data stored in multiple unconnected hospital information systems (HIS) operated by autonomous departments.

We present a datamart system which integrates and consolidates all research relevant data, taking into account the modifications of the source data base structures. At this any considerable operational or maintenance contract liability risk for the existing HIS is minimized. The partial consistency and partial redundancy of the data within and across departments is tackled by plausibility checks and further used to derive attributes with semantic definitions which are not present in the original data sources.

With the datamart clinical reporting, quality assessment and data preparation for comprehensive studies are enormously simplified. We

Bert Arnrich (✉) · Jörg Walter
AG Neuroinformatik
Universität Bielefeld
Postfach 100131
33501 Bielefeld, Germany
E-Mail: heart@joerg-walter.de

Alexander Albert · Jürgen Ennker
Klinik für Herz-, Thorax-
und Gefäßchirurgie
Herzzentrum Lahr/Baden
77933 Lahr, Germany

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

■ **Schlüsselwörter** Herzchirurgie – klinische Informationssysteme – Datamart – konsolidierte Forschungsdatenbank

■ **Key words** Heart surgery – hospital information systems – datamart – consolidated research data base

Einleitung

Die wachsende Verfügbarkeit der Informationstechnologie ermöglicht in immer stärkerem Maße den vielfältigen Einsatz von Datenverarbeitungssystemen – auch in der medizinischen Domäne [1]. Häufig handelt es sich dabei um heterogene und nicht für eine Zusammenarbeit konzipierte Systeme, die in verschiedenen Abteilungen zum Einsatz kommen.

Gleichzeitig wird zunehmend erkannt, das große Datensammlungen bei entsprechender Konsolidierung und Nutzung sehr wertvoll sein können. Aus wissenschaftlicher Sicht ist eine doppelt blinde, randomisierte Studie die beste Methode neue Erkenntnisse zu gewinnen – allerdings auch die teuerste und zeitaufwendigste Prozedur, die gerade in der Herzchirurgie aus ethischen Gründen häufig nicht durchführbar ist. Durch den hohen medizinischen Entwicklungsstand befassen sich viele Forschungsfronten mit speziellen Fragestellungen und damit verbunden mit seltenen Merkmalskonstellationen. Die Erzielung von Fortschritten in diesen Bereichen erfordert meist große Datensammlungen.

In diesem Artikel berichten wir über ein Datamart-basiertes Informationssystem im Herzzentrum Lahr und zeigen die Einsatzmöglichkeiten in den Bereichen

- medizinische Forschung
- Qualitätssicherung
- präoperative Risikoschätzung

■ Klinikmanagement (risikoadjustierter Inter- und Intra-Klinik Vergleich).

Weiterhin erfüllen wir die Forderung nach der Beschreibung von empirischen Dataming Ansätzen basierend auf medizinischen Daten [2]. Wir beschreiben die Konstruktion unseres Datamart Systems und diskutieren die Herausforderungen und den Nutzen in der besonderen Anwendungsdomäne Herzchirurgie.

Methoden

Die Klinik für Herz-, Thorax- und Gefäßchirurgie Lahr ist eine hoch spezialisierte Einrichtung, die jährlich etwa 2000 Operationen am offenen Herzen durchführt. Dabei entfallen etwa 70% auf isolierte Bypassoperationen, 15% auf isolierte Herzklappeneingriffe, 10% auf kombinierte Operationen und 5% auf verschiedene andere, meist wesentlich komplexere Eingriffe.

Aus historischen Gründen arbeitet das Herzzentrum Lahr seit seiner Gründung 1995 mit unabhängigen arbeitenden klinischen Informationssystemen (KIS), die zur Zeit der Anschaffung auf dem neuesten Stand der Technik waren und beträchtliche Investitionen darstellten.

