Die approbierte Originalversion dieser Dissertation ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (http://www.ub.tuwien.ac.at).

The approved original version of this thesis is available at the main library of the Vienna University of Technology (http://www.ub.tuwien.ac.at/englweb/).

TU WIEN

Ph.D. Thesis

Towards a Sustainable Data Warehouse Approach for Evidence-Based Healthcare

Conducted for the purpose of receiving the academic title
´Doktorin der technischen Wissenschaften´

Advisors

Prof. A Min Tjoa
Institute for Software Technology and Interactive Systems / E188
Faculty of Informatics
Vienna University of Technology

Prof. Thomas Mück
Institute of Information Systems / E184
Faculty of Informatics
Vienna University of Technology

Submitted to the
Vienna University of Technology
Faculty of Informatics

by

Nevena Stolba
9326716
Auhofstrasse 162/16
1130 Wien

Vienna, September 27th 2007 _____

Acknowledgments

It is my pleasure to thank the many people who made this thesis possible.
I would like to thank my thesis advisor Prof. A Min Tjoa for his encouragement, guidance, and for giving me precious professional and moral support during last four years.
I am grateful to Prof. Thomas Mück for his valuable suggestions and for giving me insight into the healthcare industry life.
Acknowledgement is also due Prof. Gerti Kappel for giving me the opportunity to achieve my goals in her team and all my colleagues at WIT who accompanied me throughout my thesis work.
I am very grateful to Prof. Silvia Miksch for mentoring, motivation and moral support she gave me.
Most of all, I want to thank my family for giving me the emotional support and for their endless patience and understanding.

Nevena Stolba

Kurzfassung

Die Gesundheitsfürsorge-Industrie ist eine der größten, sich am schnellsten entwickelnden und informationsintensivsten Industrien der Welt [KFF, 2006]. Im Gegensatz zu den anderen Industrien, wo Data Warehouse (DWH) Technologie in vielen Bereichen erfolgreich angewandt worden ist, hat diese Technologie im Gesundheitswesen bisher vor allem die administrativen und logistischen Aspekte der Informationsverarbeitung angesprochen. Das wachsende Bedürfnis nach der einheitlichen Gesundheitsfürsorge trägt vermehrt dazu bei, die umfassenden klinischen Entscheidungsunterstützungssysteme (*decision support system - DSS*) einzuführen.

Das schnelle Wachstum von Informationstechnologie hat immense Möglichkeiten für die Entwicklung und Verbreitung von *evidence-based guidelines (EBG)* gebracht. Diese, auf Beweisen gegründete medizinische Erkenntnisse, werden durch die statistischen Verfahren und Analyse der verteilten, in heterogenen Datenquellen gelagerten, Patientendaten gewonnen.

Evidence-based medicine (EBM) bietet eine Sammlung der bestbewehrten Praxis-Richtlinien an, um Medikamente und Behandlungen zu empfehlen. Diese Dissertation stellt fest, dass DWH Technologie das Praktizieren der EBM zweifach fördern kann: (1) durch das Unterstützen des EBG - Entwicklungsprozesses und (2) durch zur Verfügung Stellen der Richtlinien für die optimale Entscheidungsfindung der behandelnden Ärzte.

In Bezug auf (1) zeigen wir, dass DWH und Data Mining die Entwicklung der EBG ermöglichen, indem sie eine Plattform und die Werkzeuge für die Wissensfindung und Mustererkennung bereitstellen. Große Datenmengen können analysiert werden, um bekannte Tendenzen in Daten zu bestätigen oder unbekannte Korrelationen zu entdecken.

Bezüglich (2) behaupten wir, dass der Behandlungsprozess von der Anwendung der DWH Technologie wesentlich profitieren kann. In Anbetracht einer einheitlichen Wissensbasis, in die eine breite Vielfalt von patientenbezogenen Datenquellen integriert ist und in die EBG eingebettet sind, bietet unser Ansatz eine statistisch aufbereitete und bedarfsorientierte Entscheidungshilfe für die behandelnden Ärzte in ihrer täglichen Arbeit.

Um die Vertraulichkeit der Patientendaten im heutigen informationsbasierten und verteilten Gesundheitsfürsorgesystem zu sichern, empfiehlt diese Dissertation den föderierten anstatt des zentralisierten DWH Ansatzes. Wir befürworten die Anwendung von Depersonalisierung, Pseudonymisierung und rollen-basierten Zugriffsmechanismen für den Schutz der sensiblen medizinischen Daten.

Diese Doktorarbeit postuliert die Regeln, die zu befolgen sind, wenn ein umfassendes, föderiertes, EBM-basiertes Decision Support System entwickelt wird. Das Einhalten dieser Regeln führt zu einer nachhaltigen Entscheidungshilfe, die die Informationsbedürfnisse des medizinischen Personals, der Behörden und der Patienten zufrieden stellen kann.

Abstract

The healthcare industry is one of the world's largest, fastest-developing and most information-rich industries [KFF, 2006]. Rapid growth of information technologies has brought immense opportunities for patient data sharing, development and dissemination of evidence-based medical knowledge and analysis across distributed, heterogeneous healthcare data sources.

In contrast to other industries, where data warehouses have been successfully applied, healthcare is an area in which the information technology had only been able to permeate the administrative and logistic aspects of information processing. The growing need for integrated healthcare has led this industry to open towards adoption of extensive clinical decision support systems.

Evidence-based medicine (EBM) offers a collection of proven best practice guidelines for recommending drugs and medical treatments. This thesis states that the way data warehouse (DWH) technology can facilitate EBM is twofold: (1) by supporting the rule development process, and (2) by providing the EBM-enriched knowledge base to support the decision-making process of the care givers.

With respect to (1), we explain that data warehousing and data mining support the creation of the evidence-based rules by providing a platform and tools for knowledge discovery. Large amounts of data can be analysed to confirm known or discover unknown trends and correlations in data.

Regarding (2), we argue that the care giving process can benefit significantly from the application of DWH technology at the point of care. Given an integrated knowledge base, built upon a broad variety of patient-related information sources and incorporating evidence-based rules, our approach offers a unique decision support for the practitioners in their every-day work.

In order to guarantee the confidentiality of patient data in today's increasingly information-based, multi-site health delivery environment, this thesis recommends a federated DWH approach instead of collecting data from remote sources into a centralized system. We endorse the application of de-personalisation, pseudonymization and role-based access mechanism for protection of sensitive healthcare data.

This dissertation is intended to provide a roadmap for achieving sustainable healthcare decision support system based on federated data warehouses, facilitating evidence-based medicine that safeguards patient's personal privacy. It postulates four rules to follow when building a modern medical decision support system and its advisory nature will prove to be helpful in designing future healthcare project[*].

[*] This research has been partly funded by the Austrian Federal Ministry for Education, Science and Culture, and the European Social Fund (ESF) under grant 31.963/46-VII/9/2002.

Contents

List of Figures

List of Tables

1 Introduction

In the past decade, data warehouse (DWH) technology has been successfully applied in telecommunications [Walters, 2003], [SYBASE, 2007b], retail [Teradata, 2004], finance services [Britt, 2005], travel [Ferrell, 2007] and many other industries. Companies have been deploying DWH-based decision support systems in order to not only support business management high-level decision making, but also to support a wide range of applications throughout the enterprise.

The telecommunication sector is a good example of the importance of DWH for strategic business management in a highly competitive market. As mobile telephony market was growing very rapidly, business decisions had to be made equally fast. The accuracy of decisions was highly dependent on the availability and quality of operational data they were based on. DWH has been recognized as the ideal platform for integration of heterogeneous data sources into a company-wide decision-support system. Since DWH provided the competitive advantage in the marketplace, new emerging Telco companies planned and built their DWH systems even before they had their first registered subscriber. This way, they could rely on a mature decision support system to be used from the start of their business. In the past few years, DWH proved to be successful in all domains of Telco-business, most of all in marketing, revenue assurance, fraud detection, billing, churn prevention and customer care [Teradata, 2003], [Donovan, 2005], [SYBASE, 2007a].

Although DWH effectively broke into most markets, there have been almost no attempts to establish it in the healthcare domain. There are few reasons for this:

- In most European countries, healthcare is a public service. Since the government is assuring at least the basic healthcare for the citizens, the health insurance market is not as competitive as in most other industries. The absence of competitiveness resulted in no need for DWH deployment.
- Until recently, most of the healthcare data has been stored in unstructured or even hand-written notation. Integrating such data sources into a single data repository and preparing the data for analytical purposes is a major research challenge.
- Healthcare data are extremely complex and voluminous. Building a DWH would therefore be extensive and costly undertaking.

These reasons for not implementing DWH solutions in healthcare environment are going to lose their relevance in the coming years.

1.1 The Need for Data Warehousing in the Healthcare Environment

The treatment and welfare costs are expected to increase significantly as a result of growing and aging population. Increasing patient expectations and rising budgetary pressure are forcing governmental health authorities to rapidly adopt research-based innovations into the healthcare system. These innovations are required for a major shift in efficiency and quality of communication between individual health system participants.

The patient health records are usually distributed over different locations at care giver sites. Most healthcare providers have developed their local information systems independently (using different, mutually incompatible data formats) and are exchanging their information only sporadically. As a result, each care provider possesses only a partial knowledge of the patient's health history. [dbMotion, 2005] mentions the report published by the Institute of Medicine which states that in 1999, between 44.000 and 98.000 annual patient deaths in the United States were caused by preventable human medical errors. These errors also caused total costs of between $17 billion and $29 billion annually in hospitals alone. The report states one reason to be the decentralized nature of healthcare provision, preventing any single caregiver from receiving the complete patient information.

At this time, we are witnessing a paradigm shift in the healthcare information systems development and implementation - from individual, local decision support systems to collaborative, shared-knowledge based approach. Building a global patient healthcare record by gathering all available data stored at remote locations is an inevitable, but challenging task. Protecting the confidentiality of highly sensitive patient data is the essential pre-requisite for successful sharing of data among healthcare institutions. Although it is obvious that the comprehensive IT support is needed for information acquisition, gathering and consolidation, there is still no consensus on the approach.

The integration of evidence-based medicine rules into clinical decision-support systems would both improve quality and reduce costs of care, by recommending only the most efficient treatments and medications. Practicing evidence-based medicine often relies on published clinical evidence in form of books, magazines, journals, healthcare protocols, clinical trials and best practice guidelines [Abidi S.S.R., 2001]. In order to obtain the best evidence for a given disease, external clinical expertise, as well as internal clinical experience, must be available to the healthcare practitioners at the right time and in the right manner. Consolidation of heterogeneous medical information sources and building of a huge medical data repository are the prerequisites for practice of evidence-based medicine.

In summary, there are two major requirements for the information technology in order to improve the quality of care:

- to enable the inter-operability across heterogeneous medical information systems,
- to integrate evidence-based knowledge into medical decision making process.

1.2 Contribution

The consolidation of distributed patient data and building of a sustainable e-health system is gaining more attention as the need for collaboration in the healthcare domain increases. Standardized data structures and corresponding underlying infrastructure are needed in order to overcome the institutional barriers for change across sectors. We argue that DWH, already proven to be beneficial in other industries, represents the best suitable platform for consolidation of diverse health information systems and can therefore sustain the pressure of governmental authorities.

DWH offers a comprehensive support for gathering, analyzing and presenting medical data. The system complexity, the heterogeneity of healthcare data sources, massive volumes of medical data and a high number of concurrent users are the main reasons for use of a *federated DWH integration model*. The essential concern of the proposed approach is the data integration of the heterogeneous, underlying data warehouses, databases, and diverse legacy systems with the external EBM sources.

This thesis shows how federated DWH facilitating EBM can be used for building sustainable healthcare systems which contribute to more effective healthcare. The benefits of such system would be as follows:

- **QUALITY OF CARE:** EBM relies entirely on proven medical experience. Incorporating EBM into the clinical decision-making process ensures that only impartial and scientifically verified knowledge will be offered to physicians.
- **INTEROPERABILITY:** In order to give the best possible care to the patient, physicians need insight into the patient's complete health record. During their lifetime, patients receive healthcare from diverse medical institutions that keep the records of the individual treatments for a mandatory period of time. Utilizing wrapper and mediator technology and semantic integration, federated DWH acquires all available patient data from diverse medical data sources, transforms it, and creates a unique patient's health record which is then made available to the decision makers in a user-friendly manner.
- **DECENTRALIZED DATA STORAGE:** An electronic health record (EHR) is not necessarily stored as a single physical entity in a centralized system. Instead, it can be aggregated into a single coherent record from data stored at various geographical locations, when required.
- **SCALABILITY:** A federated DWH utilizes a component-based architecture. Each new data source can be easily included into the federation without redesign of the existing system.
- **ADOPTION OF INTERNATIONAL STANDARDS:** Our approach recommends the application of internationally adopted standards in order to enable seamless transmission and understanding of healthcare data among health providers.
- **SECURITY:** Due to high confidentiality of healthcare data, guaranteeing data privacy is a fundamental requirement for the successful data integration. Any disclosure of health data, especially when related to a particular person, could be irreparably harmful, and their

protection is even legally regulated. In our federated DHW approach, the data from participating healthcare organizations are only virtually tied in a network and not physically duplicated into a central data store. Therefore, the introduced solution addresses possible security threats to the highly sensitive medical data. This approach is a significant step towards decentralization of security assurance.

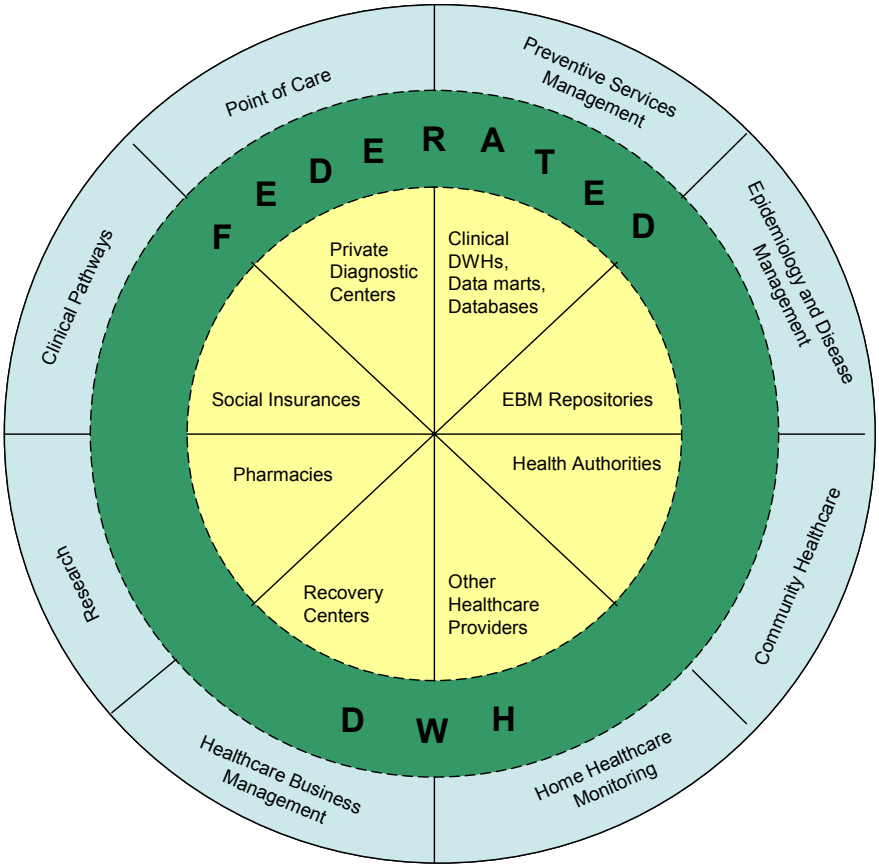


Figure 1: Federated DHW Conceptual Model

Figure 1 illustrates the conceptual model of our approach. The main healthcare players that might participate in the federated DWH are shown in the center of the picture. The outer frame represents the most important applications fields.

We will show that the use of a federated DWH, which incorporates evidence-based knowledge, is a sound basis for successful decision support. Significant improvement of the healthcare quality could therefore be achieved by reducing medical errors, increasing operating efficiency, giving advice about staffing and resource plans, supporting clinical management to negotiate care contracts based on accurate data on resource utilization, supporting development and controlling of clinical pathways and reducing treatment costs (i.e. by avoiding unnecessary and duplicate examinations).

1.3 Outline

This thesis is organized as follows:

- **Part I** presents the fundamental technologies applied in the thesis and gives an overview of the related work.
- **Part II** represents the kernel of the thesis and gives a detailed description of our approach in form of rules.
- In **Part III**, two application scenarios are used to demonstrate how the proposed approach could be used in practice.

A significant portion of the work summarized in this thesis has been published and presented at international conferences. Here we describe the structure of the thesis and we reference our own work according to the content of the publications:

PART I

Chapter 1 describes the necessity of data warehousing in the healthcare environment. It gives an overview of the present status of IT in this area and states the requirements for improving the quality of care.

Chapter 2 gives an overview of data warehouse technology. It does not cover the basic features of the DWH, as it assumes that the reader is familiar with them, but it focuses on the DWH methodology and evolution instead. In this chapter, we commence our running example that will be used throughout the thesis.

Chapter 3 introduces evidence-based medicine. It describes the concepts and references the most important sources of evidence-based knowledge. It then explains how clinical decision support systems can be enriched by integrating EBM. The description of three major application fields rounds up this chapter.
Parts of the material from this chapter were published in [Stolba and Tjoa, 2006b].

Chapter 4 presents relevant work in the field and contrasts it to the approach proposed by this thesis. It describes the opportunities for federated DWH technology in the healthcare area. Furthermore, four basic rules for the creation of a sustainable federated DWH supporting evidence-based healthcare are proposed.
Part of the issues from this chapter were published in [Stolba and Schanner, 2007].

PART II

Chapter 5 covers the first rule and explains the federated DWH technology and its place in the healthcare environment. The application of this technology is illustrated using our running example.
Parts of the material from this chapter were included in the publication [Stolba et al., 2006c].

Chapter 6 explains the second rule and introduces the most common standards used in the healthcare arena to enable seamless communication between diverse information systems.
Part of the issues from this chapter were published in [Stolba et al., 2006c].

Chapter 7 introduces the third rule. It describes how wrapper-mediator architecture is applied to support integration of medical data coming from heterogeneous sources. Our running example illustrates this issue through a scenario at the point of care.
Parts of this chapter were published in [Stolba et al., 2007a]

Chapter 8 presents the fourth rule, security assurance. This chapter introduces major security threats, concepts and measures and reviews the most relevant international privacy regulations for healthcare. Then, it focuses on preserving privacy in the federated environment through depersonalization, pseudonymization and role-based access. Once again, this is demonstrated through our running example.
This chapter builds on the publications:[Stolba and Tjoa, 2006a] and [Stolba et al., 2006c].

PART III

Chapter 9 provides an application scenario for the federated DWH as a model for supporting interoperability of national healthcare IS in Austria. It describes MAGDA-LENA, the Austrian framework for electronic exchange of patient data, and describes how federated DWH can be built according to the MAGDA-LENA recommendations.
This chapter builds on the publication [Stolba et al., 2007b].

Chapter 10 delivers another application scenario for the proposed model, the support of real-time sense and response for effective healthcare. This scenario illustrates the

application of a federated DWH facilitating EBM and incorporates the sense and response mechanism in emergencies, for support of emergency room decision making. *This chapter is based on the publication [Stolba et al., 2007c]*.

Chapter 11 concludes by describing lessons learned from the healthcare projects investigated in the course of this thesis and gives an overview of the open issues and future work.

Part I

Technological Foundation

2 Technical Background of Data Warehousing

A data warehouse is a central repository containing data from variety of source systems and needed for decision-making across the whole organization. Integrating these heterogeneous data sources into a consistent framework is a major challenge. In this chapter we describe the basic features of the data warehouse technology and we introduce the running example that will be used throughout the thesis.

2.1 Concepts of Data Warehousing

The term data warehouse was coined by W. H. Inmon: "A warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process" [Inmon, 1992].

In contrast to traditional On-Line Transaction Processing (OLTP) applications, decision support places some different requirements on database technology. Data warehouses provide storage, functionality, and responsiveness to queries that are beyond the capacity of OLTP databases. They contain historical, summarized, and consolidated data over potentially long periods of time. Their size can be from hundreds of gigabytes to terabytes. There is a great need to provide decision-makers at all levels of management with information at the desired level of detail, to support their decision-making. Apart from performing regular, predefined reporting activities, a number of parallel users are submitting ad hoc, complex queries. These queries require access to millions of records and cause numerous scan, join, and aggregate operations across the warehouse. As a result, query throughput and response times are main issues in multi-user decision-support systems.

Figure 2 shows the DWH architecture, from collecting source data to data delivery to the decision-makers.

The source data is usually stored in different source-system-specific formats. During the *extraction phase,* data is collected from operational systems or external sources. The external data is often stored into spreadsheets, personal databases, web logs etc. It can be accessed directly or indirectly (in case of proprietary systems).

In the *transformation phase*, selected data is cleaned and converted into the format and structure compatible with the warehouse. Syntactic and semantic distinctions between operational sources are adjusted and local logical models are mapped and integrated into the global DWH data model. The mapping characteristics are captured and stored in the DWH as metadata.

In the *data storage phase*, new data is loaded into the warehouse and merged with the existing stored data. During this phase, data is usually restructured for optimized querying. Data loads need to be run on a regular basis in order to keep warehouse data accurate.

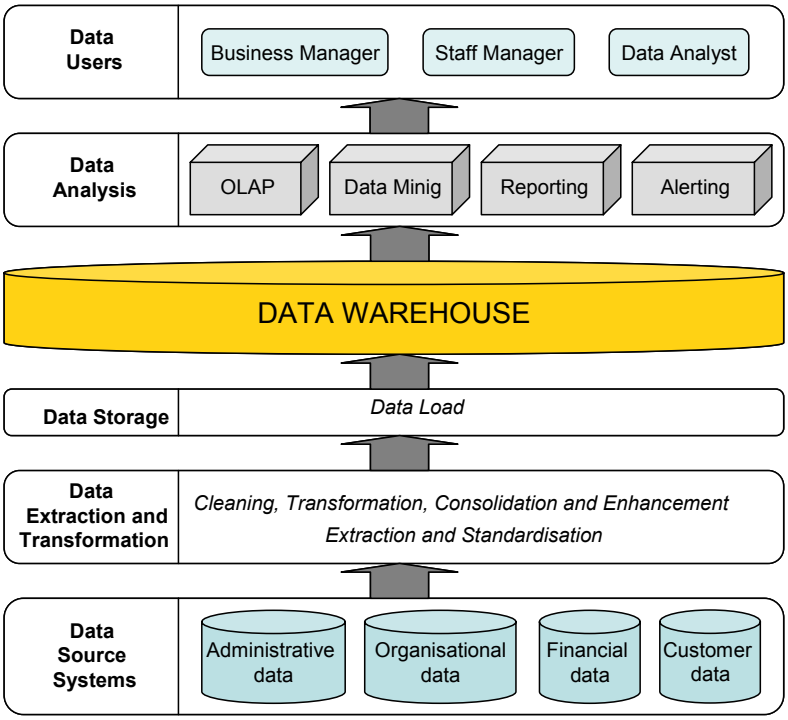


Figure 2: DWH Architecture

After being integrated into the DWH, data is ready for querying and analysing by OLAP and data mining tools. Decision-makers and staff managers are supported by pre-defined reports or can retrieve desired information in an ad-hoc manner.