Während des gesamten Klinikaufenthaltes erheben verschiedene Fachabteilungen fallspezifische Daten mit teilweise unterschiedlichen Zielsetzungen und verwalten diese in „ihren“ KIS:

1. Chirurgische Datenbank (Qualitätssicherung Herzchirurgie): medizinische Basis-, Therapie- und Verlaufsdaten;
2. Anästhesiologische Datenbank: präoperative Anamnese und operativ-anästhesiologische Daten;
3. Klinische Chemie Datenbank: bedarfsorientierte Bestimmung von über 60 Laborwerten;
4. Administrative Datenbank: Rechnungswesen und Verwaltung.

Zusätzliche Datensammlungen in diversen anderen Formaten stammen von speziellen medizinischen Studien (z. B. eine Excel-Datei die für alle perioperativen Apoplexereignisse 40 weitere Merkmale erfasste).

■ Typische Probleme und Herausforderungen

Einige der folgenden Punkte die bei der Konstruktion einer umfassenden, auf medizinische Forschung fokussierten Datenbank beachtet werden müssen sind allgegenwärtig, andere beruhen auf Besonderheiten der jeweiligen Einrichtung:

1. isolierte Datenquellen die häufig in unverbundenen KIS von autonomen Abteilungen betrieben werden (die KIS unterstützen selten eine einfache Interprozesskommunikation),
2. redundante Datenaufnahme mit partieller Konsistenz,
3. Präferenz für Autonomie der Fachabteilungen, Minimierung des Arbeitsablafrisikos und Schutz getätigter Investitionen (durch Haftpflichtregelungen und Wartungsverträge sind Modifikationen an den bestehenden KIS nur sehr schwierig realisierbar).
4. Einhaltung von Datenschutzbestimmungen.
5. Integration der wertvollen Altdaten erfordert die Anpassung

aller relevanten historischen Änderungen der Datenbankstrukturen in allen KIS (ausgelöst durch Software-Updates, Erweiterung der Erhebungsbögen, etc.) und den speziellen Datenformaten.

Diesen Herausforderungen begegnet das im folgenden vorgestellte Konzept.

■ Eine Lösung: das Datamart-System

In Abbildung 1 ist das entwickelte Konzept illustriert. Sämtliche KIS in den Abteilungen bleiben hierbei unverändert. Durch den ausschließlichen Lesezugriff auf die Datenbanken wird das Störungsrisiko im Arbeitsablauf praktisch auf Null gebracht (siehe obige Punkte 1 und 3). Spiegelungsprozess kopieren alle relevanten Daten und speichern diese in Hilfstabellen. Hierbei werden keine personenbezogenen Daten, sondern nur pseudonyme Fallidentifikatoren verwendet. Aus Sicher-

heitsgründen befindet sich der Zielrechner in einer isolierte Zone des Klinikintranets (DMZ).

Der erste Schritt bei der Konstruktion eines Datamart-Systems ist die Sammlung und detaillierte Dokumentation der geeigneten Quelldatenbanken und Tabellen. Dabei ist die Verfügbarkeit eines Domänenexperten in dieser wichtigen und zeitaufwendigen Phase essentiell. Das Expertenwissen ermöglicht auch eine Bewertung der verschiedenen Quellattribute bzgl. Genauigkeit der medizinischen Diagnose und Aufnahmequalität. Die hier erzielten Resultate bilden die Basis für das Datenverständnis und die folgenden zwei Teilschritte: Entwurf der Inkonsistenzerkennung und der Transformationsregeln.

Inkonsistenzerkennung

Die erste Phase im Extraktions-Transformations-Prozess widmet sich der Sicherstellung der Datenqualität. Inkonsistenzen werden mittels einer Vielzahl von Plausibilitätskontrollen erfasst. Wie in

Abbildung 1 dargestellt werden Inkonsistenzberichte, und wenn verfügbar, Korrekturvorschläge generiert. Die erkannten Konflikte müssen durch medizinisches Personal in den Quelldatenbanken via KIS korrigiert werden und fließen dann automatisch in das Datamart-System ein. Dieses Systemkonzept erhöht also iterativ und persistent die Datenqualität aller integrierten Datensätze.

Regelbasierte Transformation

Für die nächsten Transformati-onsschritte werden Prüf- und Integrationsregeln definiert. Diese Regeln beschreiben im einfachsten Fall einen Kopierprozess oder eine univariate Transformation der gespiegelten Daten zur Zieldatenbank. Die im ersten Schritt gewonnene Bewertung der Quellattribute wird hier bei Vorliegen redundanter Informationen genutzt, um Zugriffsketten zu definieren.

Einerseits ist die Kombination von verschiedenen Daten logisch eindeutig (z.B. ist für alle Labor-

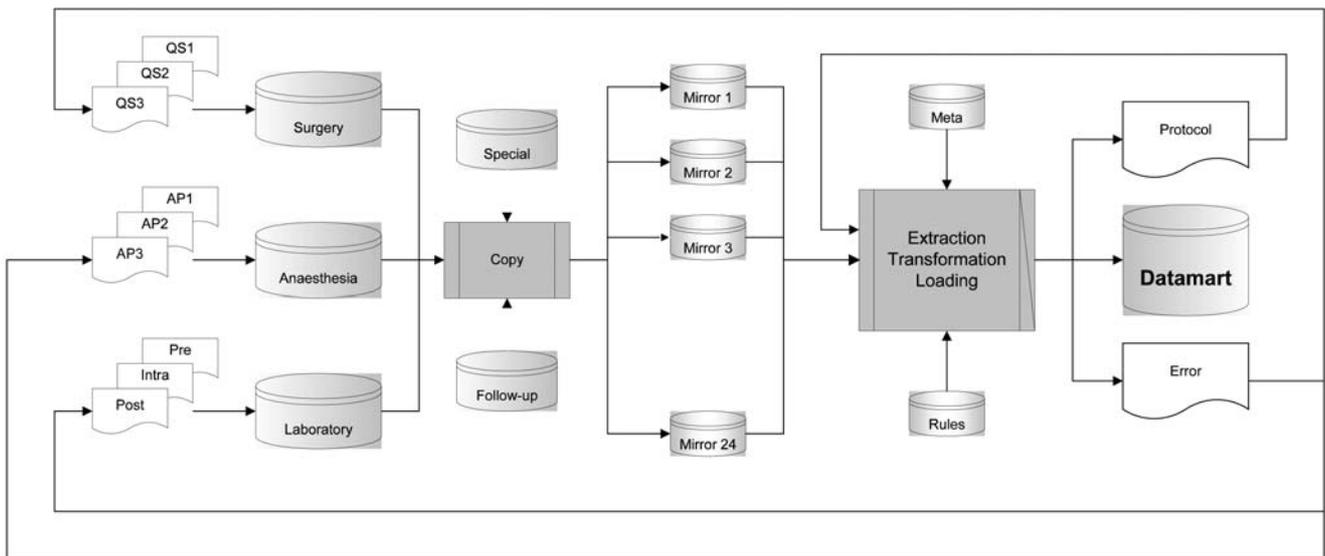
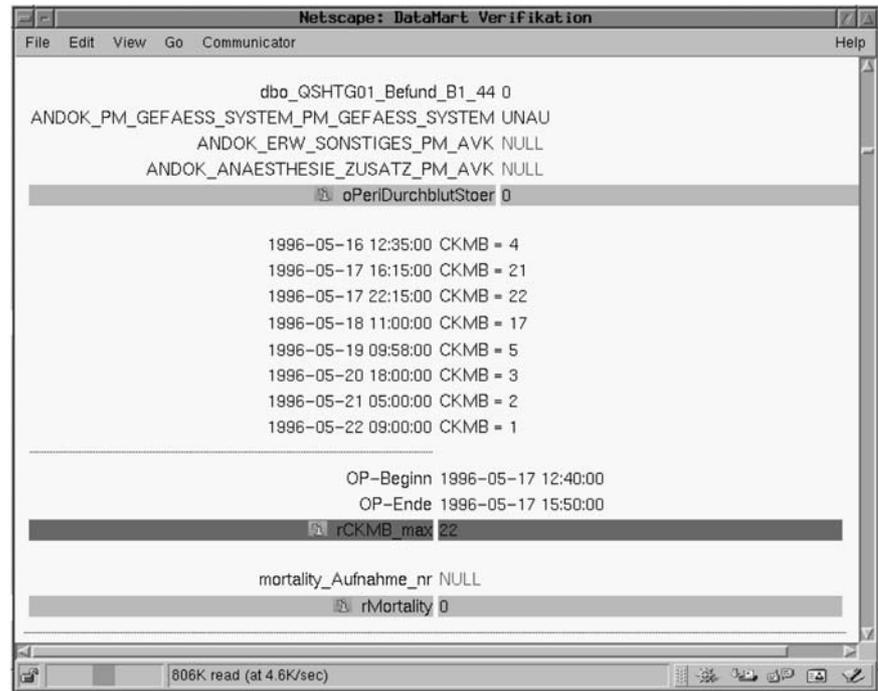