2.2 Data Warehouse Design Methodology

Designing and rolling out a data warehouse is a complex process, consisting of the following activities [Kimball, 1996]:

1. Define the architecture, do capacity planning, and select the storage servers and OLAP servers and tools.
2. Integrate the servers, storage, and client tools.
3. Design the warehouse schema and views.
4. Define the physical warehouse organization, data placement, partitioning, and access methods.

5. Connect the sources using gateways, ODBC drivers and other wrappers.
6. Design and implement scripts for data extraction, cleaning, transformation, load, and refresh.
7. Populate the repository with the schema and view definitions, scripts, and other metadata.
8. Design and implement end-user applications.
9. Roll out the warehouse and applications.

Analogically, Elsmari and Navathe [Elsmari and Navathe, 2004] describe the main phases of DWH design and implementation as illustrated in Figure 3.

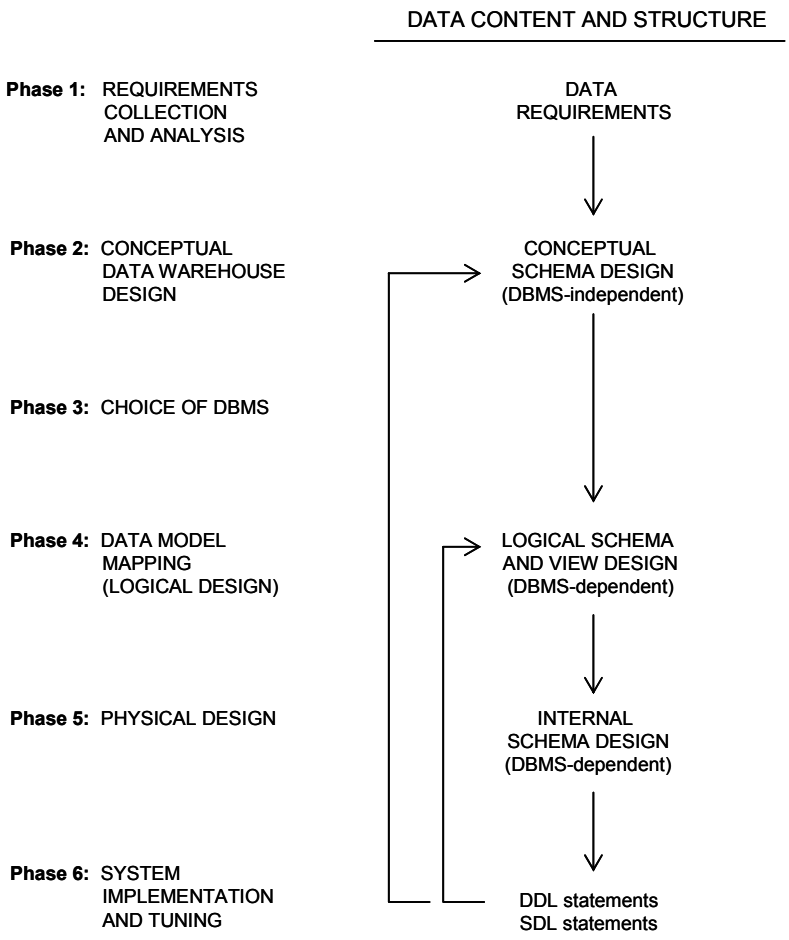


Figure 3: Phases of DWH Design and Implementation, adopted from [Elsmari and Navathe, 2004]

Phase 1 involves collecting information about the intended use of the DWH. Phase 2 (conceptual DWH design) produces a conceptual schema for the DWH that is independent of a specific DBMS. High level data models, such as Entity Relationship (ER), enhanced ER (EER), and ADAPT [Symmetry_Corporation, 2007] model are often used during this phase. In Phase 4 (which is also called logical DWH design), the conceptual schema from the high-level data model used in Phase 2 is mapped into the data model of the chosen DBMS. During this phase, mappings between the logical data models (LDM) of the source databases and the target DWH LDM are created. In Phase 5, physical storage structures, record placement, and indexes are specified. During Phase 6, the DWH

and application programs are implemented and tested. This typically reveals opportunities for "data tuning" - physical design changes, data indexing, re-organisation, and different placement of data. Tuning is an ongoing activity – a part of system maintenance that continues for the life cycle of the DWH as long as the DWH keeps evolving and performance problems are detected.

Figure 4 represents an even more detailed DWH Methodology road-map, given by the Teradata. Teradata Solutions Methodology is comprised of eight phases. For each service, a set of tasks shows in detail the work required, including necessary inputs and outputs, for the successful task completion.

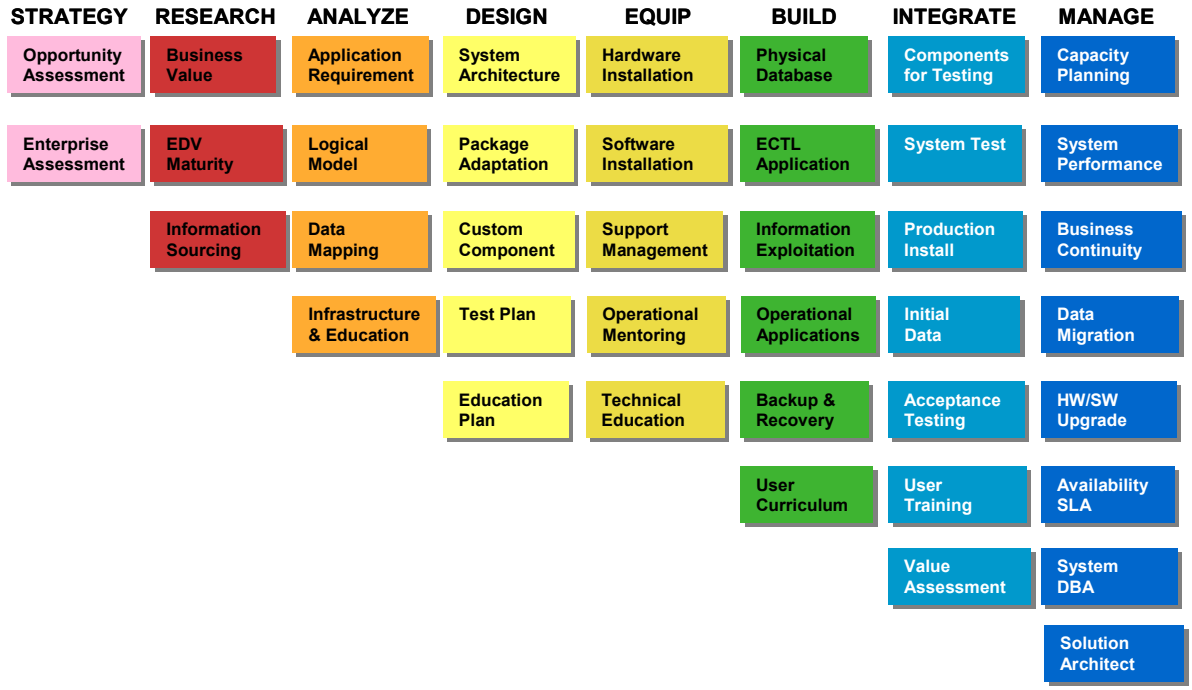


Figure 4: Teradata Solutions Methodology 5.0 [Teradata, 2006]

The first three phases: *Strategy*, *Research*, and *Analyze,* involve activities related to strategic planning and definition of specific DWH opportunities. After business goals, processes, available data and organizational support have been analysed, business functions which need to be addressed by the DWH have to be prioritised. The result is a roadmap of business improvement opportunities and a high-level data model of the defined project scope. It demonstrates the flow of data from operational systems and the impact of privacy and security requirements.

The next three phases: *Design*, *Equip*, and *Build* comprise the physical implementation of the system, including hardware and software installation, ECTL (Extract, Correction, Transformation and Load) development, and application development.

Integrate is the phase which encompasses data loading, testing, user training, and demonstration of value based on quantified metrics.

The last phase, *Manage*, is designed to provide seamless expansion of new applications, accommodate a growing number of users, and deliver continuing education, capacity and contingency planning as well as 24/7 support [Teradata, 2006].

2.3 Main Stages of Data Warehouse Evolution

In the early nineties, the main concern of data warehousing was to support the knowledge workers in their strategic decision-making. Data warehousing technology evolved significantly since these early beginnings. According to [Brobst and Rarey, 2001] the evolution of data warehousing can be described through the following five stages: *reporting, analyzing, predicting, operationalizing,* and *activating*. Table 1 gives a short overview of these evolution stages, with a special focus on deployment in the healthcare environment.

#	Stage	Description	In Healthcare
1	Reporting	DWH integrates large amounts of heterogeneous data sources within an organization. It is primarily used for reporting where the business questions are mostly known in advance. For this reason, the performance of DWH queries can be optimized even when the queries require huge amounts of information.	Clinical decision support systems are created for hospitals. Data about the patients, diagnoses, drugs, treatments, staff, room occupation etc. are united into a single repository. Clinical management and clinical administration is using pre-defined reports to make strategic decisions.
2	Analyzing	The focus of decision makers at this stage is less on *what happened* and more on *why it happened*. DWH tools are used to drill down to the detailed data level and to slice and dice data for in-depth analysis. A DWH at this stage supports ad-hoc queries. Since the questions to a DWH can not be known in advance, performance is much more important here. Database optimizers play a crucial role in allowing flexible access to information within acceptable response times. The challenge of this stage is to support the concurrent query execution and large numbers of users.	High level business users (clinical business management) are accessing a DWH to analyse detailed data. Some examples of strategic queries supported at this stage are: what causes delays in patient admission phase, why does the treatment of some diseases creates above-than average costs, etc.
3	Predicting	The main request of a DWH at this stage is to predict what will happen next in the business. Data mining tools are required for	Development of evidence-based rules relies heavily on data mining techniques. Knowledge workers are

		building predictive models using historical data. Although the number of end-users is much smaller than in prior stages, workloads associated with model construction and scoring are heavy. Complex mathematical functions and access to detailed data are pre-requisites for the application of advanced data mining methods.	applying these methods to discover patterns in huge amounts of available historical data. Apart from its application in the area of evidence-based medicine, data mining is a valuable tool for prediction in the field of epidemiology and disease management.
4	Operationalizing	Unlike stages 1-3, which are dealing with strategic decision-making, this stage focuses on tactical decision support. Instead of providing information necessary to make long-term decisions, tactical decision support is utilized by people in the field, who execute the decisions. This discrepancy implies the difference in data freshness as well: while strategic decision support relies on data that is loaded once per month or once per week, tactical decision support requires up-to-date information. Furthermore, only short query response time is acceptable in order to accommodate the realities of decision-making in an operational field environment.	A DWH at this stage of evolution is used to support immediate decision-making at the point of care, in an emergency room etc. Since the physician is accessing the DWH while examining the patient, his (her) queries must be answered in a short time and return accurate patient data.
5	Activating	The most advanced stage of a DWH nowadays reflects the business need to automate decisions when humans do not add significant value. The Active Warehousing is based upon event-based triggering. Its goal is to allow decision-making to take place as near real-time as possible and so to increase the speed and accuracy of business decisions. The query response time in this stage is measured in seconds or milliseconds. Active DWH supports the coexistence of both strategic and tactical decision support.	An example of deployment of an active DWH in the healthcare area would be for automatic data integration in the field of EBM: Every time that new, verified research results about the treatment of a particular disease are available, a DWH would be triggered and new knowledge would be inserted into an evidence-based guidelines repository. Another example is in intensive care, where each change of the patient's health condition triggers a decision making process in the DWH.

Table 1: Stages of DWH Evolution Applied to Healthcare Environment, based on [Brobst and Rarey, 2001]

2.4 Running Example – Building a DWH for a Health Insurance Company

The running example, which we introduce in this subchapter, will be used throughout the thesis in order to illustrate the topics discussed. Initially we will use it to represent the requirements collection and analysis phase, as well as the conceptual DWH design in the healthcare environment.

In our example, we show how a DWH will be built for a social insurance company located in Vienna, Austria. Around 350 000 freelancers and about 150 000 retirees are currently obtaining their social insurance from this company. The social insurance company is providing data about the insurants and their co-insured relatives to the "Hauptverband". ([Hauptverband, 2007] is the main association of Austrian social security institutions.) This data is centrally processed and forwarded to the e-Card operator. (e-Card is a robust and secure smart card-based patient identification solution that is in use in Austria nation-wide.) Physicians can access this data through their e-Card readers and find out if a particular patient has a valid health insurance, his (her) social insurance provider and the eligible insurance benefits.

Business management is striving to deliver best medical treatment to the insurants by applying the most effective therapies. In order to cope with future technology challenges and newest medical achievements, as well as to reduce administrative and treatment costs, this social insurance company wants to operate a data warehouse. The purpose of the DWH will be to make all the information about the insurants available, their health history, therapies, and drugs they received in the past, as well as the information about the physicians and institutions where they were treated. Additionally, charging information like drug and treatment costs, or insurant account details, are playing the crucial role in answering the main business questions and will therefore be added to the warehouse.

The DWH will deliver all available data in order to enable the detection of treatments which proved to be the most efficient for a given illness. In the first phase, business management concentrates on treatments for Diabetes Mellitus (E 11.9 Typ II and E 11.9 Typ I).

The approach to answering this business question is to identify the insurants who are suffering from Diabetes Mellitus. In case that the data quality in the diagnoses tables is not sufficient for a reliable conclusion, the most often prescribed Diabetes Mellitus drugs can be used for patient detection.

Table 2 gives a short overview of the most important data sources for the example DWH.

Data source	Data	Load cycle
Partners (hospitals, statutory health insurance physicians, etc.)	The majority of data originates from the health insurance partners and is stored in a mainframe DB2 database. Mainframe is used for complex calculations (i.e. for social security contribution) and the results of these calculations are stored in the DWH.	On a daily and on a weekly basis.

Chamber of Pharmacies	Prescription charging	On a monthly basis.
Hauptverband	"Erstatungskodex" reference table, which is necessary for the reimbursement calculation.	On a monthly basis.
External accounting partner	Accounting data	On a monthly basis.
Internal	Internal reference tables (tariff tables, physicians' directories, etc.) are maintained manually by internal company staff.	Continuously.

Table 2: Data Sources for the DWH

Main users of the DWH will be upper *business management*. They are interested in tracking the business development and thus need a DWH for analysing and forecasting their business processes.

The *controlling department* would be another heavy user of the DWH. This department is responsible for verification of reports and for internal controlling, e.g., for calculating the volume of work per employee. The *marketing department* is interested in analysing the drug consumption behaviour. They are interested in reduction of medication and treatment costs and deploy DWH facilities for knowledge discovery on this subject. *Specialist departments* are also potential users, and last but not least, a DWH may be frequently used for statistical purposes: for creation of monthly and yearly reports for the Hauptverband.

Figure 5 shows a proposed relational logical data model created with XCase modelling tool [RESolutionLtd, 2007]. The social insurance DWH is holding personal data about the insurant (*patient)*. A patient is uniquely identified by the social security number and possesses additional attributes: name, date of birth, gender etc. A patient is characterised by the relationships to other entities. A patient may have one or more drug *prescriptions*. The set of possible drugs is given in a *drug* entity, giving information about drug characteristics, like description, size, or pharmaceutical form. Each patient may have one or more *diagnoses*. These are specified by diagnosis code and classification. Since one patient can have many diagnoses, one diagnosis is only valid within a certain period of time. For each patient and each diagnosis made, there is a responsible clinician, stored in the *clinician* entity. For each diagnosis, there are one or more *therapies* assigned. Apart from therapy descriptive attributes, *charging* information is available as well. Information about the medical institutions (*clinic*) where the patients are treated and clinical facilities (*equipment*) are also presented in this model. In addition to data represented in this LDM, the DWH will contain data about staff skills and staff availability and data about the provided facilities.

Querying data that is not consolidated in a DWH is an exhausting task. For the time being, regular recurring queries, with slight differences, are written manually and are therefore hardly reusable. The introduction of OLAP-Tools would significantly simplify reporting and ad-hoc querying of data stored inside the DWH. This would lead to an increased number of warehouse users and therefore would be beneficial to the social insurance company as a whole.

The sample DWH which has been introduced in this chapter will be used as a starting point for examples in the following chapters.

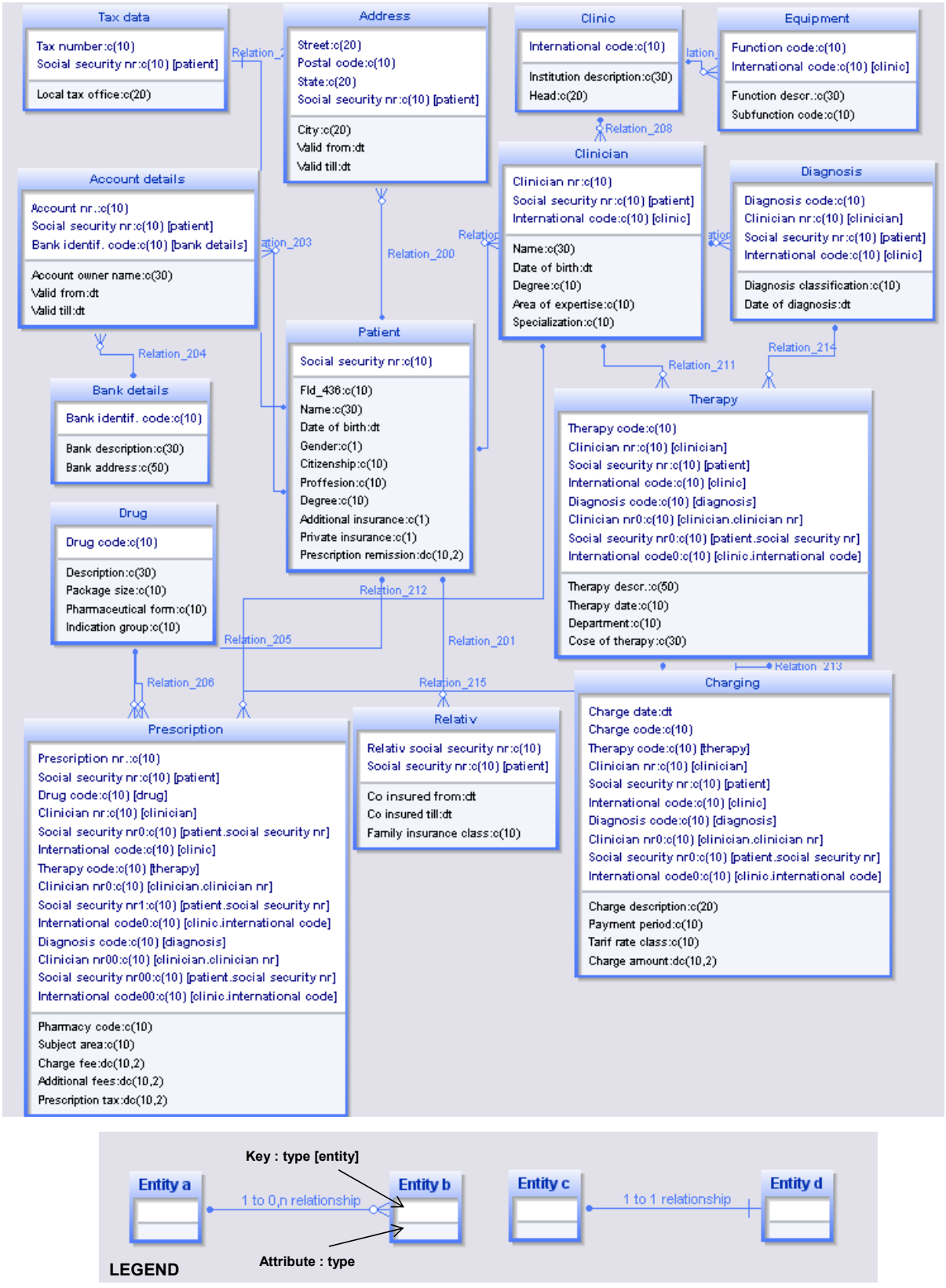


Figure 5: Logical Data Model for a Data Warehouse of a Social Insurance Company

An alternative approach, which has not been used in our running example, would be the multidimensional model for the DWH, as illustrated in Figure 6.

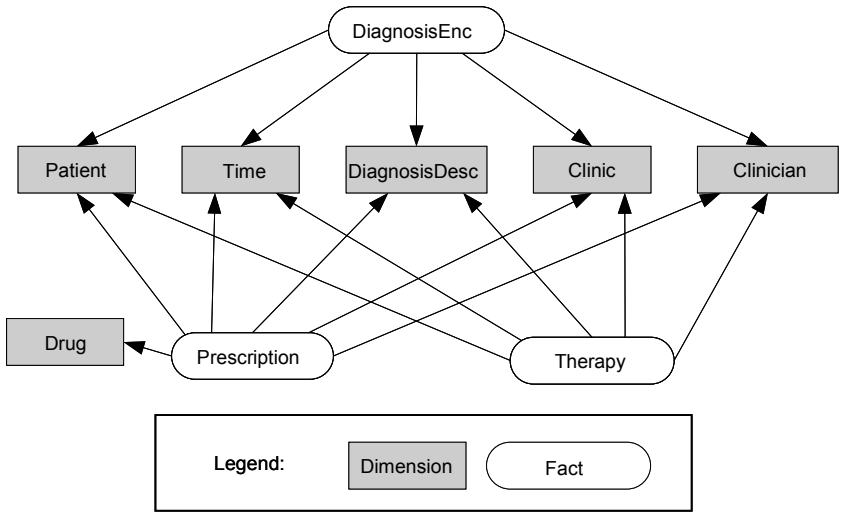


Figure 6: Multidimensional Model of the Sample Social Insurance Data Warehouse

The owners of the warehouse pointed out that the therapy charging and drug prescriptions are the most interesting issues which include crucial financial parameters. *Therapy* and *prescriptions* would in this model represent facts (with financial attributes as measures) while *patient*, *clinician*, *clinic*, and *drug* are typical dimensions for both facts. The necessary *time* dimension is also represented. Meanwhile, diagnosis can either be a dimension (particular therapy or prescription is the result of a diagnosis) or a fact, if observed as the process of asserting a diagnosis (as part of a patient encounter). Such fact can also be described by patient, clinician, clinic, and time dimension. It is necessary to separate the descriptive data characterizing the stated disorder (disease), which form a *DiagnosisDesc* dimension, from those illustrating the encounter, which are presented as the *DiagnosisEnc* fact. DiagnosisDesc would also be a dimension for Prescription and Therapy facts.

3 Data Warehouse Facilitating Evidence-Based Medicine

Evidence-based medicine is a new direction in modern healthcare. Its task is to prevent, diagnose, and medicate diseases using medical evidence. In order to obtain the best evidence for a given disease, external clinical expertise as well as internal clinical experience must be available to the healthcare practitioners at the right time and in the right manner. External evidence-based knowledge can not be applied directly to the patient without adjusting it to the patient's health condition. Integration of external evidence-based data sources into the existing clinical information system, and finding an appropriate therapy for a given patient and a given disease is a great research challenge. In this chapter, we explain the role a data warehouse can play in the area of evidence-based medicine and we describe its contribution to development and dissemination of evidence-based knowledge. A clinical data warehouse that facilitates evidence-based medicine is a reliable, powerful, and user-friendly platform for strategic decision making and has a great relevance to the practice and acceptance of evidence-based medicine.