Abb. 1 Das Datamart-System spiegelt, extrahiert und konsolidiert alle relevanten Daten von den vorhandenen, unverbundenen KIS und den zusätzlichen Daten von verschiedenen Spezialstudien. Bislang werden 28 Tabellen mit insgesamt 389 Quellattribute verwendet. Inkonsistenzen in den Quell-

daten können ggf. mit automatisiert generierten Korrekturvorschlägen berichtet werden. Zur Zeit enthält die Datamart-Datenbank 277 prä-, intra- und postoperative Attribute

Abb. 2 Beispiel einer fall-basierten Verifikation von drei Datamart-Parametern. Oben: periphere Durchblutungsstörung (oPeriDurchblutStoer) ist aus der chirurgischen (erste Zeile) und der anaesthesiologischen (restliche drei Zeilen wegen KIS software update) Datenbank abgeleitet; die Werte „0“, „UNAU“ und die letzten beiden Marker für fehlende Werte „NULL“ resultieren in 0. Mitte: Bestimmung des maximalen Wertes der CKMB (rCKMB_max) nach Operationsende. Unten: Mortalitätsstatus ist 0, da der Fall nicht in der Mortalitätsliste vorhanden ist



werte der genaue Operationsbeginn für eine Einteilung in prä-, intra- und postoperative Werte nötig; siehe unten und auch mittleren Teil von Abbildung 2), in anderen Fällen ist zur Konstruktion der Zielgröße eine Übersetzung verschiedener Quellattribute erforderlich. Zum Beispiel wird der Parameter „Kritischer präoperativer Zustand“ aus insgesamt 15 Werten ermittelt (siehe auch Abbildung 2: der Datamartwert periphere Durchblutungsstörung ist aus chirurgischen und anesthesiologischen Daten abgeleitet).

Bedingt durch historische Veränderungen in den KIS-Datenbankstrukturen hängen Transformationsregeln z.T. vom Zeitpunkt der originalen Datenerfassung ab. Für die iterative Entwicklung dieser Regeln ist die enge Zusammenarbeit mit den Domänenexperten unerlässlich.

Tools zur Verifikation

Ein wichtiges Hilfsmittel ist ein effektives Werkzeug für die In-

spektion des kompletten Integrationsprozesses. Die Verwendung einer Vielzahl von Quellen mit historischen Veränderungen in den Datenbankstrukturen erfordern ein effektives Werkzeug für die Inspektion des kompletten Integrationsprozesses. Während des Zusammenbaus der Datamart Datenbank werden alle relevanten Quellwerte für jedes Attribut und für jeden Fall in einer Verifikationsdatenbank gespeichert. Sie dienen der schnellen iterativen Verifikation aller wesentlichen Aspekte des Extraktions- und Transformationsprozesses.

Abbildung 2 zeigt einige Integrationsdetails eines Falles. Das dargestellte web-basierte Tool erlaubt neben der fall-basierten Verifikation auch die Inspektion von Fallgruppen und Regeln.

Inspektion fehlender Werte

Fehlende Daten sind ein häufiges Problem bei der Realisierung von medizinischen, auf Patientendaten basierenden Studien. Die ad hoc-Methoden bei der Analyse unvoll-

ständiger Daten beschränken sich auf die Nichtbeachtung von Fällen mit fehlenden Werten oder auf die Ersetzung durch plausible Werte [7]. Korrelationen zwischen der Existenz fehlender Werte bei einem Attribut und der Wertausprägung eines anderen können signifikante Abweichungen in den Analyseresultaten bewirken.

Die Verwendung von verbreiteten Statistikprogrammen für die bi-variate Analyse fehlender Werte erzeugt schon bei einer geringen Anzahl von Attributen eine überwältigende Menge an Ausgaben. Eine kompakte Präsentation der beobachteten Beziehungen ist somit erforderlich. Wir entwickelten ein web-basiertes Werkzeug mit dem die Verteilung einer Menge von Attributen im Fall des Nichtvorhandenseins anderer Attribute inspiziert werden kann. Die Ergebnisse von allen hier verfügbaren statistischen Methoden (u.a. Chi-Quadrat, Student-t, Entropie-basierte Messungen, etc.) werden in konfigurierbarer und kompakter Form dargestellt.

Ergebnisse

Bislang sind Daten von über 13 000 Herzoperationen mit 277 prä-, intra- oder postoperativen Attributen pro Fall für verschiedene Auswertungen verfügbar.

■ Datamart-Informationsportal

Registrierte Chirurgen können das Datamart-System über ein Informationsportal im klinischen Intranet nutzen.

1. Teilmengen des konsolidierten Datensatzes können ausgewählt werden. Der Export von größeren Datenmengen erfordert die Autorisierung durch den Datamart-Administrator. Im Exportprozess wird automatisiert die Fallidentifikation durch eine sog. One-way-hash-Kodierung ersetzt und somit der Datensatz komplett anonymisiert. Damit entfallen die zeit- und arbeitsaufwendigen Schritte der Sammlung, Vorbereitung und Konsolidierung der Daten für retrospektive Studien in der medizinischen Forschung.

2. Für die Überwachung der klinischen Performanz können verschiedene Auswertungen online erzeugt und dargestellt werden (siehe auch Abbildung 3). Häufige Abfragen (z.B. bestimmte Zeitabschnitte, Operationstypen, Chirurgen, etc.) sind für autorisierte Benutzer direkt verfügbar.
3. Risikoadjustierte zeitliche Performanzdarstellungen in ausgewählten Untergruppen können dynamische generiert werden (z.B. VLADs).

Im Folgenden werden drei ausgewählte medizinische Studien welche besonders von den Möglichkeiten des Datamart-Systems profitierten vorgestellt.

■ Vermeidung eines Schlaganfalls

■ Medizinischer Hintergrund.

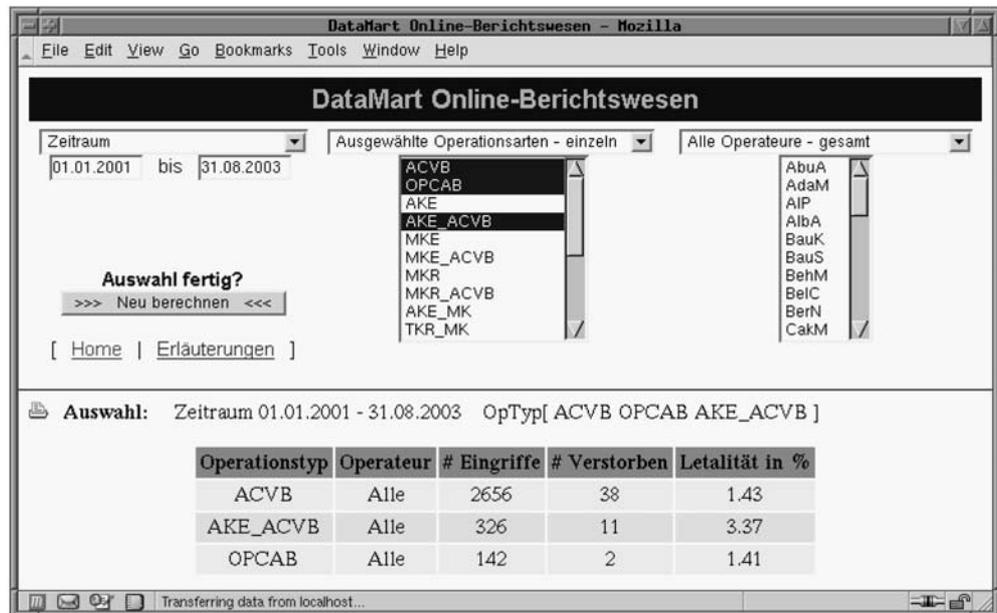
Schlaganfall ist die zweithäufigste Ursache von Mortalität und Morbidität in der westlichen Welt. Während und nach einer Herzoperation ist das Risiko eines