Since it is a difficult task for clinicians to gather all necessary knowledge about given diseases, the practice of evidence-based medicine would not be imaginable without IT support. With rapid changes taking place in the field of healthcare, decision support systems play an increasingly important role. Healthcare institutions are deploying DWH applications as decision-support tools for strategic decision-making.

The combination of data warehousing technology and evidence-based medicine opens an innovative application field of information technology in healthcare industry. Medical institutions, as well as health insurance companies, are primarily interested in increasing the patient healing rate and reducing treatment costs. In the long term, the application of DWH in the area of evidence-based medicine could prove economical by avoiding the duplication of examinations, saving time through automation of routine tasks, and simplifying the accounting and administrative procedures.

Caused by the growing and aging population, chronic illnesses are going to become the major concern of the healthcare industry. Diseases of elderly people, like Diabetes, Alzheimer's disease, cardiac insufficiency, and sight loss (macula degeneration) will cause more treatment effort and therapy costs than the treatment of most difficult illnesses (cancer and heart attack) generate nowadays. Since these diseases can be treated more efficiently and more cost-effectively when detected in the early stages, the mission of modern medicine is to become able to recognize the patterns of disease formation and development. Evidence-based medicine deals with the analyses of the existing medical records, clinical studies etc., and searches for the recurring samples of disease symptoms. Data warehousing and data mining techniques play a crucial role in acquisition and gathering of existing medical experience from diverse data sources and in statistical analysis of that data. The extracted experience values and formulated knowledge (evidence-based guidelines) are used for more efficient prediction, discovery, and treatment of diseases.

In this chapter, we present some application fields which are relevant to the clinical knowledge management, especially:

- Developing new knowledge – the data warehouse-supported creation of evidence-based guidelines and clinical pathways
- Knowledge sharing – the data warehouse as an easy to use platform for knowledge dissemination among healthcare decision makers.

3.1 Evidence-Based Medicine

Testing the outcome of medical interventions has been performed for hunderds of years. During last century, this effort started to impact all fields of welfare and healthcare. One of the founders of evidence-based practice was professor Archie Cochrane, a Scottish epidemiologist, whose angagement in this field resulted in an increased acceptance of the concepts behind evidence-based medicine. He was the first one to point out and promote the vital importance of use of medical evidence resp. randomised controlled trials for improving the effectivness of treatments. A *randomized controlled trial* (RCT) is a scientific procedure most commonly used in testing medicines or medical procedures. It is a trial that uses randomized control. This is considered the most reliable form of scientific evidence because it eliminates all forms of spurious causality [Wikipedia, 2007a].

Cochrane's work was honoured through naming of centres of evidence-based medical research — The Cochrane Centres — and an international organisation, The Cochrane Collaboration, after him. The Cochrane Centers coordinate activities, primarily in language-defined regions, and are the main contact point for the public. The Cochrane Collaboration [Cochrane_Collaboration, 2007] is a world-wide endeavour dedicated to tracking down, evaluating, and synthesising RCTs in all areas of medicine. It is a major force in the EBM movement. The Cochrane Collaboration provides the

Cohrane Library [Cochrane_Library, 2007], a collection of medical databases that contain high-quality, regularly updated independent evidence to support healthcare decision-making. It includes reliable evidence from Cochrane and other systematic reviews, clinical trials, and more.

The explicit methodologies used to determine "best evidence" were largely established by the McMaster University (Canada) research group led by David Sackett and Gordon Guyatt. According to this group, *evidence-based medicine* is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research. By individual clinical expertise we mean the proficiency and judgement that individual clinicians acquire through clinical experience and clinical practice [Sackett et al., 1996]

Since 1992, when the evidence-based medicine research group at McMaster University was founded, the number of articles about evidence-based practice has grown exponentially (from 1 publication in 1992 to about a thousand in 1998) and international interest has led to the development of 6 evidence-based journals (published in up to 6 languages) that summarize the most relevant studies for clinical practice and have a combined world-wide circulation of over 175.000 [CEBM, 2004].

3.1.1 Concepts of Evidence-Based Medicine

Most clinical practice is based on limited evidence, like textbook information, sometimes defective research or case studies, unverified reviews, and personal experiences. In their everyday practice, clinicians constantly strive to offer the best suitable treatment to their patients. Traditionally, they would consult their manuals, textbooks or their senior, more experienced colleagues in order to solve the problem. With time, they would gain a lot of expert knowledge themselves and would act as advice-givers to their junior colleagues. This way of developing medical expertise is natural, but not always the most optimal for the patients.

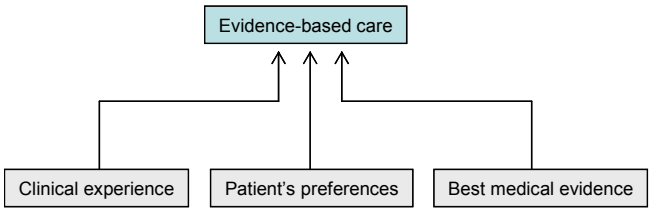


Figure 7: Basis of Evidence-Based Care

The evidence-based medicine is built on another idea. Its task is to complement the existing clinical decision-making process with the most accurate and most efficient research evidence (Figure 7). For example, when treating a Diabetes patient suffering from a progressive liver disease, his (her) clinician has to find the most efficient therapy which does not conflict with the patient's ongoing Diabetes treatment. The clinician searches through evidence-based guidelines to find current best evidence for

treating liver diseases, and verifies whether the proposed method fits into the Diabetes patient's health risks.

Evidence-based guidelines explicitly define the decision points to which this valid evidence needs to be integrated with the individual clinical experience in deciding on a course of action. Thus, they don't inform the clinician which decision to make and take away his (her) authority for decision-making. Instead, they identify the range of potential decisions and provide the physicians with the evidence which, when added to individual clinical judgement and patient's values and expectations, will help them make their own decisions in the best interest of the patient [Audiimoolam et al., 2005].

Although it is always stated that evidence-based medicine is based on evidence, this does not mean that traditional medicine does not rely on it. The traditional medicine uses both research experience and evidence as well, but the quality of this information (in terms of accuracy and timeliness) is much lower than of one electronically stored in the EBG repositories.

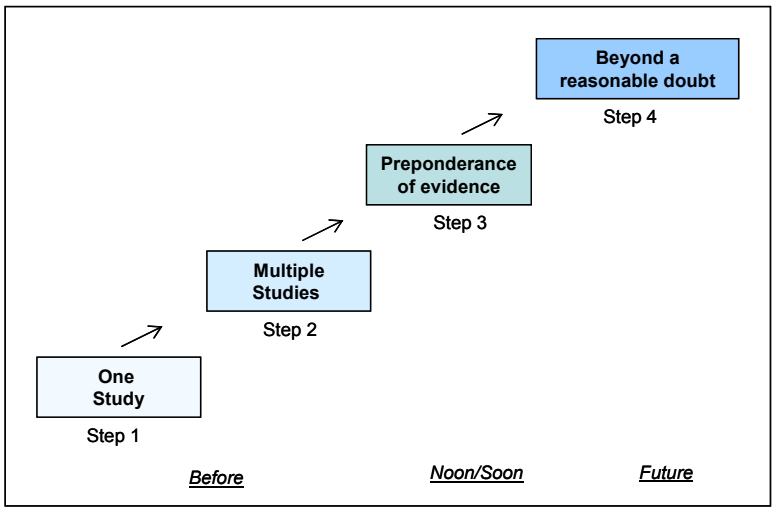


Figure 8: Hierarchy of Evidence [Carner®, 2007]

One of the greatest achievements of evidence-based medicine has been the development of systematic reviews and meta-analyses, methods by which researchers identify multiple studies on a certain topic, separate the best ones and than critically analyse them to come up with a summary of the best available evidence [White, 2004]. The quality and the reliability of evidence can be classified as shown in Figure 8.

The main information sources providing accurate medical evidence are presented in Figure 9. After the information is collected, it is analysed and used for generation of evidence-based guidelines. The guidelines are used for prevention and treatment of diseases, including the forecast of relevant patient state parameters. This way, they will push for effective therapies and measures to replace ineffective ones.

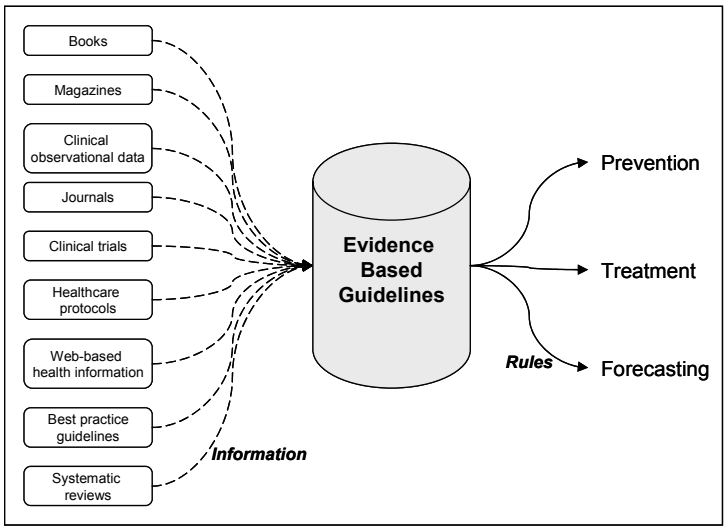


Figure 9: Data Flow in Evidence Based Medicine

Usually, the guideline can be represented as a tree, with choice nodes based on tests or inquiries and leaves describing actions. An example is given in Figure 10, representing a selection of pacemaker systems for patients with sinus node dysfunction. (AV = atrioventricular).

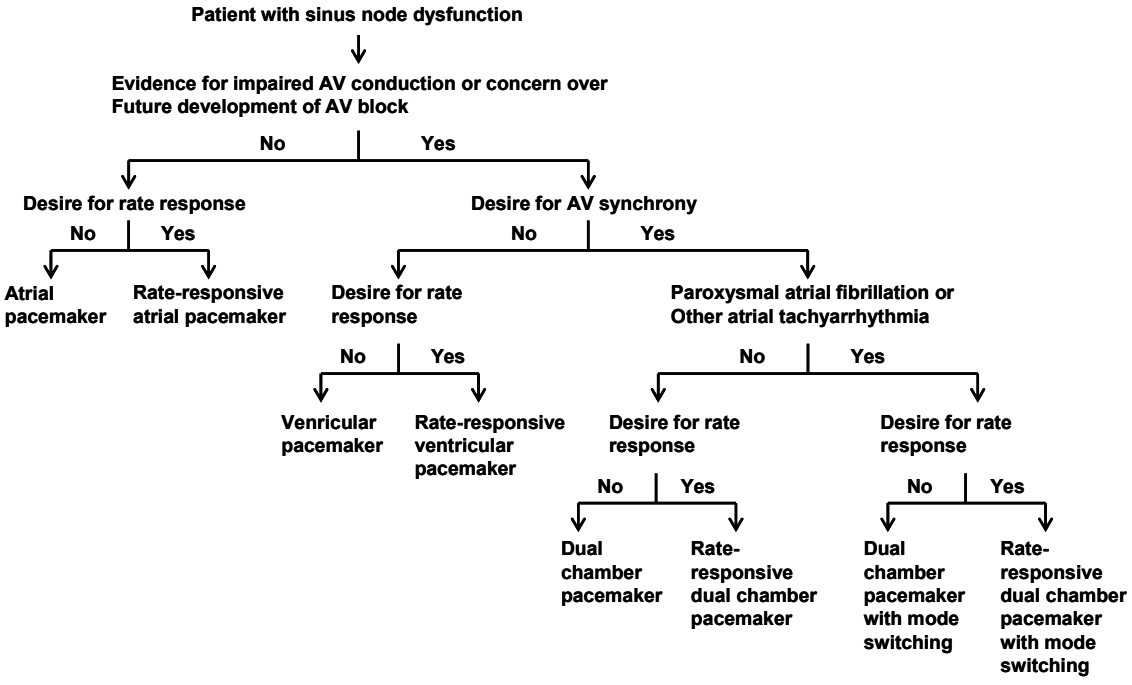


Figure 10 : Example EBG: Selection of Pacemaker Systems [Gregoratos et al., 1998]

[Craig et al., 2001] summarizes advantages and disadvantages of the adoption of evidence-based guidelines in decision-making process as follows:

Advantages:

- Very comprehensive: summarizes all relevant research information about all possible interventions for a common clinical problem; improved power to detect small and important differences.
- Very useful applicability information, explores the trade-off of benefit and harm.

Disadvantages:

- Can be difficult to use if not formatted with the end-user in mind.
- May quickly become out of date.

3.1.2 Practising Evidence-Based Medicine

The major steps in practising evidence-based medicine are:

1. formulate answerable questions
2. find the best evidence
3. evaluate the evidence for validity and usefulness under concrete circumstances
4. apply the evidence
5. evaluate the performance

Evidence-based medicine is required when:

- *a diagnosis needs to be established*

  A clinician, who may be overwhelmed with the variety of symptoms, can consult evidence-based rules to determine the right diagnosis.

- *a patient needs to get appropriate therapy*

  Here, a clinician is searching for the best-fitting therapy that is proven to be efficient. Evidence-based guidelines will usually offer a few treatments, but only those that do not conflict with the patient's preferences and health risks will be considered.

- *a prognosis is needed*

  The external knowledge about the patient's future health state is provided. For example, if an elderly patient is suffering from osteoporosis, a prognosis of his (her) quality of life can be generated comparing the option of patient undergoing a surgery to them deciding to take medications.

- *a etiology needs to be clarified*

  In this case, evidence-based medicine provides experts knowledge about the possible causes of a given disease. For example, how much nicotine consumption increases the risk of a heart attack.

3.1.3 External Sources of Clinical Evidence

External clinical evidence is available in numerous medical databases. One of the most important sources is the Cochrane Library [Cochrane_Library, 2007]. Evidence-based rules can also be found in journals with evidence-based focus, high-quality textbooks, protocols, medical journals (e.g. British Medical Journal [BMJ, 2007], the Journal of American Medical Association [JAMA, 2007], and The Lancet [Elsevier, 2007]), in databases with an evidence-based focus, web resources, clinical guidelines, clinical audits, clinical accreditations, clinical appraisals, clinical governance, and health technology assessments.

Table 3 [White, 2004] shows some of the leading clinical evidence resources available to clinicians.

Name	Description	Publisher/Sponsor
JOURNALS		
ACP Journal Club http://www.acpjc.org	Bimonthly journal that analyzes the content of over 100 clinical journals and summarizes those articles found to have scientific merit and relevance to medical practice.	American College of Physicians
American Family Physician http://www.aafp.org/afp	Twice monthly clinical review journal that contains evidence-based components, such as POEMs (patient-oriented evidence that matters), Cochrane for Clinicians and Point-of-Care Guides.	American Academy of Family Physicians
Bandolier http://www.jr2.ox.ac.uk /bandolier	Monthly journal that searches PubMed and the Cochrane Library for systematic reviews and meta-analyses published in the recent past and summarizes those that "are both interesting and make sense."	Produced from Pain Research at Oxford University with multiple sponsors
The Journal of Family Practice http://www.jfponline.org	Monthly clinical review journal that contains evidence-based components, such as its online archives of POEMs.	Dowden Health Media
EVIDENCE SUMMARIES		
Clinical Evidence http://www.clinicalevidence.com	A compendium of systematic reviews, gathered from Cochrane, MEDLINE and other sources, updated and expanded every six months.	BMJ Publishing Group
The Cochrane Database of Systematic Reviews	Arguably the most extensive collection of systematic reviews.	The Cochrane Collaboration

http://www.cochrane.org/cochrane/revabstr/mainindex.htm		
DynaMed http://www.dynamicmedical.com	A database of summaries of the evidence drawn from sources such as Clinical Evidence and the Cochrane Library.	DynaMed
FIRSTConsult http://www.firstconsult.com (formerly PDxMD)	A database of evidence summaries drawn from Cochrane, Clinical Evidence, the National Guideline Clearinghouse and others.	Elsevier
InfoRetriever http://www.infopoems.com	A search engine with access to evidence-based sources such as POEMs, Cochrane, clinical rules, a diagnostic test database, practice guideline summaries and Griffith's Five-Minute Clinical Consult; subscribers also receive Daily POEMs via e-mail.	InfoPOEMs Inc.
SUMSearch http://sumsearch.uthscsa.edu/	A search engine that gathers evidence-based clinical information from MEDLINE, DARE and the National Guideline Clearinghouse.	The University of Texas Health Science Center
TRIP Database (Turning Research Into Practice) http://www.tripdatabase.com	A search engine that gathers evidence-based clinical information from MEDLINE, DARE, the National Guideline Clearinghouse and many other evidence-based Web sites.	Gwent, Wales
The York Database of Abstracts of Reviews of Effects (DARE) http://www.york.ac.uk/inst/crd/darehp.htm	A collection of abstracts of systematic reviews.	Centre for Reviews and Dissemination, University of York
CLINICAL GUIDELINES		
Institute for Clinical Systems Improvement (ICSI) http://www.icsi.org/knowledge/	Guidelines for preventive services and disease management developed by ICSI, an independent, nonprofit collaboration of healthcare organizations, including the Mayo Clinic, Rochester, Minn.	Institute for Clinical Systems Improvement
National Guideline Clearinghouse http://www.guidelines.gov	Comprehensive database of evidence-based clinical practice guidelines.	The Agency for Healthcare Research and Quality
U.S. Preventive Services Task	Recommendations for clinical preventive	USPSTF

Force (USPSTF) Recommendations http://www.ahrq.gov/clinic/uspstfix.htm	services based on systematic reviews by the U.S. Preventive Services Task Force.	
OTHER		
DailyPOEMs http://www.infopoems.com	Daily e-mail update with approximately 30 POEMs per month; subscription includes InfoRetriever access online or via PDA (above).	InfoPOEMs Inc.

Table 3: Clinical Evidence Resources [White, 2004]

3.2 Clinical Decision Support Systems Embracing Evidence-Based Guidelines

The simplest version of a clinical data warehouse facilitating evidence-based medicine is presented in Figure 11.

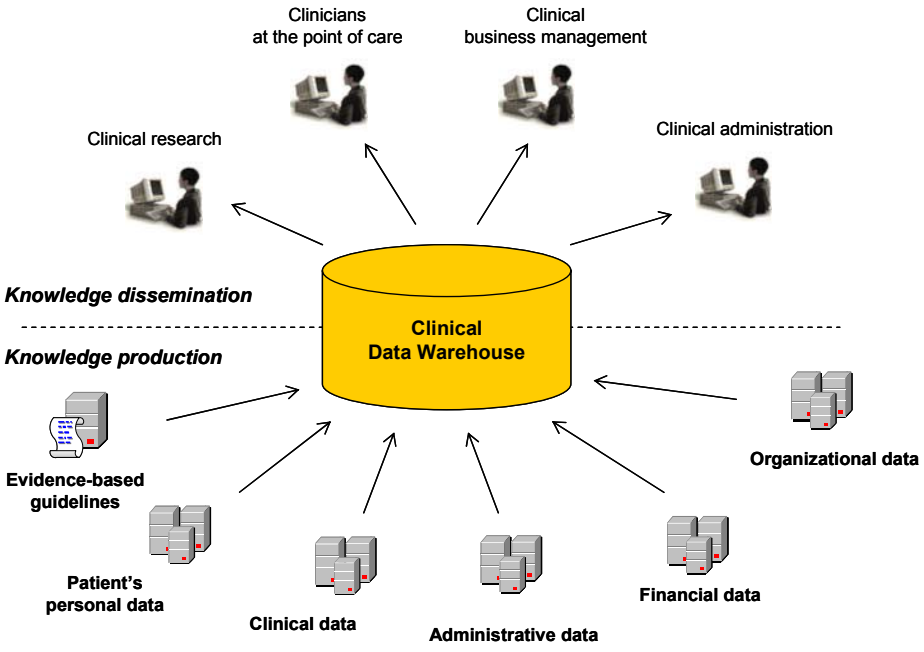


Figure 11: Sources and Users for Clinical DWH

The medical DWH above unifies data from diverse departments as well as evidence-based guidelines. Ideally, in such a DWH, data is prepared to be queried and analysed in any way required. Clinical management is often interested in finding out which treatments and medications led to more rapid and more economic patient convalescence. Data mining and OLAP functions would support business

decision-makers in creating the most effective business strategies that satisfy both patients' expectations and the requirements for financial optimisations. If administrative data is available in the DWH, it could be combined with evidence-based medicine recommendations in order to give advice regarding the right resources needed, e.g., number of skilled staff needed for certain medical treatments. Such information could be used further for work and treatment scheduling, which is essential in supporting medical decision-makers in the area of human resources.

Clinical decision support systems (CDSS) are active knowledge systems which use two or more items of patient data to generate case-specific advice [Wyatt and Spiegelhalter, 1991]. CDSS are deployed in order to reduce medical errors, increase the efficiency and quality of healthcare and to decrease treatment costs. The effectiveness of such systems is highly dependable on the strength of the underlying evidence base. A CDSS facilitating evidence-based medicine can be fully successful only if it is able to follow and integrate the changes coming out of the clinical research. This means that CDSS should be capable of:

1. monitoring the literature for high quality research findings,
2. extracting the best clinical evidence,
3. incorporating new knowledge into the existing decision support knowledge base.

The tasks listed above are still not fully automated. It is an open research issue to create machine-interpretable sources of evidence and thus automate the whole process of clinical evidence integration into a CDSS.

[Perreault and Metzger, 1999] outline the four functions of electronic clinical decision-support systems:

- *Administrative*: Supporting clinical coding and documentation, authorization of procedures, and referrals.
- *Managing clinical complexity and details*: Keeping patients on research protocols; tracking orders, referrals follow-up, and preventive care.
- *Cost control*: Monitoring medication orders; avoiding duplicate or unnecessary tests.
- *Decision support*: Supporting clinical diagnosis and treatment plan processes, promoting use of best practices, condition-specific guidelines, and population-based management.

3.2.1 Systems Integration

The DWH is one major part of CDSS, providing an easy to use platform to support decision-making processes of care-givers and clinical managers. EBM-oriented CDSS can vary in their scope. The simplest systems would be fed by data about diseases and best practice guidelines to support care delivery. More sophisticated systems additionally include various clinical internal and external data sources. In order to build a DWH to support evidence-based medicine, two basic operations need to be assured:

- Properly integrated source information systems
- Easy access to data and decision support tools for the wide area of users.

Here we give a more detailed list of the relevant data sources for the CDSS, based on the high-level overview shown in Figure 11:

- **Evidence-based guidelines**
- **Clinical data**
  - patient data
  - pharmaceutical data
  - medical treatments
  - length of stay
- **Patient's personal data**
  - Demographic data (age, gender, height, weight…)
  - Health risks (allergies, insufficiencies, …)
  - Personal preferences
- **Administrative data**
  - staff skills
  - overtime
  - nursing care hours
  - staff sick leave
- **Financial data**
  - treatment costs
  - drug costs
  - staff salaries
  - accounting
  - cost-effectiveness studies
- **Organisational data**
  - Room occupation
  - Facilities
  - Equipment

3.2.2 Decision-Support Tools

Decision-support tools support the practitioner's ability to correctly identify the problem when examining the patient and to formulate that problem in a clear question. When the answer to a defined problem is retrieved from the CDSS, decision support tools represent the result in an understandable manner. In order to find the best-fitting treatment, it is essential to assess the health differences between the examined patient and those involved in the study.

An important concern of evidence-based practice is that guidelines can not be applied to patients without checking if they fit into the patient's health risks. Problems faced so far in practice are that the physicians often avoid evidence-based approaches because they feel forced to apply recommended treatments and not free to create personalised, tailored care. The most efficient way to diminish this resistance is by proving that the evidence-based medicine significantly helps to achieve better results and by providing the practitioners with easy-to-use decision-support tools. In this sense, the Internet is becoming an innovative way to easily extract and mine data.

Apart from just being queried on demand, decision-support tools can act proactively, for example by alerting physicians when a change in a patient's condition occurs or when duplicate medications are issued. Another example is that they can send an alert when a physician prescribes a drug that doesn't fit into the entire range of patient's health risks [Briggs, 2005].

3.2.3 Application Fields

We think that the most relevant functions of DWH in the area of evidence-based medicine are marked by the ability:

1. to support the generation process of the evidence-based guidelines,
2. to support the clinicians at the point of care delivery, by making evidence-based rules available,
3. to support the control of clinical treatment pathways,
4. to support the administrative and management tasks, by providing evidence-based knowledge as well as diverse organizational and financial data.

Apart from these most common application fields, it could be used for processing huge amounts of medical data in the area of epidemiology. In the following subchapters, we are going to explain these ideas in more detail.

3.3 Deploying Data Warehouses in Generation of Evidence-Based Guidelines

For practitioners treating patients it is not feasible to access and evaluate all primary studies applicable to an individual patient. Given the volumes and complexity of the clinical data, creation of evidence-based guidelines requires a computer-based information management system.

Data warehousing and data mining can support creation of evidence-based rules by providing a platform and tools for knowledge discovery and pattern recognition. Large amounts of data are analysed to confirm known or discover unknown trends and correlations. The discovery process in medical research would benefit enormously from data mining facilities.

Figure 12 represents our rule-generation process based on clinical evidence. In our model, data originating from different medical sources is extracted, transformed, and prepared for loading into the existing DWH structure. The DWH contains diverse patient, clinical, and pharmaceutical data. Supported by a DWH, medical knowledge workers are able to analyse and mine a vast quantity of available data.

Data mining techniques enable identification of trends and recognition of best practices for different disease treatments. Knowledge workers are interested in finding new associations and rules which are hidden in the data. The new rules need to be exhaustively checked for their correctness before being declared valid.

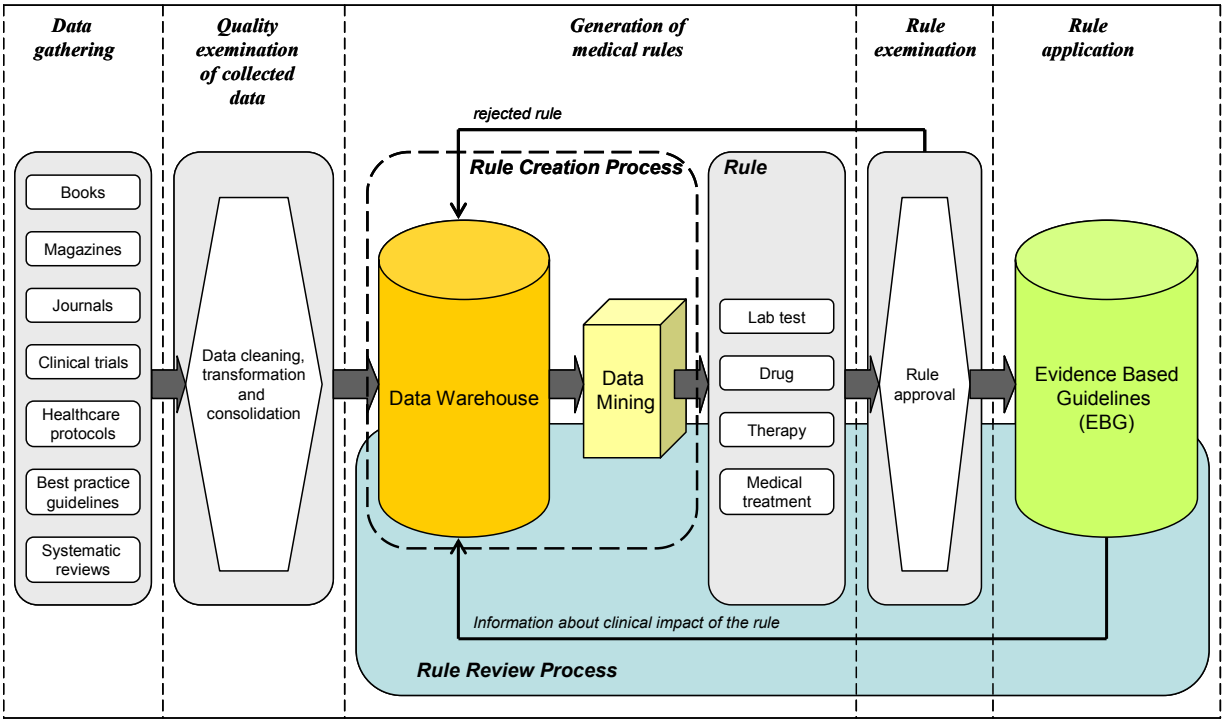


Figure 12: Generation of Evidence-Based Guidelines

.After a rule (in form of laboratory tests, recommended drugs, therapies, or medical treatments) has been created, it needs to be examined and approved by a higher authority. In case of rejection, the proposed rule is sent back to the DWH for further development. If the rule is approved, it is added to the database holding evidence-based guidelines. Once added to the evidence-based guidelines, rules undergo a systematic review process.

Schuerenberg describes the governance process to develop and review clinical guidelines, as it is practiced at the University Health Network (Toronto), as follows: The IT clinical advisory committee starts the process by recommending which lab tests should be ordered for which patient conditions, and why. Its recommendations are sent to an enterprise-wide medical advisory committee that has the final approval on which evidence-based rules and alerts are programmed into the decision support system. University Health's IT department then enters the new rule and its supporting evidence into the application. The medical advisory committee measures the clinical impact of each new automated

rule six months after it is programmed. Rules are often modified during the review process and then reviewed again in another six months [Schuerenberg, 2003].

Development of evidence-based guidelines using DWH and data mining goes far beyond just reviewing the literature. These guidelines are a unification of best evidence with clinical expertise and patient values. They are reported in sufficient detail to allow clinicians to make judgements about the validity of their recommendations and to improve patient care.

3.4 Controlling Clinical Treatment Pathways with Data Warehouses incorporating EBM

Healthcare organisations are searching for methods to rationalise their processes, to improve healthcare and ultimately also to reduce costs. As stated by the Health Informatics Research (HIR) group [HIR-group, 2006], traditional business process modelling tools and business process execution tools, such as workflow management systems, lack support for the complex, multi-organisation, dynamic, and large scale patient treatment processes that exist within the healthcare system. New research programs are launched to develop methodologies, tools, and techniques that can be applied to the more complex clinical pathway process.

Clinical Pathways are structured, multidisciplinary plans of care which are designed to support the implementation of clinical guidelines and protocols. They are built to support the overall clinical management, the clinical and non-clinical resource allocations, and last but not least the clinical audit and financial management [Open_Clinical, 2006].

Clinical practice guidelines, which often underlie clinical pathways, are primarily the responsibility of professionals, while chain management is that of managers. Therefore, clinical pathways can be considered to be a joint effort designed with the motive to improve the patient outcomes and enhance the quality of diagnosis, interventions, and management [Audiimoolam et al., 2005].

An example of improved effectiveness of the healthcare delivery during last few years are Australian hospitals, which have been increasing application, standardisation and observation of clinical pathways [Uni_Münster, 2001], [Dowsey et al., 1999].

In the area of clinical pathways, the DWH facilitating evidence-based medicine could be used for controlling the clinical processes, from patient's admission to his (her) release. For frequently occurring diseases (like Diabetes mellitus, pneumonia, hernia etc.), the whole treatment process, from diagnosis to therapy, could be verified against clinical pathways. By analysing all relevant data stored in the DWH, it would be possible to find out if prescribed levels (operation, recovery duration) were reached on time. In case of a significant delay, the DWH could alert the responsible clinician. The discrepancy between the clinical pathway and the actual treatment can be caused by unnecessary modifications of the therapy, by management problems, but also by an incorrect rule. Clinical pathways are reviewed on a regular basis (once or twice a year). Frequently occurring deviations of a

pathway are analysed, and, in case of legitimate causes, the pathway may be reformulated. This consequently contributes to the pathway refinement process.

The deployment of data-mining techniques in discovery of patterns within historically applied treatment processes would be an efficient method for clinical pathway development. Based on patient record data, administrative data, clinical log data, and evidence based rules, the mining process could be applied. This way the structure of clinical paths and the sequence of activities could be detected, a task that could hardly be done manually.

The development of clinical pathways is knowledge-intensive and it requires the cooperation among knowledge workers, clinicians, nurses, and clinical management. Data miners aim to combine the experience from the evidence-based guidelines with the concrete clinical data and to identify (time) dependency patterns. Lin et al. [Lin et al., 2000] state that by obtaining the time dependency patterns, the paths for new patients can be predicted when they are admitted to a hospital, and, in turn, the care procedure can be designed more effectively and more efficiently.

The goal of the use of DWH, data mining, and evidence-based medicine in the domain of clinical pathways is to improve the coordination of care between different clinical departments and therefore to improve the quality of care and to reduce the length of patient's hospitalisation. Apart from this, deploying clinical pathways for negotiations with cost units proved to be very effective, since they represent a detailed description of the range of services and benefits for a given patients' group. Cost units get a better view of activities offered and hospitals can better support the argument about the estimated costs of treatments. For this reason, clinical pathways represent the perfect foundation for discussions between clinical management and sponsors. Consequently, they oblige the clinical economist to take responsibility for clinical decisions, since the financial factors influence the quality of treatments and procedures proposed by a clinical pathway.

Open investigation areas include the development of new clinical pathway transformation frameworks, pathway modelling techniques, and new clinical pathway process management systems.

3.5 Evidence-Based Decision Support at the Point of Care

In order to illustrate the use of DWH at the point of care, we will use an example scenario that is based on our example DWH introduced in chapter 2.

Figure 13 illustrates our idea to use a DWH facilitating EBM at the point of care. In this scenario, the clinician is querying the DWH while examining the patient, so the answers he (she) is expecting must come quickly and be presented clearly. The clinician is interested in finding the best-fitting treatment for the given patient and given disease.

The clinical question will be defined based on the patient's disease. The clinician uses an OLAP tool in order to query the DWH. Standard, predefined reports, as well as ad-hoc queries, may be used. After this, selected tables would be joined inside the DWH on the fly.

The clinician wants to analyse patient's historical health data stored in the DWH (entities *diagnosis* and *therapy*), as well as the *drugs* which were prescribed in the past. Patient's clinical dossier, accompanied by patient's age and lifestyle habits (originating from entity *patient*), would than be combined with evidence-based guidelines. Supported by the DWH functionality (aggregation, slice and dice etc.) as well as by OLAP and data mining techniques, those treatments which proved to be the most efficient in similar cases would be detected. It is important to point out that the guidelines are only suggestions and not obligatory steps in patient's treatment. The practitioner has the choice of acting differently from the proposed guidelines. Having all necessary information, the clinician can decide how the patient will be treated.

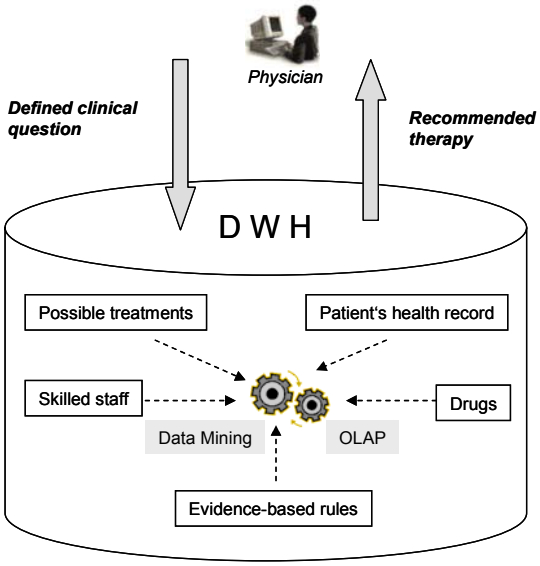


Figure 13: Decision Support at the Point of Care

Since the DWH in our scenario integrates clinic-wide data from different departments, the clinician would be able to immediately check if the proposed treatments can be carried out (if necessary facilities (entity *equipment*) or skilled stuff are available).

4 Related Work and Opportunities for Data Warehousing

This chapter reviews related work on building effective and lasting healthcare information systems for exchange of patient medical data between care providers, social insurances and other participants of the healthcare network. Since there has been almost no work done to address facilitation of EBM in this context, we will specify the opportunities that DWH technology would bring in this field. Concluding, we formulate the basic rules to be followed when designing a sustainable federated DWH for support of EBM.

4.1 Related Work – Existing Approaches to Medical Information Sharing and Clinical Decision Support

Over the last decade, many countries have recognized the value of information exchange in the healthcare sector and have started initiatives for creation of shared electronic health records. Some of them went a step further by integrating decision support systems into their healthcare networks, primarily in order to support physician's decision making process at the point of care.

Although the advantages of healthcare information integration are clear, divergence still exists about how such integration should be achieved. Because of the complexity of clinical information as well as very heterogeneous and rapidly growing source databases involved, an integrated system needs to fulfil some important requirements like flexibility, maintainability, and scalability. In this subchapter, we will give an overview of the most relevant related work.

4.1.1 Service Oriented Architecture (SOA)-Based Network Healthcare System

Canadian healthcare system is comprised of more than 700 hospitals, almost 400.000 general practitioners, local pharmacists, nurses, and many community care centres. *Canada Health Infoway* [Infoway, 2007a] is an organisation that aims to improve the collaboration and accelerate the use of

electronic health information systems and electronic health record across Canada. Since 2001, this organisation supports a common, nation-wide framework of electronic health systems, which are increasingly providing care givers with rapid access to accurate patient information, enabling better decision support at the point of care. The goal of the *Infoway* is to integrate the provincial health information systems and connect them to a pan-canadian health network. By 2010 a new, modern, and secure healthcare information system should be put in place and provide 50% of Canadian population with EHRs [Infoway, 2007b].

For the implementation of their healthcare network, Canadian authorities chose to create a series of *hub-and-spoke repository systems,* which store the relevant patient information in jurisdictionally coordinated repositories. The main reason for choosing this architectural approach was the fact that there were no strong existing electronic medical structures and, even more importantly, that the national government had a great financial control over the healthcare system.

4.1.1.1 Service Oriented Architecture (SOA)

Canada Health Infoway has proposed an information infrastructure for the network healthcare system, one that is based on service oriented architecture (SOA) and provides standards for sharing data and services. SOA integrates heterogeneous information systems by means of services that represent different systems' functionality, independent from local platforms and programming languages. Its main concepts are illustrated in Figure 14 and described by [Sartipi et al., 2007] as:

- *Application frontend*: use the business processes and services within the system. Among others, these may be components for the physicians, pharmacy, lab, public health services etc.
- *Service*: consists of implementation, service contract, functionality, constraint specification, and service interface. It is distributed among different parts of SOA.
- *Service repository*: stores service contracts. Figure 14 shows three main service repositories: "registries data and services", "ancillary data and services", and "EHR data and services".
- *Service bus*: connects fronted to the service. (HIAL in Figure 14)

The integration of the heterogeneous local healthcare systems in a network is done within the Infoway's HIAL communication mechanism. Following techniques are proposed:

1. *Technology gateway*, which plays the role of the mediator between two technologically distinct components. It facilitates interoperability between such modules by using each other's services.
2. *Facade* includes a common service access layer which provides standard access to the services of the legacy systems.

Canada Health Infoway is expected to extend its domain by adding two new general services to the SOA: decision-support system component and mined-knowledge services [Sartipi et al., 2007].

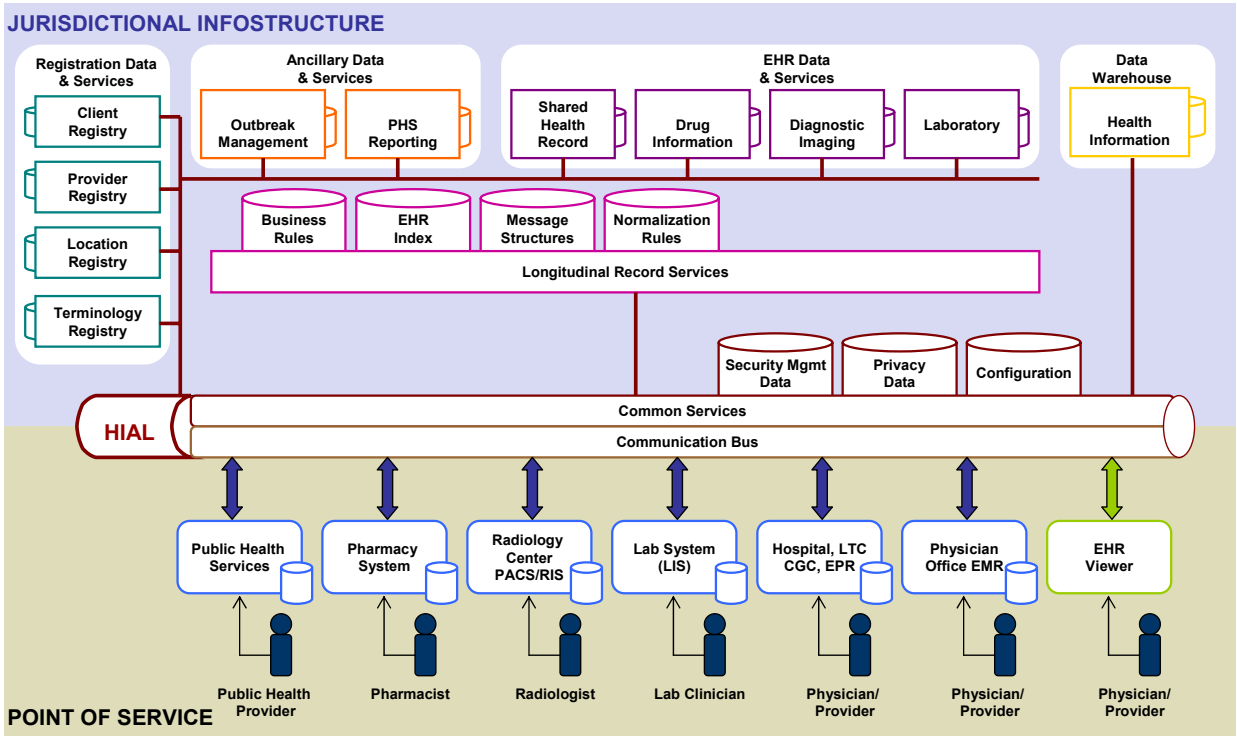


Figure 14: Canada Health Infoway Infrastructure [Sartipi et al., 2007]

4.1.1.2 Decision Support System and Mined Knowledge Services

The idea of adding these new knowledge-based components to the SOA infrastructure is to provide the knowledge available in a specific application domain through standard services.

The services provided by the Decision Support System component would be twofold: (1) standard workflow services in form of clinical guidelines and (2) adaptable workflow generation services which would allow users to define workflows for their organisations.

Mining knowledge services would be implemented in order to enhance decision making process. Data mining service would be located in the "EHR Data and Services" part, while clinical guidelines (service repository for data mining) would be located in the "Registries and Services" part of the Infoway infrastructure [Sartipi et al., 2007].

4.1.1.3 The Assessment of This Approach

The described SOA based approach for support of nation-wide healthcare information exchange is doing well in providing seamless access to remote services of the underlying components. It is scalable and allows for easy addition of new components, and therefore the building of a nation-wide EHR seams to be realistic goal. It may be used for purposes of tele-medicine (which is certainly an important issue in such a large country), but it has major limitations when it comes to decision support.

This infrastructure is primarily designed for transporting of clinical documents through a communication framework and it is not suitable for performing any automatic statistical analysis of patient data. The described approach lacks the possibility to store historical patient data (centralized or decentralized) on atomic level and to facilitate a huge knowledge base which could be queried in any desired way by the care givers, business, and administrative staff. Although the intended expansion of the system by adding mined-knowledge services opens new application possibilities, such as exploration of spread of epidemics or distribution patterns for certain groups of patients or health services, it does not shift its focus to application of evidence-based medicine in everyday handling.

4.1.2 Integrating the Healthcare Enterprise (IHE) Approach

In Lower Austria, there are 22 hospitals distributed over 27 locations and encompassing approx. 8.200 beds. Currently, 7 different health information systems from 4 different vendors are in use. In addition to the hospitals, around 3.000 general practitioners are providing healthcare to approx. 1.5 million inhabitants [Krenn, 2006]. At the time the NÖMED WAN project was started, no unique patient identifier existed, medical records were isolated and stored at the particular care giver devices. Clinical documentation was hardly re-used and exchange of patient information occurred rarely and only on request.

The Lower Austrian Health and Welfare Fund Area of Health (NÖGUS), which is in charge of financing, quality assurance, and supply planning for the region's healthcare system, decided to develop a shared network for exchange of patient's healthcare documents. This seminal, successfully implemented project, endorses the use of the future-proof Integrating the Healthcare Enterprise [IHE, 2007a] standard, which protects the investments through its vendor-independent extendibility.

4.1.2.1 The IHE Initiative

Integrating the Healthcare Enterprise (IHE) is an initiative designed to stimulate the integration of the information systems that support modern healthcare institutions. IHE initiative recommends the use of existing standards, such as HL7, DICOM, IETF, and others, instead of defining new ones.

IHE is strongly supported by the industry: more than 160 companies have developed IHE-compliant systems between 1999 and 2005 and participated in cross-vendor testing events, Connect-a-thon [IHE, 2007b],[IHE, 2007c] organized by IHE. This means that standards recommended by IHE have a high probability of a quick uptake in the medical market [Eichelberg et al., 2005].

IHE is organized by clinical and operational domains. For each domain, integration and information sharing preferences are defined. The aim of each IHE domain is to promote the implementation of standard-based interoperability solutions in its specific area, to improve information sharing, workflow, and patient care. Following active IHE domains are available: Cardiology, Eye Care, IT Infrastructure, Laboratory, Patient Care Coordination, Patient Care Devices, and Radiology.

4.1.2.2 IHE IT Infrastructure Technical Framework

IHE IT Infrastructure Technical Framework (ITI TF) is an ongoing expanded document which defines specific implementations of established standards, in order to support optimal patient care through beneficial sharing of medical information. This document illustrates IHE functionality showing the transactions organized into functional units called integration profiles [IHE, 2007a].
IT Infrastructure domain consists of following integration profiles:

- Retrieve Information and Display (RID)
- Enterprise User Authentication (EUA)
- Patient Identifier Cross-referencing (PIX)
- Patient Synchronized Applications (PSA)
- Consistent Time (CT)
- Patient Demographics Query (PDQ)
- Audit Train and Note Authentication (ATNA)
- Personal White Pages (PWP)
- Cross-Enterprise Document Sharing (XDS)

Last integration profile mentioned (XDS) specifies how to manage and share electronic clinical documents that participating healthcare providers are willing to share. Since this corresponds to our research issue, we will describe it in more detail in the next subchapter.