Schlaganfalls bekanntlich durch die Manipulation der Herz und Gehirn versorgenden Arterien, die Verwendung der Herz-Lungen-Maschine und die beeinträchtigte Hirndurchblutung erhöht. Zu den seit längerem bekannten Schlaganfall-Risikofaktoren zählen erhöhtes Alter, Arteriosklerose der Hirnarterien, Diabetes, etc. In letzter Zeit werden verstärkt hämodynamische Faktoren untersucht, die möglicherweise das Auftreten eines Schlaganfalls beeinflussen.

■ **Klinische Situation.** Schlaganfall in der Herzchirurgie ist nach wie vor eine verheerende Komplikation.

■ **Datamart-Gewinn.** Die Schlaganfallhäufigkeit im Herzzentrum Lahr ist glücklicherweise nur 1,5% gemittelt über alle Operationen. Für den Nachweis signifikanter Beziehungen wird bei solch seltenen Ereignissen eine genügend große Fallanzahl benötigt. Durch die Integration aller relevanten historischen Daten und aller Laborwerte eines Patienten

Abb. 3 Online-Überwachung der klinischen Performanz durch autorisierte Benutzer



während des gesamten Klinikaufenthaltes in das Datamart-System konnten wir Blutzellveränderungen vor dem Auftreten eines Schlaganfalles untersuchen. Dies ist besonders wertvoll, da Blutparameter anderweitig nur sehr selten direkt vor einem Schlaganfallereignis verfügbar sind. Eines unserer Ziele ist, zum allgemeinen Verständnis über die Entstehung von Schlaganfällen beitragen zu können. In [4] gelang uns zum ersten Mal der Nachweis einer signifikanten Korrelation zwischen hohen präoperativen Leukozytenwerten und dem Auftreten eines Schlaganfalles während oder nach einer Herzoperation.

■ Operative Qualitätsmessung

■ **Medizinischer Hintergrund.** Die Bewertung der Qualität herzchirurgischer Operationen durch den Vergleich der Mortalitätsraten zwischen Operateuren und Kliniken ist von steigender Bedeutung. Für einen fairen Vergleich müssen die unterschiedlichen Patientenprofile in den statistischen Analysen berücksichtigt werden (siehe auch Albert et al.).

■ **Klinische Situation.** Es besteht der Bedarf an einem auf objektiven Parametern basierenden Risikomodell mit dem eine genaue präoperative Schätzung der Mortalität und Morbidität möglich ist.

■ **Datamart-Gewinn.** Ein bewährtes Risikomodell ist das „European System for Cardiac Operative Risk Evaluation“ (EuroSCORE) [5]. Für die Schätzung des individuellen Mortalitätsrisikos werden 17 verschiedene Parameter verwendet. Die Datamart-Datenbank ermöglichte uns die retrospektive Berechnung des EuroSCORE's im

Herzzentrum Lahr in 75% aller Fälle obgleich die Mehrzahl der erforderlichen Parameter nie nach den genauen Definitionen in [5] erfasst wurden. Damit können wir risiko-adjustierte Vergleiche zwischen Kliniken und Operateuren durchführen. Weiterhin haben wir den Einfluss des Lebensalters als Determinante der Mortalität in der Herzchirurgie in unserer Patientenpopulation untersucht [6].