4.1.2.3 Cross-Enterprise Document Sharing (XDS)

The task of Cross-Enterprise Document Sharing (XDS) integration profile is to facilitate registration, storing, and sharing of healthcare documents across healthcare enterprises. Thereby, XDS is not concerned with the content of the documents – it handles standard or formatted text as well as images or structured clinical information.

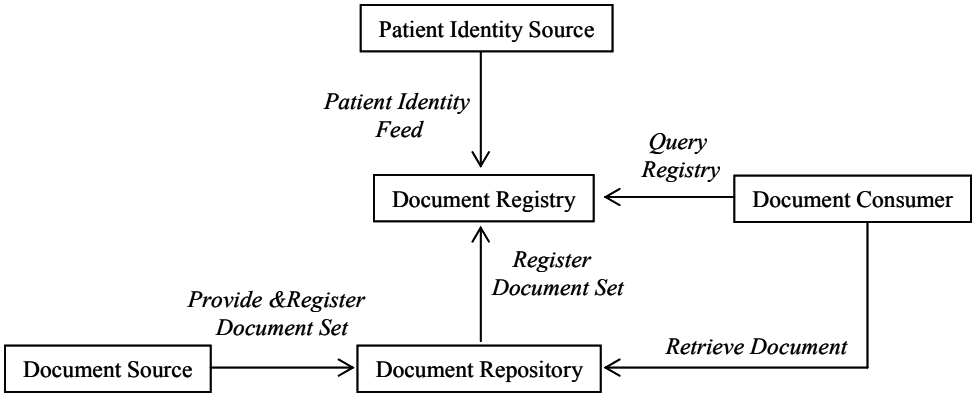


Figure 15: Cross-Enterprise Document Sharing Diagram [IHE, 2006]

XDS acts on the assumption that healthcare enterprises participating in the document exchange network build a group called Clinical Affinity Domain. A Clinical Affinity Domain is a group of healthcare enterprises that have agreed to work together using a common set of policies and share a common infrastructure [IHE, 2006]. The polices include specification of patient identifiers, definition of document format, structure, and content, agreement on access control as well as assignment of metadata representation of the clinical documentation. These are necessary prerequisites for ensuring interoperability between heterogeneous healthcare information systems.

Actors involved with XDS and the transactions between them are depicted in Figure 15. [IHE, 2006] describes the diagram as follows:

Document Source is the healthcare institution participating in the document sharing network. Document source provides the documents to be shared and the corresponding metadata, which is used for document registration.

Document Repository stores the documents to be shared. Since these documents contain highly sensitive data, Document Repository can not be queried. Users (*Document Consumers*) need to know document's unique ID (stored as metadata in the Document Registry) in order to retrieve the desired document.

Document Registry contains and administers the document metadata and can be queried in order to detect the designated Document Repository.

Patient Identity Source is responsible for assigning patient identifiers.

4.1.2.4 NÖMED WAN Patientenindex Project

The goal of *NÖMED WAN Patientenindex* project is to provide direct access to clinical documents at the local document repositories by using standardized IHE XDS integration profiles [Krenn, 2006].

Patient Index and Medical History

Since very heterogeneous information systems and patient identification mechanisms are in use at participating institutions, the unique patient identification is the essential prerequisite for establishing interoperability between various healthcare providers. The aim of this project was not to rebuild the existing systems, but to create a superstructure which would allow the hospitals to keep working with their existing, proven structures. At the same time, this construct would support communication and message exchange between the heterogeneous participating healthcare information systems.

As presented in Figure 16, the project consists of two phases [Schanner, 2006]:

1. Patient Master Index (MPI)
2. Electronic Patient Record – Index (EPA-I)

The hospitals deliver their patient identification numbers to the central systems. Here, a unique patient identifier (MPI) is assigned to each patient. The MPI is confidentially stored in the central system and remains invisible to the originating hospital.

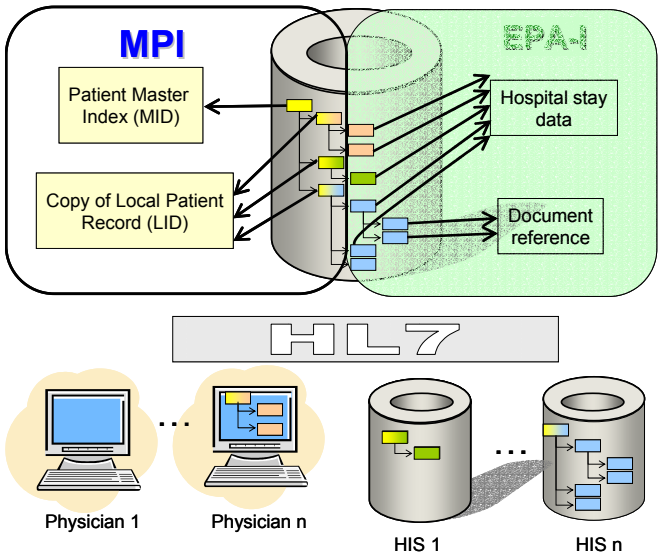


Figure 16: Patient- and Medical History Index: Overview – Data Management [Schanner, 2006]

The Electronic Patient Record – Index identifies the collection of patient's clinical documents. This is the history of all patient's interactions with the participating healthcare providers. It includes medical summaries, ambulatory treatments, clinical, and lab findings, X-ray images, medications etc. Hospitals, as well as the private operating physicians, have access to the electronic patient record. When a patient is admitted to the hospital or visits the physician, he (she) is identified by his (her) e-Card. When the patient is identified and the access rights of the physician in charge are verified, the physician is eligible to get insight into patient's clinical documents.

This project has been implemented by T-Systems company [T-Systems, 2007]. It started in March 2005 and had its first phase (pilot project) completed by the end of the year. The goal of the pilot project was to connect five hospitals and a few private medical practices into a healthcare network for exchange of patients' clinical documents. In the subsequent phase, which started in summer 2006, remaining hospitals and private practices joined the network.

The experience gained in this project in Lower Austria will be used for nation-wide connection of care providers.

Cross-Enterprise Document Sharing in NÖMED WAN Patientenindex

To achieve the goal of the NÖMED WAN Patientenindex project - to facilitate clinical document sharing among the healthcare providers in Lower Austria - the following two requirements are placed upon the clinical documentation: change over to online availability of all relevant patient's data and creation of a life time patient's health record.

Figure 17 shows the separation of the XDS-related activities caused by these requirements [Schanner, 2006]:

- MAKE = Creation of documents
- STORE = Storage of documents
- SEARCH = Retrieval of documents or particular contents

- VIEW = Presentation of documents

This kind of separation in radiology is due to DICOM standard, an already established way of document handling. Due to IHE model, it can be applied to all healthcare areas.

Cross-enterprise clinical document sharing, as represented in Figure 17, has been applied in Lower Austria. Clinical documents are created (MAKE) in the hospitals which are participating in the document exchange network. Document structure and format are created according to the clinical affinity domain specification. Only pure clinical documents (clinical referrals, hospitalisations, admission, transfer, and discharge) are created here, with no additional medical data, which might by interesting for further statistical analyses.

Clinical documents that are released to be shared among other healthcare providers are stored in document repositories (STORE). Following occurrences of document repositories are possible:

- Per organisation (hospital → hospital group → association of doctors with private practice)
- Per region (state → province → region)
- Per data category (discharge letter → images → medical findings)

MAKE and STORE activities are performed by document producers (hospitals), while SEARCH and VIEW are executed by document users (hospitals and private practices).

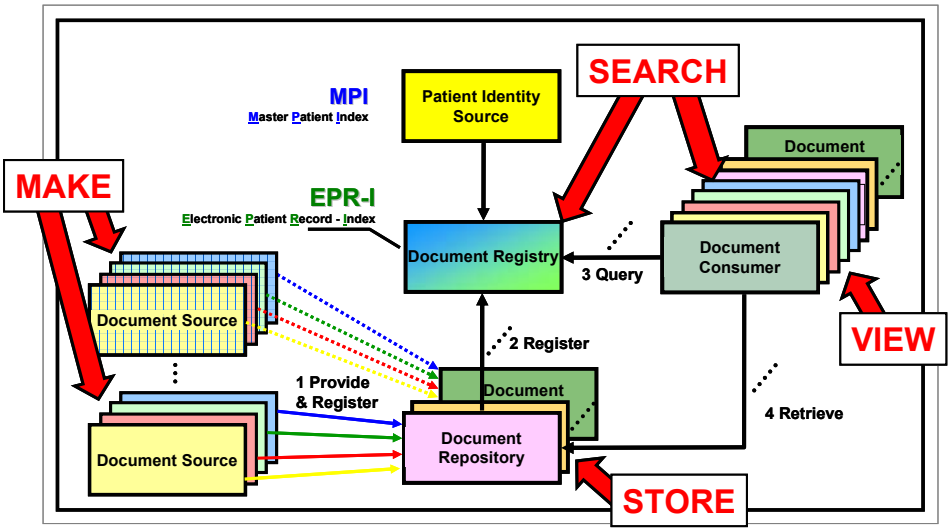


Figure 17: XDS in „NÖMED WAN Patientenindex" Project [Schanner, 2006]

In order to retrieve a document (SEARCH), document consumer has to know document identification number (EPA-I) and patient master index (MPI). Social security number (SVNR) was a natural but not the best candidate for MPI because it can be ambiguous. Moreover, it represents a security threat since it can be misused to reveal personal, financial, political etc. data of a patient. MPI and EPA-I are created centrally, by the trusted third party, and stored in metadata files residing in document registries.

After the document consumer has identified the desired document, it can be viewed or downloaded (VIEW). In this project, one new model of data retrieval is currently being considered: document

consumers subscribe to desired documents (i.e. a physician subscribes to all clinical data related to his (her) patient). As soon as such documents are registered in one document registry, subscribers are notified and given a link to the particular document ("push" instead of "pull" mechanism).

4.1.2.5 The Assessment of This Approach

The IHE-based approach, XDS integration profile in particular, enables seamless data exchange beyond clinical and local healthcare boundaries. It guarantees mobility of the patients, especially chronic disease patients. Through improved patient identification and patient demographics data management by the central patient index system, better efficiency can be achieved in the patient admittance process. As we have seen in the example of our case study project, this approach is capable of improving the quality of care through efficient access to patient's medical history.

The privacy of sensitive patient data is achieved through the adoption of a role-based access model, patient self-determination about his (her) clinical document, and the creation of a unique patient identifier which remains centrally stored and hidden from originating information systems.

When analysing the NÖMED WAN Patientenindex project, two main disadvantages become obvious:

(1) Data stored in the document repositories are not suitable for querying, and no statistical analysis can be run on this data. Medical records, lab tests etc. are still stored as static documents (.pdf files) and can be helpful to the practitioner to re-assemble patient medical history. Still, such information does not offer appropriate foundation for building a sophisticated healthcare decision support system.

(2) The value of evidence-based medical knowledge can not be asserted in this environment. In order to do so, patient electronic health record would have to be accessible at a grain level of aggregation. Through wider adoption of new releases of the healthcare standards (i.e. HL7) in the future years, this approach could move forward in the direction of direct data access, which would open new perspectives for processing of shared patient data.

4.1.3 Grid Technology in Healthcare Environment

The goal of the grid computing is to provide information communication technology for sharing computer power, data storage, software, and tools over the Internet. In recent years, increased interest in deploying grid computing in medical environment, especially in the area of bioinformatics, is noticeable. In this subchapter, we will explain the concept of application of Grid technology in healthcare domain on the example of BIOPATTERN Grid project.

4.1.3.1 The BIOPATTERN Grid Project

The BIOPATTERN Grid is designed to facilitate seamless sharing of geographically distributed bio profile databases and to support the analysis of bioprofiles to combat major diseases such as brain

diseases and cancer within a major EU-funded research project, BIOPATTERN [BIOPATTERN, 2007]. The aim of this project is to facilitate a Europe-wide, coherent, and intelligent analysis of a citizen's bioprofile and to enable remote access to it for the caregivers and the patients. The utilisation of bioprofiles can be beneficial in fighting major diseases, such as cancer or dementia.

A bioprofile is a personal "fingerprint" that fuses together the person's current and past medical history, biopattern, and prognosis. It combines data, analysis, and predictions of possible susceptibility to diseases. The joint activities of the project include making information from distributed databases available in a secure way over the Internet, and providing on-line algorithms, libraries, and processing facilities [Sun et al., 2006].

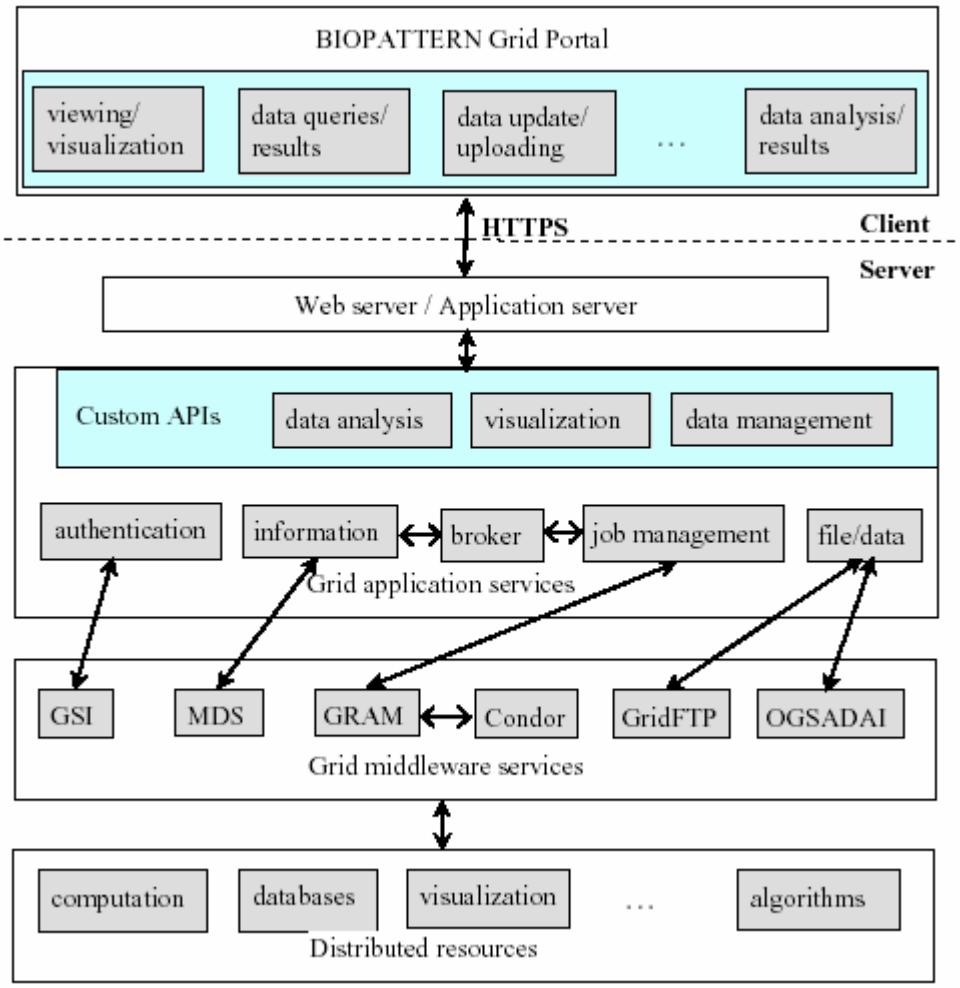


Figure 18: BIOPATTERN Grid Architecture [Sun et al., 2006]

4.1.3.2 Architecture of BIOPATTERN GRID

We describe the architecture of the BIOPATTERN Grid which consists of 5 layers, shown in Figure 18 as it is described in [Sun et al., 2006]:

The Client Application Layer provides the Grid portal access to the end user via a web browser.

The Web Server/Application Server Layer contains web servers and related components and is responsible for authentication of end-user information and for establishing connections to the lower layer. This layer handles incoming HTTP(S) requests and forwards them to the lower layer. After the requests have been processed, it retrieves the results and sends them back to the end-users.

The Grid Application Services Layer transforms the higher layer requests into detailed action requests (e.g. job creation or file transfer), according to the predefined rules for grid services, and sends them for execution. Further, this layer is responsible for efficient management of grid resources and jobs. Each of its components provides different functionality:

- *Authentication service* checks the users' credentials.
- *Information service* translates the resource information obtained from the grid resource monitoring tools and passes them on for various purposes.
- *Broker* is in charge of finding appropriate resources based on the job requirements and the resource information provided by Information Services, and it generates jobs together with job management services.
- *File and data services* manage file and data transfer.

The Grid Middleware Services Layer accepts the request coming form the Grid Application Layer and transforms them into real actions. The central component of this layer is GT4, which consists of Web Service mechanisms for building distributed systems. The main elements of GT4 are:

- *The Grid Security Infrastructure* (GSI), which enables secure authentication and communication over the open network.
- *GridFTP*, which offers high performance, secure, reliable data transfer.
- *The WS-Grid Grid Resource Allocation Management* (WS-GRAM), which is in charge of submitting, monitoring, and cancelling of jobs on grid resources.
- *Condor*, which deals with job queuing mechanism, scheduling policy, local priority scheme, and local resource management.
- *Open Grid Service Architecture Data Access and Integration* (OGSADAI), which provides generic grid data services for access and integration of data stored in relational databases and XML repositories.
- *WS-Monitoring and Discovery System* (WSMDS), which offers services to monitor and discover resources and services on Grids.

The Distributed Resources Layer contains computational resources (e.g. CPUs), data resources (e.g. relational databases), and knowledge resources (e.g. software implementations of scientific algorithms).

4.1.3.3 Application of Bioprofiling over Grid for Early Detection of Dementia

The described grid architecture is applied in this project for the prototype used for early detection of dementia. *Dementia* is the progressive decline in cognitive function due to damage or disease in the brain beyond what might be expected from normal aging [Wikipedia, 2007f]. One acceptable and affordable method for the early detection of dementia is Electronencephalogram (EEG), which is a non-invasive, real time depiction of electrical activities in the brain [Sun et al., 2006].

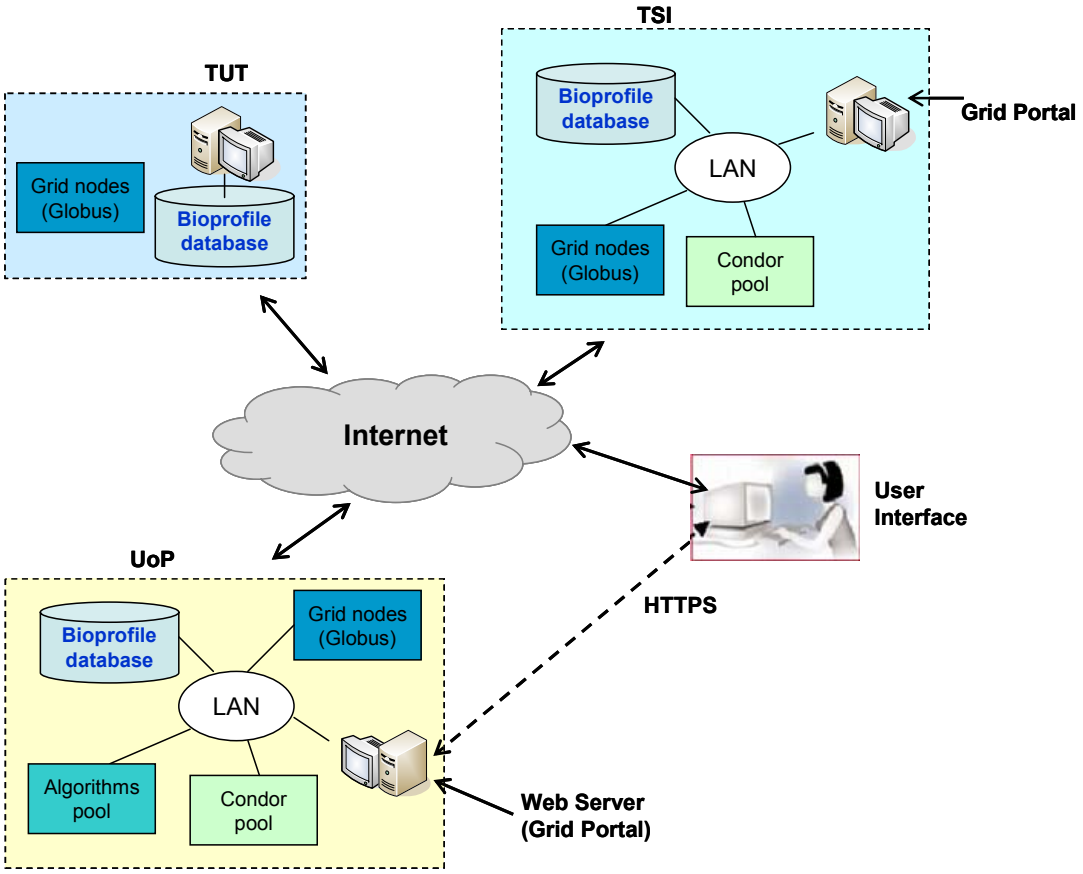


Figure 19: BIOPATTERN Grid Prototype for Early Detection of Dementia [Sun et al., 2006]

The idea of the *individualized care* is that patient's bioprofile is used for analysis and diagnosis, by comparing patient's current and previous condition in order to discover trends. In this case study, the authors argue that this method is more effective than compairing patient's condition to what is generally normal within the population. For this purpose, the BIOPATTERN Grid is used in order to analyse patient's EEG stored in heterogeneous databases, distributed over the network. The clinical data resources include patient information, algorithms, EEGs, and related meta-data.

Physicians are accessing the prototype through a Grid Portal and are able to query, update, upload, and analyse data (Figure 19). After the user has submitted the query containing patient identification, the EEG analysis service locates the distributed EEG files which contain data about the specified patient. Further, the user can select the EEG analysis algorithm from the grid for the analysis of the EEG data. After the analysis jobs have been performed, the results are presented to the user.

4.1.3.4 The Assessment of This Approach

The grid approach for interoperability in healthcare area, which was illustrated in this subchapter on the example of the BIOPATTERN grid project, demonstrates the benefits that can be achieved using throughput computing needed by today's medical analysis tasks. Still, there are some open technological, ethical, regulatory, security, and privacy issues. This approach mostly lacks capabilities needed for communication and knowledge sharing. Strong privacy and Quality of Service constraints at the comprehensive level need to be defined in order to enable seamless exchange of sensitive patient data between heterogeneous medical systems.
In summary, the main benefit of this approach is the performance increase in case of analysing distributed medical data whereas system complexity and costs represent its main disadvantages.

4.1.4 Centralized Summary Care Record Approach

In England, National Health Service (NHS) has created the Connecting for Health (CfH) agency [NHS, 2007], whose role is to develop an integrated, patient-centred IT system. This healthcare system will connect over 100.000 doctors, 380.000 nurses, and 50.000 other health professionals and give patients access to their personal health and care information [Spronk, 2007]. One of the CfH's programs is *the Spine*, which is related to creation of an electronic care record for more than 50 million of England's patients.

4.1.4.1 The Spine

The Spine is a national, central database, where summary patient records are stored. When fully implemented, the summary patient records will be automatically uploaded with relevant medical information coming from local records [NHS, 2005].
According to the NHS Care Record Service plan, each patient will have two care records:

- *full local record*, stored on the care giver site,
- *summary record*, stored on the Spine.

The summary record will provide the accredited care givers across the country with important demographic data about the patient and his (her) medical history (medications, treatments received, lab test results etc.). Detailed information, held by the local records, will be linked to the summary record, so that the care givers will be able to drill down to this data, when needed.
Apart of the supporting the practitioners at the point of care, anonymized and pseudo-anonymized summarized care record can be pulled for secondary uses services, such as research, business reports, and planning.

4.1.4.2 The Spine Architecture

The architecture of the Spine is based on a centralized summarized care record, supported by directory services and HL7 V3 messaging. Figure 20 represents its structure.

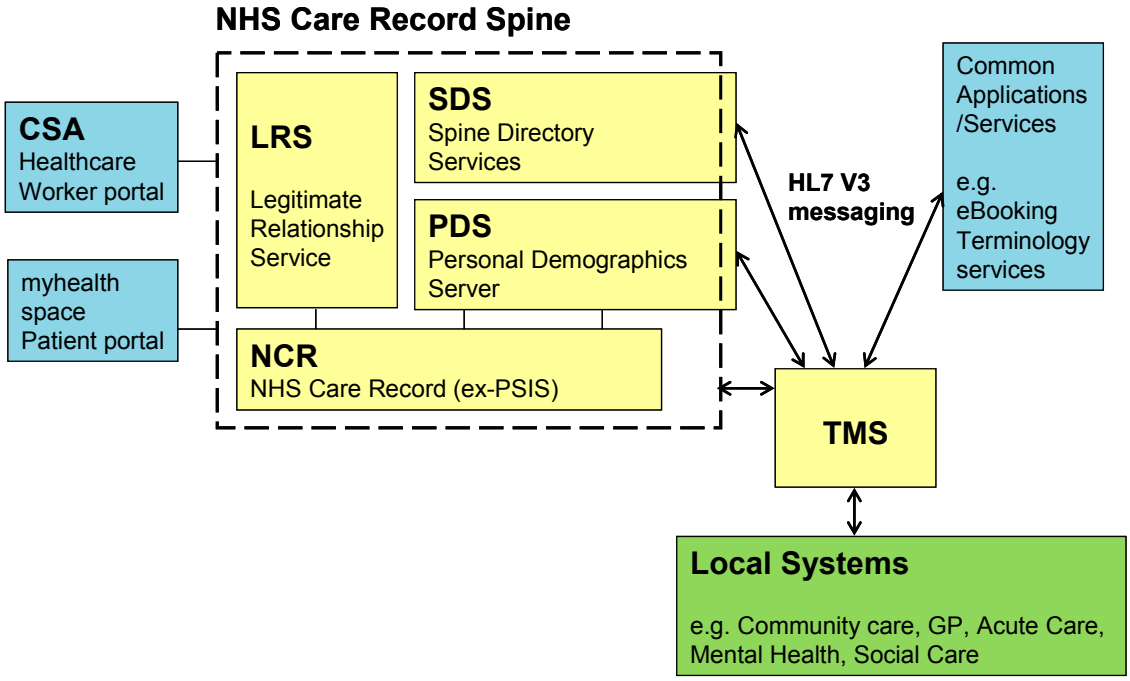


Figure 20: The Spine Architecture [Spronk, 2007]

The major components of the Spine are [Spronk, 2007]:

- **Personal Demographics Services (PDS)**

  The PDS is going to be the central source of NHS patient demographic information, such as name, data of birth, NHS number, address etc. The patient identifiers, which are currently not unique and are created and maintained locally by diverse local and national organizations, will be migrated. In order to protect the patient demographic data, the role-based access mechanism of National Care Record will be used to control PDS queries. When anonymized and pseudo-anonymized for security reasons, data stored in this database will be available to the researches, business, and governmental users, for studies, analysing and planning activities.

- **National Care Record (NCR)**

  NCR is a central repository for summarized clinical records. A Summarized Care Record receives accurate medical information from various local systems where the patient receives healthcare. It is aware of the existence of detailed information, which is stored in the Detailed Care Record, in a provider's local system. This means that when a user is querying for the specific information, the NCR will be checked first. If the searched data is not stored there, the

information will be obtained from one of the local healthcare systems, which is referenced by the NCR. Note: NCR was formerly known as NHS Care Record, the Personal Spine Information Service (PSIS) and Integrated Care Record Service (ICRS).

- **Legitimate Relationship Service (LRS)**
  In NHS, the patient's consent is required for care professionals to view their NHS Care Record. The LRS controls the access rights on a patient's clinical data. The LRS will provide a single log-in and a record of each healthcare professional accessing a patient's NHS Care Record. All information will be provided on a need-to-know basis and based on user's role [NHS, 2005].
- **Spine Directory Services (SDS)**
  SDS provides various Directory Services (e.g. organisational details of GP practices). It contains Spine User Directory and Spine Accredited Systems, which ensure that transactions and messages are only performed by the legitimate users and is therefore essential for providing the security of the Spine.
- **Transaction and Messaging Spine (TMS)**
  TMS is responsible for message routing between local systems and the NCR. All messages transmitted via TMS will be based on HL7 V3 messaging standard.

4.1.4.3 The applications attached to the Spine

Applications can be attached to the Spine via TMS or by other means. Applications attached to the Spine by means of the exchange of messages via TMS [NHS, 2005] are:

- **Common applications & services:** Services offered at a national/regional level to all healthcare providers, e.g. the support for e-Booking (provides a mechanism for GPs and other primary care staff to register a request for a service for a patient and either book an appointment there and then or allow the patient to book a suitable appointment themselves at a later date through a number of routes) and e-Prescription.
- **Local systems:** Systems used by healthcare providers that are *Spine-enabled*, i.e. linked to the Spine. Examples include GP, Acute, Community, Mental health, and Social Care IT systems.

Clinical Spine Applications Service (CSA) and *myhealthspace* are not part of the Spine and are attached to it by means other than TMS [NHS, 2005]:

- **Clinical Spine Applications Service (CSA):** A web front-end that allows healthcare professionals to have access to the NCR/PDS data if their system isn't Spine-enabled (yet). The CSA can't be used to communicate with Local systems or other Common Applications & Services.

- **Myhealthspace:** A web front-end that allows the patient to have access to their own NCR/PDS data. Myhealthspace can not be used to communicate with Local systems or other Common Applications & Services.

4.1.4.4 The Assessment of this Approach

The described project appears to be relying on a fairly mature architecture. It aims to provide a commonly accessible patient care record by interconnecting local health providers and by centrally storing summarized patient medical information. By keeping demographic and medical data separated and by providing researchers and high-level users with the anonymized and pseudo-anonymized patient data, it ensures the security of access to national and local systems. By adopting the international standard for exchange of health messages (HL7), this British healthcare information network will be open for future international cooperation projects. For the time being, no statistical processing of patient data or incorporating of evidence-based rules to support clinical decision making is planned.

The described project is a long lasting undertaking. It started in 2006 and is expected to be completed in 2010. According to the [AERZTE_ZEITUNG, 2005], the cost of this project will amount to about six billion pounds (more than nine billion euro). Apart from high costs, one of the major obstacles to the adoption of this approach in other countries is the fact that the privacy laws and decentralized management of healthcare in general do not meet the necessary preconditions for such an implementation.

4.2 Opportunities for the Federated DWH Approach in the Field of EBM

After review of the work that has been done in this area, the opportunities for DWH approach present themselves as follows:

- DWH, OLAP tools, and data mining technology, provide the best support for development of evidence-based guidelines.
- DWH offers clear and feasible methodology for nation-wide acquisition, transformation, and load of heterogeneous and complex medical data, with the goal of building a life-time electronic patient record, accessible to the wide range of users.
- Adding of evidence-based guidelines to the DWH, which contains all available data about the patient health history, opens new horizons for enhancing the quality of care.
- Leaving sensitive data at their origin and uploading them on demand makes federated DWH the preferable and faster accepted consolidation approach in the healthcare industry.

- Federated DWH solution strengthened by depersonalization, pseudonymization, and role-based access mechanisms offers the necessary level of protection of the patient data privacy.

4.2.1 4 Rules for Creation of Sustainable, Federated DWH – based DSS for EBM

In the next four chapters, we will describe the prerequisites for the creation of a federated DWH, which will support the interoperability among various heterogeneous data sources for the purposes of EBM and for facilitating decision support in the healthcare environment.
First of all, we formulate the four basic rules for the creation of a sustainable federated DWH for support of evidence based medicine[†]:

Rule 1: The local autonomy rule:

Each component warehouse must operate independently of the federation.

Existence of a federation must not have impact on local users of the component warehouse. The sharing process generally includes only a part of the component warehouse data and is under selective control of the local administrators. This issue is handled in chapter 5.

Rule 2: The generalisation and portability rule:

The highest level of generalisation and portability for the federated DWH must be achieved by the deployment of an international standard that covers the required areas of healthcare.

In our approach, we adopt the Health Level Seven [HL7, 2007] standard protocol for data exchange, which defines format and content of messages that healthcare applications must use when exchanging data with each another. It is widely independent from underlying hardware and network infrastructure as well as database and applications used. Chapter 6 deals with this subject.

[†] The creation of these rules was motivated by the "Codd's 12 rules for defining a fully relational database" [Codd, 1985].

Rule 3: The physical and logical data independency rule:

Changes to the physical level (how the data is stored) and changes to the logical level (logical data models, tables, relations) must not require changes to an application.

Due to the wrapper-mediator architecture, any changes on the local physical and logical DWH models can be bypassed. Rule 3 will be addressed in more detail in chapter 7.

Rule 4: The guaranteed security rule:

Patient privacy must be ensured through the evidence-based medicine specific security measures: depersonalization, pseudonymization, and role based access control.

The main ethical concern of federated DWH for evidence-based purposes is to protect the extraordinarily sensitive nature of health data while using clinical data sources for knowledge discovery and development of evidence-based guidelines. Protection of confidential patient data is subject of (inter)national regulations. Issues of privacy, software regulation, and ethical and legal aspects of data processing for the purposes of EBM are discussed in chapter 8.

Part II

The Federated DWH as Enabler for Evidence-Based Healthcare

5 Rule 1: Federated DWH for Consolidation of Healthcare Information Systems

As the need for collaboration in the healthcare domain grows, the interoperability of medical information systems becomes increasingly important. Consolidation of distributed patient data and building of a sustainable e-health system facilitating EBM is a major research challenge. In order to overcome institutional barriers and competencies for changes across sectors, standardized data structures and corresponding underlying infrastructure are needed. The goal of this chapter is to point out the advantages of virtual integration of heterogeneous medical information systems using federated data warehouse approach and abandoning the building of large physical data repositories.

When several independent medical institutions share their data for mutual purposes, they may not allow any physical copies of their data to be created in any system that is out of their full control. They may also restrict access to some data and demand to maintain full control of data access. Highly confidential healthcare records are a typical example of such data. In such situations a federated data warehouses approach offers the satisfactory solution.

The relevance and use of the federated DWH is obvious within the domain of epidemiology - for example for surveillance and tackling of epidemic diseases or in the area of evidence-based medicine. Discovering evidence-based rules requires processing of detailed data that correspond to the most detailed grain level in the component warehouses, so that all data from the components must be copied into the joint warehouse. Still, there are a lot of technical and organizational issues to be discussed.

In this chapter, we present a secure federation model and we use the health insurance example to explain the 5-level federation architecture.

5.1 Challenges and Benefits of Integration of Healthcare Information Systems

The vast majority of patients receive their healthcare from multiple healthcare providers. Hospitals, physicians, recovery centres, laboratories, pharmacies, and health insurance institutions each have their own, isolated patient records. Therefore, fragmental knowledge about the patient's health condition is stored at different sites. The absence of integrated healthcare carries the risks of medical treatment errors, duplicate examinations, lack of coordination, and increased therapy costs. In the past few years, the perception of the need for the integrated healthcare system has risen in both the provider and consumer sector. The sharing of medical data enables hospitals and single work-site physicians to provide more efficient patient care, which results in cost reduction for social insurance institutions. Additionally, patients are in many cases free to consume healthcare from different providers.

Many projects aiming at the creation of the electronic heath care record have been started all over the world. They all intend to integrate the entire available patient medical data into one single record, which would be available to the authorized physicians or medical institutions when they need it. In our federated DWH approach, we go one step further and investigate the possibility of creating a large knowledge base through joining and processing of the most detailed grain level data in the component warehouses. This knowledge base would be used in the area of evidence-based medicine, and to support public sector in preparing for the demographic change in health and welfare.

Consolidated healthcare records would serve multiple users. They are sent to the insurers to justify payment for medical services rendered and for fraud detection. They are used for quality reviews, administrative reviews, and utilization studies to manage business aspects of healthcare. Further, they are used for societal purposes, such as medical research, health management, social services and welfare system management, law enforcement, and determining life insurance eligibility.

Sharing of sensitive healthcare data implies security concerns. Although different means of protecting confidential data exist, medical institutions generally reject the idea of creating any physical copies of their data and prefer a virtual integration, as proposed by our federated DWH model. Legislations aimed at protecting personal data, like HIPPA in the USA [HIPPA, 1996], PIPEDA in Canada [PIPEDA, 2006] or EU Data Protection directive [EU, 1995], which came into effect during the last decade, give additional support to such a decision. We will offer a detailed description of privacy assurance mechanisms for healthcare data in the DWH federation in chapter 8.

5.2 Federated Data Warehouses

There are two architectural approaches for consolidation of data warehouses:

1. Traditional data warehouse approach
2. Federated data warehouse approach

The degree of integration depends on the users' needs and privacy policies applied to each warehouse. The chosen approach is setting up the abilities and the limitations of a data warehouse.

5.2.1 Traditional DWH Approach

A central (global) warehouse is generally accessed through OLAP tools and contains mostly only summarized data from the component DWHs. If detailed data are needed, the component warehouses, which contain both detailed and summarized data, must be accessed. Such a warehouse may be successfully used in a single corporation with several regional DWHs. In case of merging DWHs of different companies, or any serious data mining task, all detailed data should be copied into the central warehouse. That is a very highly priced undertaking.

5.2.2 Federated DWH Approach

According to Seth and Larson [Sheth and Larson, 1990] a federated database is a collection of cooperating database systems that are autonomous and possibly heterogeneous. A federated DWH is the integration of heterogeneous business intelligence systems set to provide analytical capabilities across different functions of an organization. It aims to integrate the key business metrics, measures, and dimensions. A federated DWH is a functional DWH, a “big umbrella”. No central, large DWH that collects data from smaller component warehouses is created: heterogeneous DWHs are functionally integrated into a single unit from the conceptual point of view using a unique common conceptual model. [Jindal and Acharya, 2004] state that a “common business model” (i.e. common conceptual model) is needed, which defines common dimensions. For example, federated DWHs are used in case of mergers and acquisitions when merging corporations that already possess their DWH.
Each component unit keeps working autonomously and the users are not aware of the existence of the federation. Since only a “single version of truth” and a unique interpretation of the joined data should exist, just a singular federation schema is needed.
Figure 21 represents the schema architecture for federated DWHs, following the “five-level schema architecture of an FDBS”, as described by [Sheth and Larson, 1990]:
Local schema of the underlying DWH is a conceptual schema of this component. Since it is expressed in the native data model, heterogeneous local schemas can be expected on this level.
Component schemas describe the different local schemas using a unique representation. Furthermore, semantics that are missing in a local schema can be added to the component schema. The process of schema translation from a local schema to a component schema generates the mapping between component schema objects and local schema objects. During the semantic integration these mappings are used to transform requests on a component schema into requests on the corresponding local schema. This principle supports the heterogeneity feature of the federated DWHs.

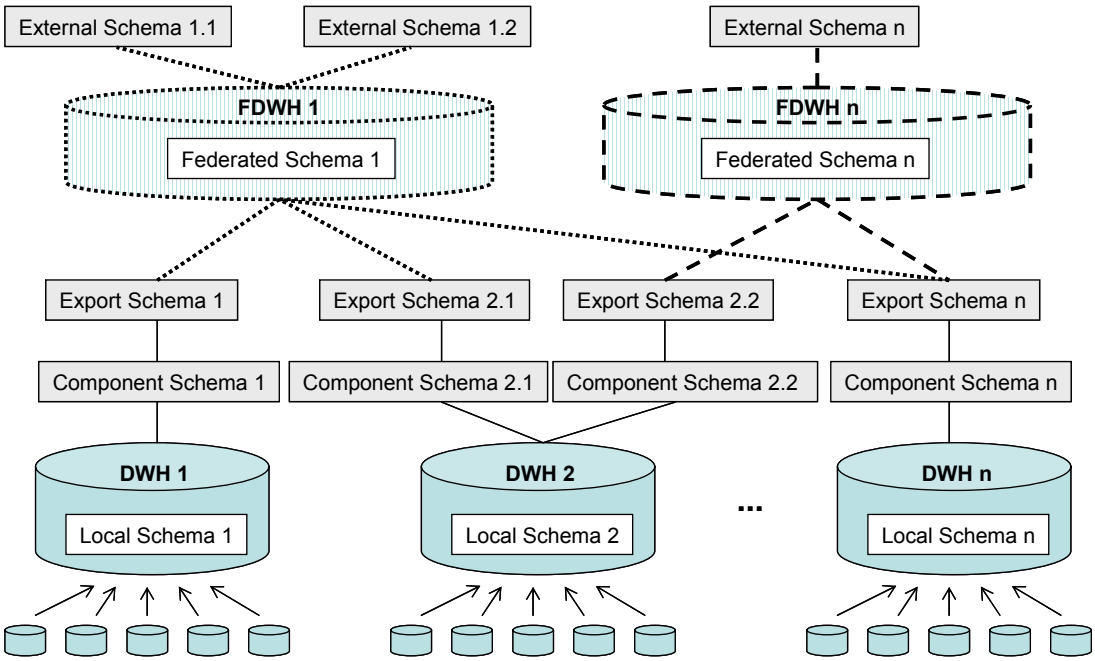


Figure 21: Federated DWH Schema Architecture

Export schema is described as a subset of a component schema that is available to the federation. It may include access control information regarding its use by specific federation users. The purpose of defining export schemas is to facilitate control and management of association autonomy. A filtering processor can be used to provide access control as specified in an export schema by limiting the set of allowable operations that can be submitted on the corresponding component schema.

Federated schema is an integration of multiple export schemas.

External schema is a user specific schema. It allows customization and supports implementation of additional access control and integrity constraints.

The major reasons for choosing a federated DWH architecture are:

- *Cross-organisational requirements*: Since the DWH federation is capable of incorporating all relevant information from the component warehouses, cross-organisational analysis can be easily performed.
- *Flexibility for adding new components:* New DWH components can be easily integrated into the federation, without having to change their existing structure.
- *Easy and cost effective implementation:* Since the federated approach does not attempt to build an additional, new DWH foundation but rather supports integration on a conceptual level, time and costs associated with its creation are far less than in a traditional approach. Furthermore, the federated approach does not influence the operation of the component systems and therefore prevents the major points of conflict in the early stage of agreement.

Table 4 shows the comparison of traditional and federated DWH approach.

	Traditional DWH approach	Federated DWH approach
Query response time	+	-
Availability	+	-
Reliability	+	-
Concurrency	-	+
Organisational effort	-	+
Administrative effort	-	+
Security	-	+
Willingness of local organisations	-	+

Table 4: Comparison of Traditional and Federated DWH Approach

5.3 HEWAF - an Healthcare DWH Federation

In this subchapter, we will use an example from healthcare field, in order to introduce HEWAF (Healthcare Warehouse Federation), an application scenario for the federated DWH for the Austrian health insurance institutions. The purpose of HEWAF is to apply the principles of evidence-based medicine in the care giving and health insurance area, which ultimately leads to new, improved, and more efficient algorithms in clinical practice, as well as cost reduction.

5.3.1 Achieving Federation

The federation schema can be created manually or semi-automatically using neural networks or other heuristic or learning-based methods. In our case, the DWH designer is given a simple map of each DWH (only the data that database administrator wants to reveal) and can make a common model based upon an existing standard health model (it might be an extension of the existing standard). The existence of a standard is especially relevant for enumerated medical data types (even the simplest, like gender, can be written as “F/M” or “0/1”), since they might be encoded differently in each component DB.

Achieving federation is accomplished through the following steps:

1. Collect the goals and requests of federated DWH users.
2. Gather the list of fact tables and dimensions of participating DWHs.
3. Receive the list attributes that component DBAs want to reveal in the federated DWH, the list of protected attributes and their data types.
4. Compare the different component DWH maps.
5. If some attributes of interest to the users of federated DWH exist only in some components, negotiate the solution of the problem and re-arrange the common model (remove the attribute

from DWH, find a default value, give all possible values and also mark the non-existing as N/E)

6. Find an existing common model for electronic patient healthcare record and create possible extensions to this model.

5.3.2 Conceptual Model of HEWAF

Traditional focus of health insurance institutions has been turned to charging inpatient hospital stays (and included therapy), outpatient encounters and therapies, as well as drug prescriptions. Figure 22 illustrates the class diagram representation of the multidimensional model for the warehouse. The basic dimensions in this model are: patient, clinician, clinic, drug and time. Therapy and prescriptions represent facts, with financial attributes as measures. Evidence-based medicine, the main purpose for creating the federation, adds diagnosis (disorders and diseases) as another crucial topic. Diagnosis can be a dimension, if the particular therapy or prescription is the result of a diagnosis (DimDiagnosisDesc in Figure 22). But, if it is observed as the process of asserting a diagnosis, as a part of a patient encounter, than it can be a fact (FactDiagnosesEnc in Figure 22). Such a fact can also be described by patient, clinician, clinic, and time dimension.

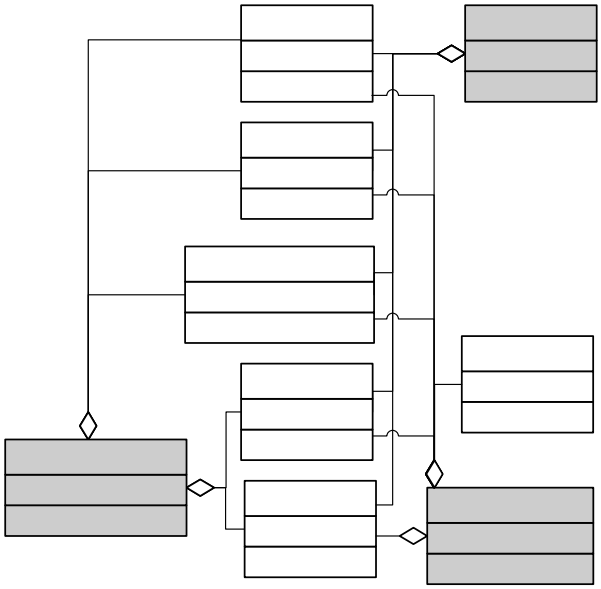


Figure 22: Multidimensional Model of the Sample Social Insurance DWH

HEWAF introduces a prototype common conceptual model for a federation of health insurance DWHs that is based on HL7 RIM model and CDA format. (We will be talking about these standards in more detail in chapter 6). It is a multidimensional conceptual model that does not point to any logical implementation (i.e. any DBMS).