■ Nierenschwäche und Creatinine Clearance (CC)

■ **Medizinischer Hintergrund.** Die Filterkapazität der Niere ist ein überlebenswichtiger Faktor – natürlich auch im Umfeld einer Herzoperation.

■ **Klinische Situation.** Im EuroSCORE-Risikomodell wird nur der binäre *serum creatinine* Wert zur Bewertung der Nierenfunktion verwendet.

■ **Datamart-Gewinn.** Der abgeleitete Parameter CC kann mittels *serum creatinine*, Geschlecht, Lebensalter und Körpergewicht geschätzt werden. Wir konnten zeigen, dass die Ersetzung des *serum creatinine* Wertes durch die CC-Schätzung im EuroSCORE-Risikomodell vorteilhaft ist. Eine variable Ranganalyse identifizierte CC als den besten einzelnen Prediktor [8].

Diskussion

Die vorgestellte Datamart Architektur hat sich als wertvoll und effektiv erwiesen. Die aktuellen und historischen Daten von allen relevanten Datenquellen werden

integriert ohne die Haftpflichtregelungen und Wartungsverträge der vorhandenen KIS zu berühren. Damit können die anfänglichen Vorbehalte der Abteilungsleiter minimiert und das Projektziel effektiv erreicht werden.

Dieser Ansatz ermöglicht die gezielte Nutzung der Redundanzen in den aufgenommenen Daten. Einerseits werden Inkonsistenzen innerhalb der Datenbanken detektiert, andererseits können Attribute die nicht direkt erfasst wurden aus verschiedenen Quellen abgeleitet werden. Eine erhöhte Anzahl von teilweise redundanten Attributen vergrößern dabei den Spielraum bei der Formulierung von Transformationsregeln. Werkzeuge zur Verifikation ermöglichen die Anpassung und Verfeinerung der Regeln.

In der Vergangenheit war die Durchführung von umfassenden retrospektiven Studien im Herzzentrum Lahr durch den hohen Arbeitsaufwand beschränkt. Einzelne Versuche bestimmte Teilmengen der Daten aus den verschiedenen KIS zu kombinieren wurden bereits unternommen. Die dazu nötigen Schritte waren abhängig von der wissenschaftlichen Fragestellung. Folglich unterschied sich die Semantik der verbundenen Daten in Abhängigkeit der jeweiligen Studie. Zusätzlich war es durch die historischen Veränderungen der Datenbankstrukturen schwierig, den Überblick über die verschiedenen Datensätze zu behalten. Mit der Implementierung des vorgestellten Datamart-Systems konnten die zeit- und arbeitsaufwendigen Prozesse der Datensammlung und Konsolidierung ersetzt werden und eine stabile Forschungsdatenbank die reproduzierbare Ergebnisse ermöglicht, realisiert werden.

Literatur

1. Beuscart-Zephir MC, Brender J, Beuscart R, Menager-Depriester I (1997) Cognitive evaluation: How to assess the usability of information technology in healthcare. *Comput Methods Programs Biomed* 54(1/2):19–28
2. Tsumoto S (2000) Clinical knowledge discovery in hospital information systems: two case studies. *Principles of data mining and knowledge discovery*, pp 652–656
3. Rayner M, Petersen S (2000) European cardiovascular disease statistics 2000 edition. British Heart Foundation
4. Albert AA, Beller CJ, Walter JA, Arnrich B, Rosendahl UP, Priss H, Ennker J (2003) Preoperative high leukocyte count: a novel risk factor for stroke after cardiac surgery. *Ann Thorac Surg* 75(5):1550–1557
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 (1999) Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. *Eur J Cardiothorac Surg* 15(6):816–822
6. Mortasawi A, Arnrich B, Rosendahl U, Frerichs I, Albert A, Walter J, Ennker J (2002) Is age an independent determinant of mortality in cardiac surgery as suggested by the EuroSCORE? *BMC Surg* 2(1):8
7. Schafer J, Olsen M (1997) Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioural Research* 33:545–571
8. Walter J, Mortasawi A, Arnrich B, Albert A, Frerichs I, Rosendahl U, Ennker J (2003) *BMC Surg* 3(1):4