The fact-to-dimension relationship in the multidimensional model is generally many-to-one (a record in the fact table is linked to a single record in each dimension). However, many diagnoses can be asserted during one patient encounter, making the cardinality FactDiagnosisEnc-DimDiagnosisDesc many-to-many (we do not want to split a unique encounter record).

fact/dimension	RIM class	RIM Backbone class	CDA XML tag
patient	Person + Patient	Entity + Role	patientRole
clinician	Person + Employee	Entity + Role	authorRole
clinic	Organisation + LicensedEntity	Entity + Role	healthcareFacility
diagnosis	Observation	Act	observation
prescription	Substance-Administration	Act	substanceAdministration
therapy	Procedure	Act	procedure
encounter	PatientEncounter	Act	encounter
drug	Material+Role (no particular sub-class)	Entity	manufacturedProduct

Table 5: Facts and Dimensions in HEWAF Conceptual Model

Among the many existing multidimensional models we chose the UML-based object-oriented conceptual model for DWHs presented by Trujillo and colleagues [Lujan-Mora et al., 2002] because of its strong formalism and a possibility to map UML fact and dimension classes and their attributes to UML-based RIM classes and their attributes. The model presents facts (grey) as aggregates and dimensions (white) as its constituting parts. It is one of the few conceptual models proposed in the literature that allows fact-to-dimension relationship to be either many-to-one or many-to-many (note the 1..* cardinality on the DimDiagnosisDesc dimension). HEWAF converts the facts and dimensions presented in Figure 22 to RIM classes with CDA notation, as shown in Table 5.

5.4 Running Example: Building a 5-level Schema Architecture of a Federated DWH

In this subchapter, we expand our running example from chapter 2 in order to explain the 5-level schema architecture of a federated DWH. Here, we consider two medical DWHs participating in one federation. We assume that these DWHs are divergent and that their local data models have been created using different representations.

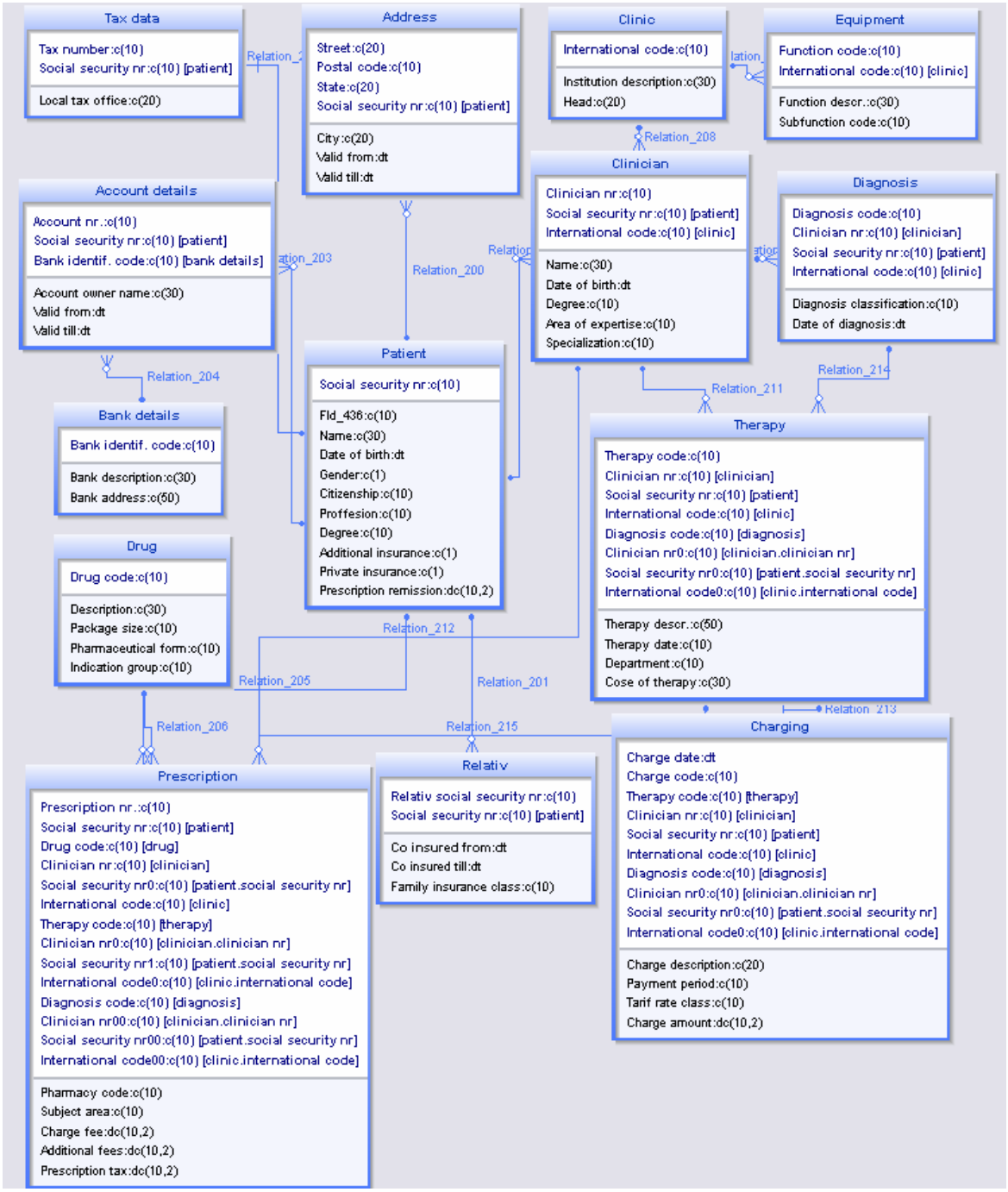


Figure 23: Relational Local Data Model of the DWH1

Local data model of the DWH1 is expressing its relational structure (Figure 23), whereas local data model of the DWH2 (Figure 25) is representing its multidimensional nature. According to the 5-level schema architecture of the federated DWHs illustrated in Figure 21, for each local DWH schema, corresponding component schemas (using a unique representation) need to be defined. Figure 24 depicts the component data model for the DWH1.

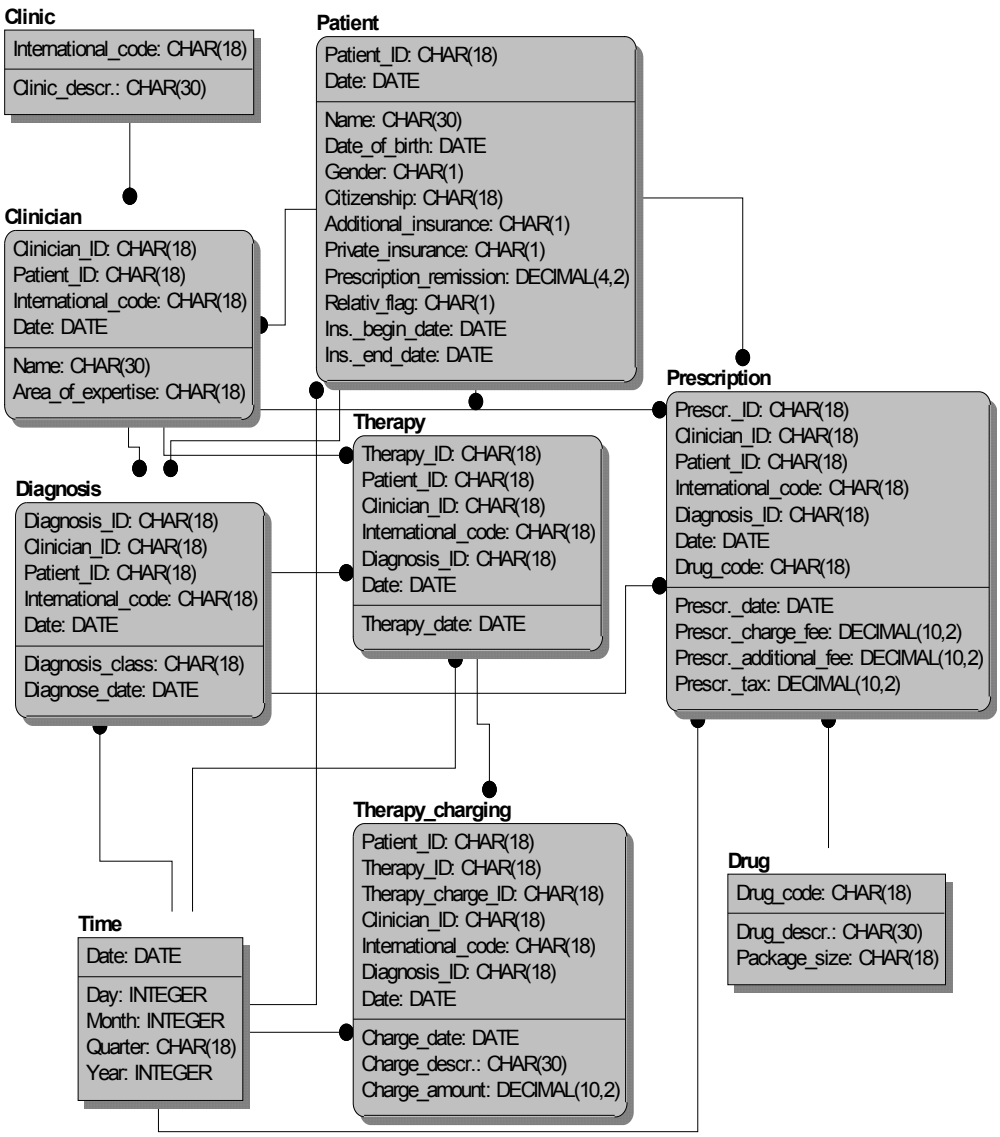


Figure 24: Component Data Model of the DWH1

During data model translation between local and component data models, the following mappings have been created (Table 6):

Component Data Model		Local data model		
Table	**Attribute**	**Table**	**Attribute**	**Condition**
Clinic	International_code	Clinic	International code	
Clinic	Clinic_descr	Clinic	Institution description	
Clinician	Clinician_ID	Clinician	Clinician nr.	
Clinician	Name	Clinician	Name	
Clinician	Area_of_expertise	Clinician	Area_of_expertise	
Diagnosis	Diagnosis_ID	Diagnosis	Diagnosis code	
Diagnosis Diagnosis	Diagnosis_class	Diagnosis	Diagnosis classification	

Diagnosis	Diagnosis_date	Diagnosis	Date of diagnosis	
Time	Date	--	--	
Time	Day	--	--	
Time	Month	--	--	
Time	Quarter	--	--	
Time	Year	--	--	
Patient	Patient_ID	Patient	Social security nr.	
Patient	Name	Patient	Name	
Patient	Date_of_birth	Patient	Date of birth	
Patient	Gender	Patient	Gender	
Patient	Citizenship	Patient	Citizenship	
Patient	Additional_insurance	Patient	Additional_insurance	
Patient	Private_insurance	Patient	Private_insurance	
Patient	Relativ_flag	Relativ		Set to "1" if existing, otherwise set to "0"
Patient	Ins_begin_date	Relativ	Co-insured from	relativ.Social security nr.= patient.Social security nr.
Patient	Ins_end_date	Relativ	Co-insured till	relativ.Social security nr.= patient.Social security nr.
Therapy	Therapy_ID	Therapy	Therapy code	
Therapy	Therapy_date	Therapy	Therapy date	
Therapy_charging	Therapy_charge_ID	Charging	Charge code	
Therapy_charging	Charge_date	Charging	Charge date	
Therapy_charging	Charge_descr.	Charging	Charge description	
Therapy_charging	Charge_amount	Charging	Charge amount	
Prescription	Prescr. ID	Prescription	Prescription nr.	
Prescription	Prescr._date	Therapy	Therapy_date	prescription.therapy_code = therapy.therapy_code
Prescription	Prescr._charge_fee	Prescription	Charge fee	
Prescription	Prescr._additional_fee	Prescription	Additional fees	
Prescription	Prescr._tax	Prescription	Prescription_tax	
Drug	Drug_code	Drug	Drug_code	
Drug	Drug_descr.	Drug	Description	
Drug	Package_size	Drug	Package size	

Table 6: Mappings between Local and Component Data Model of the DHW1

The multidimensional local data model of DWH2, depicted in Figure 25, has been created using ADAPT [Symmetry_Corporation, 2007] notation.

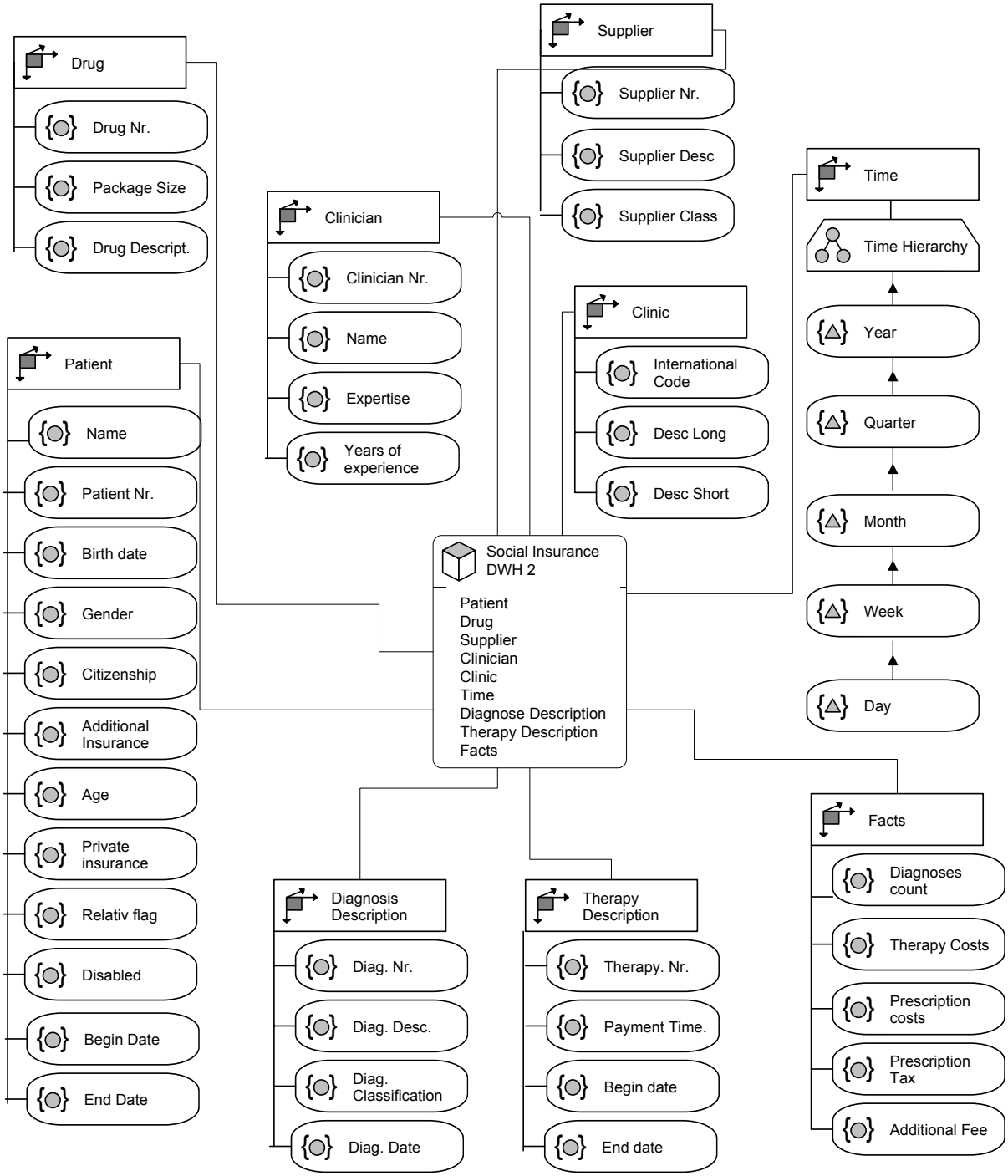


Figure 25: Multidimensional Local Data Model of the DWH2

Corresponding component data model is shown in Figure 26. Both component data models (for DWH1 and DWH2) are created using a common representation.

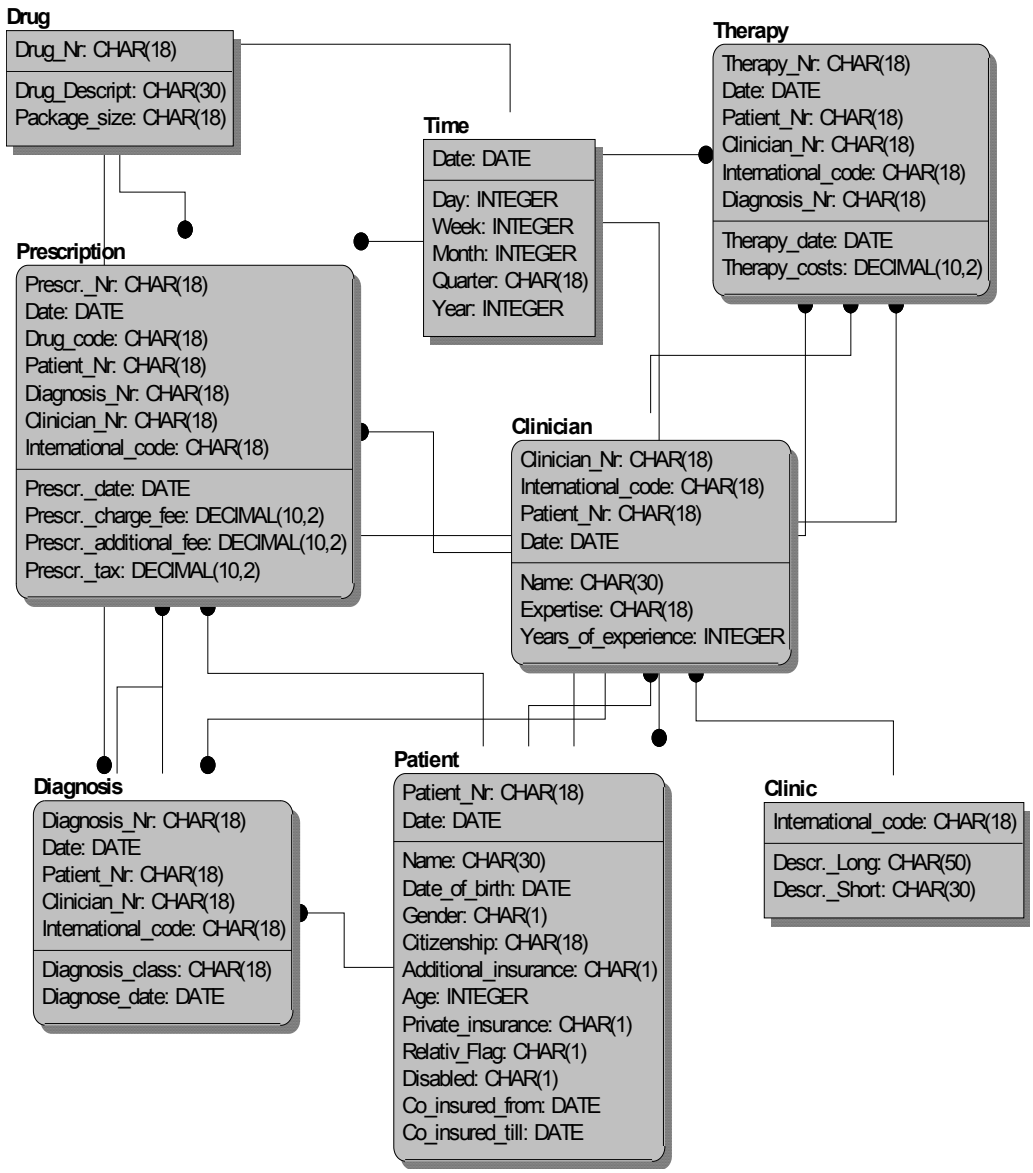


Figure 26: Component Data Model of the DWH2

Mappings between local and component data models for DWH2 are represented in Table 7:

Component Data Model		Local data model		
Table	**Attribute**	**Table**	**Attribute**	**Condition**
Clinic	International_code	Clinic	International code	
Clinic	Descr_Long	Clinic	Descr Long	
Clinic	Descr_Short	Clinic	Descr Short	
Clinician	Clinician_Nr	Clinician	Clinician Nr	
Clinician	Name	Clinician	Name	
Clinician	Expertise	Clinician	Expertise	
Clinician	Years_of_experience	Clinician	Years of experience	
Diagnosis	Diagnosis_Nr	Diagnosis Description	Diagnosis Nr.	
Diagnosis	Diagnosis_class	Diagnosis Description	Diag. Classification	
Diagnosis	Diagnosis_date	Diagnosis Description	Diag. date	
Time	Date	--	--	

Time	Day	--	--	
Time	Month	--	--	
Time	Quarter	--	--	
Time	Year	--	--	
Patient	Patient_Nr	Patient	Patient Nr	
Patient	Name	Patient	Name	
Patient	Date_of_birth	Patient	Birth date	
Patient	Gender	Patient	Gender	
Patient	Citizenship	Patient	Citizenship	
Patient	Additional_insurance	Patient	Additional_insurance	
Patient	Age	Patient	Age	
Patient	Private_insurance	Patient	Private_insurance	
Patient	Relativ_flag	Patient	Relativ flag	
Patient	Disabled	Patient	Disabled	
Patient	Co_insured_from	Patient	Begin date	
Patient	Co_insured_till	Patient	End date	
Therapy	Therapy_Nr	Therapy Description	Therapy code	
Therapy	Therapy_date	Therapy Description	Begin date	
Therapy	Therapy_costs	Facts	Therapy Costs	
Prescription	Prescr. Nr	Drug	Drug Nr.	
Prescription	Prescr._date	Therapy	Therapy_date	prescription. therapy_code = therapy.therapy_code
Prescription	Prescr._charge_fee	Facts	Prescription costs	
Prescription	Prescr._additional_fee	Facts	Additional Fee	
Prescription	Prescr._tax	Facts	Prescription Tax	
Drug	Drug_Nr	Drug	Drug_Nr.	
Drug	Drug_Descript.	Drug	Drug Descript.	
Drug	Package_size	Drug	Package Size	

Table 7: Mappings between Local and Component Data Model of the DHW2

At this stage of our running example, we assume that the component DWH management systems support proper access control security features for their external schemas and that the external schemas of a component DWHs may be used as export schemas in the 5-level architecture of Figure 21. Thus, export schema is equivalent to the component schema. In chapter 8, we will discus the possibilities for access control restriction, which can be applied to component data models through depersonalization and pseudonymization.

The federated data model is represented in Figure 27. This model integrates multiple export data models.

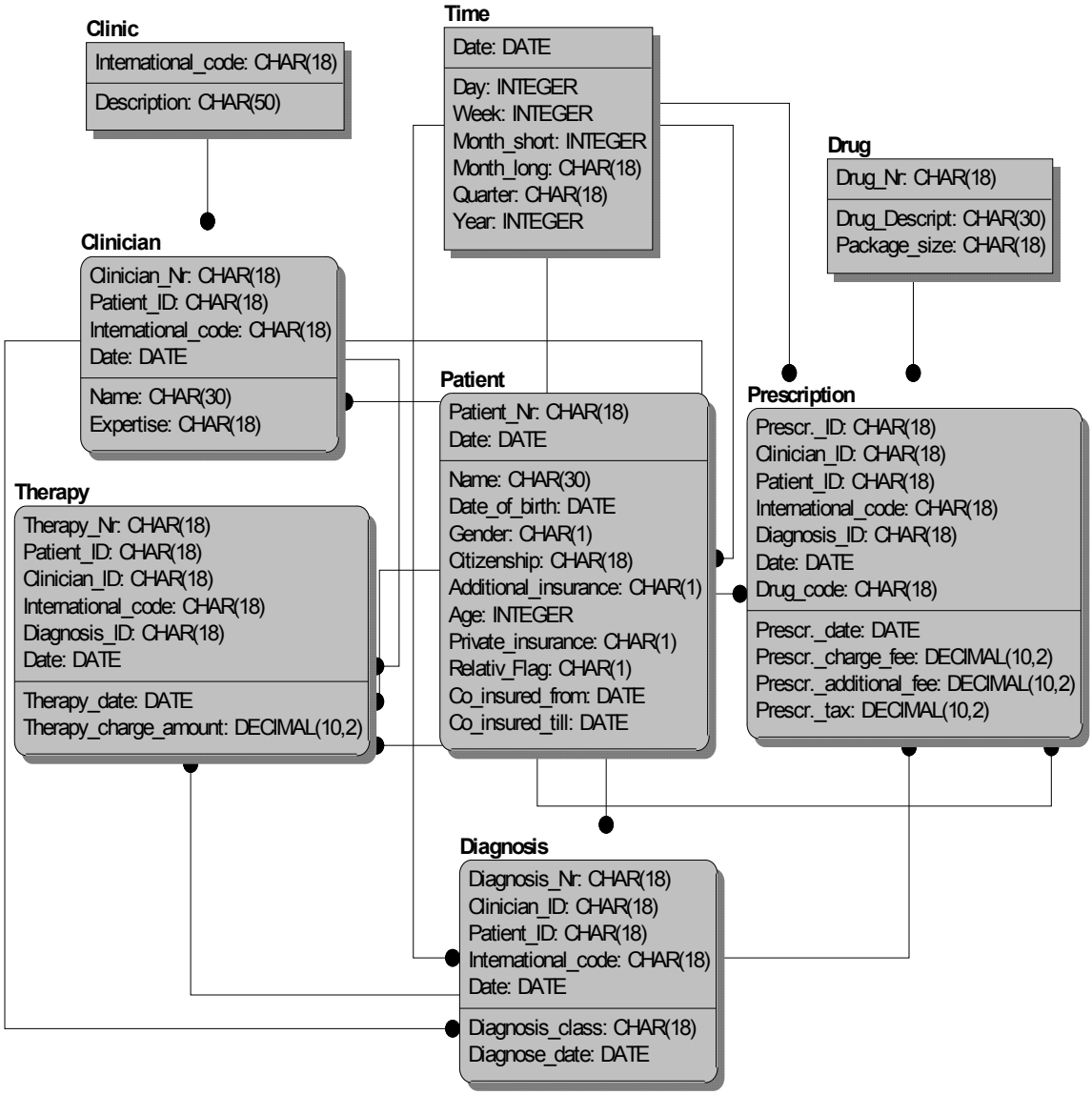


Figure 27: Federated DWH Data Model

Mappings for the federation are listed in Table 8:

Federated Data Model		Component Date Model DWH1		Component Data Model DWH2	
Table	**Attribute**	**Table**	**Attribute**	**Table**	**Attribute**
Clinic	International_code	Clinic	International_code	Clinic	International_code
Clinic	Description	Clinic	Clinic_descr	Clinic	Descr_Long
Clinician	Clinician_Nr	Clinician	Clinician_ID	Clinician	Clinician _Nr
Clinician	Name	Clinician	Name	Clinician	Name
Clinician	Expertise	Clinician	Area_of_expertise	Clinician	Expertise
Diagnosis	Diagnosis_Nr	Diagnosis	Diagnosis_ID	Diagnosis	Diagnosis_Nr
Diagnosis	Diagnosis_class	Diagnosis	Diagnosis_class	Diagnosis	Diagnosis_class
Diagnosis	Diagnosis_date	Diagnosis	Diagnosis_date	Diagnosis	Diagnosis_date
Time	Date	Time	Date	Time	Date
Time	Day	Time	Day	Time	Day
Time	Month_short	Time	Month	Time	Month
Time	Month_long	--	--	--	--

Time	Quarter	Time	Quarter	Time	Quarter
Time	Year	Time	Year	Time	Year
Patient	Patient_Nr	Patient	Patient_ID	Patient	Patient_Nr
Patient	Name	Patient	Name	Patient	Name
Patient	Date_of_birth	Patient	Date_of_birth	Patient	Date_of_birth
Patient	Gender	Patient	Gender	Patient	Gender
Patient	Citizenship	Patient	Citizenship	Patient	Citizenship
Patient	Additional_insurance	Patient	Additional_insurance	Patient	Additional_insurance
Patient	Age	Patient	DATE - Date_of_birth	Patient	Age
Patient	Private_insurance	Patient	Private_insurance	Patient	Private_insurance
Patient	Relativ_flag	Patient	Relativ_flag	Patient	Relativ_Flag
Patient	Co_insured_from	Patient	Ins_begin_date	Patient	Co_insured_from
Patient	Co_insured_till	Patient	Ins_end_date	Patient	Co_insured_till
Therapy	Therapy_Nr	Therapy	Therapy_ID	Therapy	Therapy_Nr
Therapy	Therapy_date	Therapy	Therapy_date	Therapy	Therapy_date
Therapy	Therapy_charge_amount	Therapy_charging	Charge_amount	Therapy	Therapy_costs
Prescription	Prescr._ID	Prescription	Prescr. ID	Prescription	Prescr. Nr
Prescription	Prescr._date	Prescription	Prescr._date	Prescription	Prescr._date
Prescription	Prescr._charge_fee	Prescription	Prescr._charge_fee	Prescription	Prescr._charge_fee
Prescription	Prescr._additional_fee	Prescription	Prescr._additional_fee	Prescription	Prescr._additional_fee
Prescription	Prescr._tax	Prescription	Prescr._tax	Prescription	Prescr._tax
Drug	Drug_Nr	Drug	Drug_code	Drug	Drug_Nr
Drug	Drug_Descript.	Drug	Drug_descr.	Drug	Drug_Descript.
Drug	Package_size	Drug	Package_size	Drug	Package_size

Table 8: Mappings between Component DWH Data Models and the Federated Data Model

Since the federated data model may be quite complex, external data models can be used to specify subsets of data that are relevant and can be accessed by specific groups of users. These models are less complex and therefore easier to adopt in order to accommodate changing users' needs. External schemas can use data models different from the one used by the federation. In our example, we assume that the external data model is equal to the federated data model.

6 Rule 2: Healthcare Standards for Message Exchange

Syntax and semantics are two major concepts of the interoperability. Syntax refers to the structure of communication, whereas semantics expresses its meaning. Common standards for message exchange in healthcare environment cover both notions of interoperability, so that data can be exchanged and understood by the receiver. In this chapter, we introduce some internationally adopted healthcare communications standards, in particular HL7.

Patient records contain not only well arranged and concise lab test results or hospital admission forms, but often images, personalized text descriptions and other unstructured data. In order to enable seamless transmission and understanding of medical data among health providers, social insurance companies, and governmental agencies, application of internationally adopted standards is necessary.

6.1 Categories of Standards

Current international standards used for healthcare data exchange, like HL7, ENV 13606, and openEHR have been developed in parallel since early 1990s, some of them adopting certain concepts from others. The deployment of standards allows for consistent communication between heterogeneous healthcare information systems, without losing meaning or context. Table 9 gives the basic overview of the major standard categories.

Category	Description	Examples
Data Exchange/ Messaging Standards	Contain instructions (or specifications) for format, data elements, and structure and therefore allow transactions to flow consistently between systems or organizations	HL7 for administrative data, DICOM for radiology images, NCPDP for electronic prescriptions

Terminology Standards	Provide specific codes for clinical concepts that might have varying textual descriptions in a paper chart or a transcription.	LOINC for lab results, SNOMED for clinical terms, ICD for medical diagnoses
Document Standards	Indicate what type of information is included in a document and where it can be found.	CDA (Clinical Document Architecture), CCR (Continuity of Care Record) for inter-provider communication, including: patient identifying information, medical history, current medications, allergies, and care plan recommendation.
Conceptual Standards	Allow data to be transported across systems without losing meaning and context.	HL7 RIM (Reference Information Model) for describing clinical data and the context surrounding it.
Application Standards	Determine the way business rules are implemented and software systems interact.	Standards for providing a comprehensive way of viewing information across multiple, non-integrated databases.
Architecture Standards	Define the process involved in data storage and distribution.	PHIN components of an electronic surveillance and management system for integrated bioterrorism and public health preparedness.

Table 9: Summary of Healthcare Standards, adopted from [Kim, 2005]

In the following subchapters we will address those standards that we adopted for our approach.

6.2 HL 7

Health Level Seven [HL7, 2007], is a standard for exchanging information between medical applications. "Level Seven" refers to the seventh OSI layer protocol for the health environment. HL7 is a protocol for data exchange which defines the format and the content of the messages that applications must use when exchanging data with each other. A lot of European countries (for instance Netherlands, Finland, and Great Britain) have chosen HL7 as a strategic concept for a nation-wide healthcare communication standard.

HL7 specifies the contents and the formats of the exchanged messages at the application level. It is widely independent from the underlying hardware and network infrastructure as well as the database and applications used.

Since 1997, the HL7 organization has been developing version 3.0 of the protocol. HL7 Version 3 (HL7 V3) is a complete redefinition of the HL7 standard that is intended to try and overcome some of the issues with the current standard. Version 3 will change not only the content of the messages and fields, but also the encoding rules, LLP (low level communication protocols), base data types, and even the roles of the applications participating in HL7 communications. XML is the planned medium for HL7 interchange instead of the simple ASCII text that is currently in use [InterfaceWare, 2007].

Unlike 2.X versions, which have been implemented only for hospitals, HL7 V3 is being designed to suit the needs of all healthcare network participants. HL7 V3 is based largely on a single formal object-oriented model called the Reference Information Model (RIM). The goal of RIM is to reduce the implementation costs of HL7-enabled solutions and further standardize the HL7 communication specifications between healthcare systems.

6.3 HL7 RIM Foundation Classes

HL7 RIM follows the general principles of UML. However, it cannot be regarded as a UML extension or to be totally UML-compliant. The backbone of RIM consists of six classes: Act, Entity, Role, Participation, RoleLink, and ActRelationship. Their detailed description (for RIM version 2.10) is given in Table 10. The UML class diagram containing the relationships between them is shown in Figure 28.

Class name	Definition	Example	Subclasses
Act	action that is being done, has been done, can be done, or is intended or requested to be done.	clinical observation, discharging a patient	yes
Entity	physical thing, group of physical things or an organisation capable of participating in Acts	person, animal, medical device	yes
Role	competency of an Entity participating in an Act	patient, doctor, nurse	yes
Participation	association between an Act and a Role, with Entity playing that Role	Dr. Smith prescribes a therapy for patient Doe	no
ActRelationship	directed association between a source Act and a target Act.	a biopsy procedure as a result of an observation	no
RoleLink	connection between two roles expressing a dependency between those roles	clinician (employee)– hospital (employer)	no

Table 10: The Six Backbone Classes of RIM

Participation represents the many-to-many relationship between Roles and Acts. This relationship is binary: each Participation joins a single Role (and the Entity performing it) to a single Act. In reality, healthcare procedures are complex interactions, the most common examples involving several people (Entities playing Roles) participating in a "complex" Act. For instance, a patient is examined by a clinician and a diagnosis is stated. RIM splits complex events into one-role-one-act Participations.
The patient being examined participates in a PatientEncounter Act (a subclass of Act). The clinician, examining the patient, participates in an Act of Observation (also a subclass of Act). An

ActRelationship joins the two Acts and may express a composition, sequence, source-target, or condition relationship. There is also a set of HL7 data types associated to RIM.

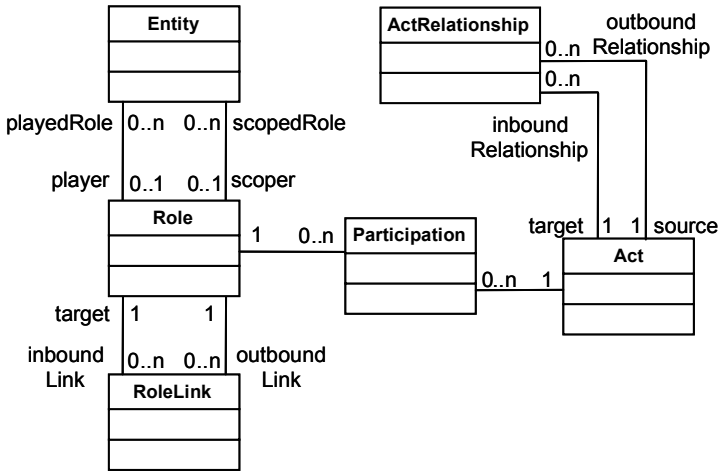


Figure 28: Class Diagram Showing the Backbone Classes of HL7 RIM

6.4 Clinical Document Architecture (CDA)

The structure and the semantic of clinical documents for the purposes of their exchange is specified by the HL7 Clinical Document Architecture (CDA), which is an XML-based document markup standard. CDA follows the concepts of RIM and uses HL7 data types.

```
<ClinicalDocument>
   <!-- Header-->
      <recordTarget> <!--(1..*)-->
         <patientRole>
            <id extension="12345" root="2.16.840.1.113883.19.5"/>
            <patient>
               <name>
                  <given>Henry</given>
                  <family>Levin</family>
               </name>
               <administrativeGenderCode code="M"
                                codeSystem="2.16.840.1.113883.5.1"/>
               <birthTime value="19320924"/>
            </patient>
        </patientRole>
     </recordTarget>
     <author><!--(1..*),sub-elements similar to recordTarget--></author>
     <custodian><!--(1..1),sub-elements --></custodian>
     <legalAuthenticator><!--(0..1),sub-elements--></legalAuthenticator>
     <!-- other roles: dataEnterer, informant, participant -->
   <!-- Body -->
</ClinicalDocument>
```

Figure 29: Structure of CDA Document Header [Beyer et al., 2004]

CDA XML document can be easily converted into a human-readable document using XSLT style sheets. CDA focuses on Acts, Entities, and Roles. Their interaction is defined in a more flexible

manner than in RIM, corresponding to the great structure freedom of XML. CDA data types describing RIM Acts, Entities, and Roles conform to the RIM model but the attribute matching is not complete. Those data types simplify the RIM model eliminating some of its optional attributes and, occasionally, introducing some new ones.

A CDA document consists of a header and a body. Structure of the header is shown in Figure 29. The header gives information about the roles participating in the described medical events: the owner of the patient record the document is attached to (normally a patient), the author (a clinician), legal authenticator of the document (a senior clinician, head of the department), and the custodian - organization that is in charge of the document (a hospital or social insurance institution). Entity is specified as a sub-element of the role element. Elements recordTarget and patientRole in Figure 29 describe the RIM Role Patient. Sub-elements of patient conform to the attributes of Person, a sub-class of RIM Entity. The example has been taken from CDA Specification [Beyer et al., 2004] and then simplified.

```
<ClinicalDocument>
   <!-- Header-->
   <!-- Body -->
     <component>
        <structuredBody>
           <component>
              <section>
                 <code code="10153-2"
          codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"/>
                 <title> Body temperature </title>
                 <text>
                    <list>
                       <item>36.9 C</item>
                    </list>
                 </text>
                 <entry>
                    <observation classCode="OBS" moodCode="EVN">
                       <code code="386725007"
                             codeSystem="2.16.840.1.113883.6.96"
                             codeSystemName="SNOMED CT"
                             displayName="Body temperature"/>
                       <statusCode code="completed"/>
                       <effectiveTime value="200004071430"/>
                       <value xsi:type="PQ" value="36.9" unit="Cel"/>
                    </observation>
                 </entry>
              </section>
              <!--other sections-->
           </component>
        </structuredBody>
     </component>
</ClinicalDocument>
```

Figure 30: Structure of CDA Document Body [Beyer et al., 2004]

The body of the document contains CDA representation of RIM Acts. As shown in Figure 30 (again the simplified example from CDA Specification from [Beyer et al., 2004]), it is a set of recursively organized components and sections that form the structure of a human-readable document. Each section is a pair of human-readable text and a number of entry elements wrapping descriptions of acts.

The entry in Figure 30 describes the process of body temperature measurement. The topic of the Observation is uniquely stated by the code sub-element of observation. Here, terminology standards LOINC [Regenstrief_Institute, 2007] and SNOMED CT [SNOMED, 2007] have been used, but other approved coding systems may also be applied. (Terminology standards are vocabularies which provide specific codes for clinical concepts such as disease, problem lists, allergies, medications, diagnoses etc. [Kim, 2005])

Relationships between acts (which are instances of RIM ActRelationship class) are generally expressed by grouping Acts into sections and components. Still, in rare cases when an explicit specification of different RIM ActRelationships types is needed, additional CDA XML tags can be used, which nest the dependent Act structure into the structure of the source Act as its sub-element.

6.5 openEHR, ENV 13606 and xDT

openEHR and ENV 13606 also introduce object-oriented reference models and a modular structure of healthcare documents. The general information model of openEHR [openEHR, 2007] describes only the nested hierarchical structure of healthcare records. Clinical data is defined separately for each healthcare domain using an ontology-defining constraint language.

ENV 13606 (proposed by the European Committee for Standardization) is currently under substantial revision due to its unnecessary complexity, which even led to some ambiguity and non-interoperability [EHCRSupA, 2007].

xDT [KBV, 2006] is a de-facto standard in Germany, used by social insurance organisations, pharmacists, and primary healthcare providers. Meanwhile, German hospitals have adopted HL7 standards. A comprehensive integration of xDT and HL7 standard has been performed by Sciphox company [SCIPHOX, 2006]. The previously used octet-encoded xDT messages have been abandoned and HL7 CDA and XML introduced. There is no general object model for xDT and its document structure is domain-dependent.

We chose to use HL7 RIM and CDA for the conceptual model of our federated DWH. Apart from HL7 RIM, no other standard offers an integrated model for all healthcare domains that can precisely define the basic structure of facts and dimensions, preventing any semantic or structural ambiguity. Fact and dimension attributes can be understood as feature descriptions having a domain precisely defined by RIM, but might actually consist of several attributes in a real database (easily defined by XML structures of CDA).

7 Rule 3: Semantic Integration in a Federated DWH Model

Semantical integration of the very complex and heterogeneous medical data structures is a demanding task. Our objective is to propose a DWH-based, semanticaly focused integration model for integration of diverse healthcare data sources which could even include external data suppliers or some third-party stakeholders' web services. The central part of our solution is the semantic integration layer, containing wrappers and a mediator.

During the last few years, healthcare organizations have been confronted with massive knowledge processing challenges. The primary cause for this is the increasing amount of complex medical data, but also the need to integrate wide range of data sources into a unique knowledge repository which could be used for EBM. Although willing to share their data with other healthcare providers, healthcare institutions prefer keeping their existing systems and metadata in place and participating in the federation in which integrated metadata is build on demand.

When several healthcare institutions try to consolidate their typically diverse metadata schemas, it is very likely that semantic conflicts will occur. Getting back to our running example, we illustrate semantic heterogeneity between two records corresponding to distinct metadata schemas in Figure 31.

Our goal is to create a metadata integration model, which will provide a uniform view and congeneric access to the wide range of heterogeneous data sources. Building such a model requires:

- common semantic representation (preferably by an ontology language)
- common data model
- query language that operates on this data model.

[Haslhofer, 2006] suggests the use of:

- OWL [W3C, 2004a] language for modelling ontologies, because it provides all the constructs and the expressiveness for describing the semantics of data,
- RDF data model [W3C, 2004b], because it is simple and yet powerful enough to allow the description of metadata of any kind originating from various heterogeneous sources,

- SPARQL [W3C, 2007c], a query language which operates on the RDF data model. Given that wrappers can translate between RDF and native data modes, SPARQL can be used for accessing data sources in an integrated fashion.

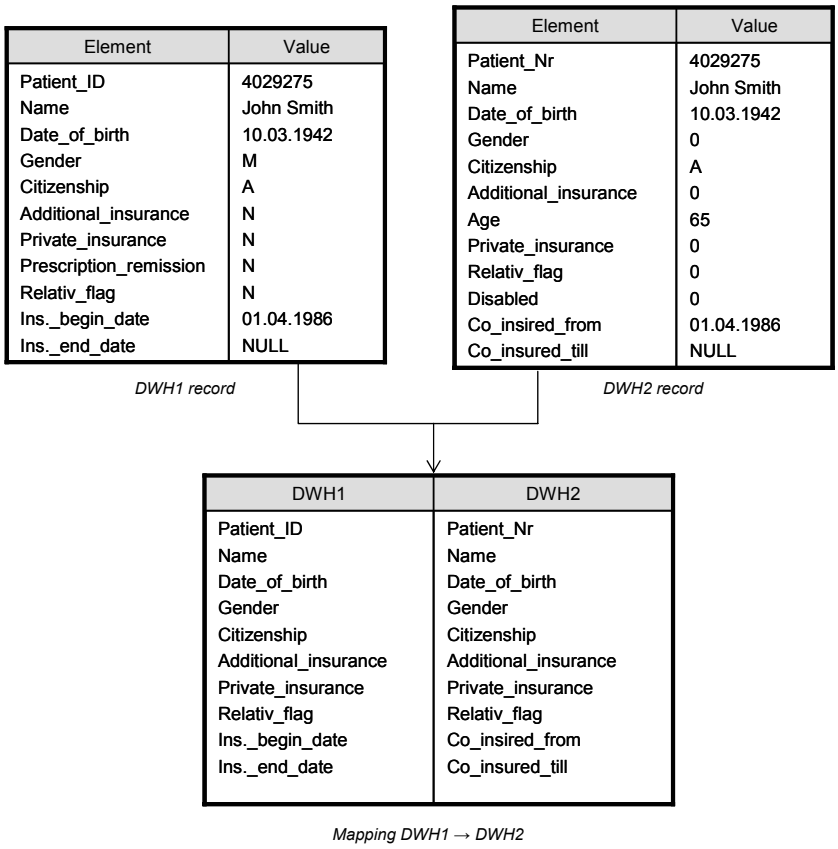


Figure 31: Semantic Heterogeneity of Two Metadata Records

7.1 Medical federated DWH Integration Workflow

The model we are describing in this subchapter is based on a mediated federated DWH architecture, with a carefully designed semantical infrastructure (Figure 32).

In our federated DWH model, the different healthcare domains, i.e. the social insurance domain, the pharmaceutical domain, and the evidence-based guidelines repository are participating in one federation. The existence of the federation is invisible to users of the source systems.

The clinical DWH contains data that originates from a wide variety of data sources, such as: clinical data, administrative data, financial data, and organizational data.

The health insurance company stores information about patient encounters, treatments, therapies, and drug prescriptions it supports. It communicates with the federation via web services.

Both clinical DWHs and the health insurance companies transfer their sensitive data to the federation, in case of a federated query. At this time, we assume that depersonalization and pseudonymization

techniques are used to protect the confidentiality of patient data. We will handle this issue in more detail in the next chapter.

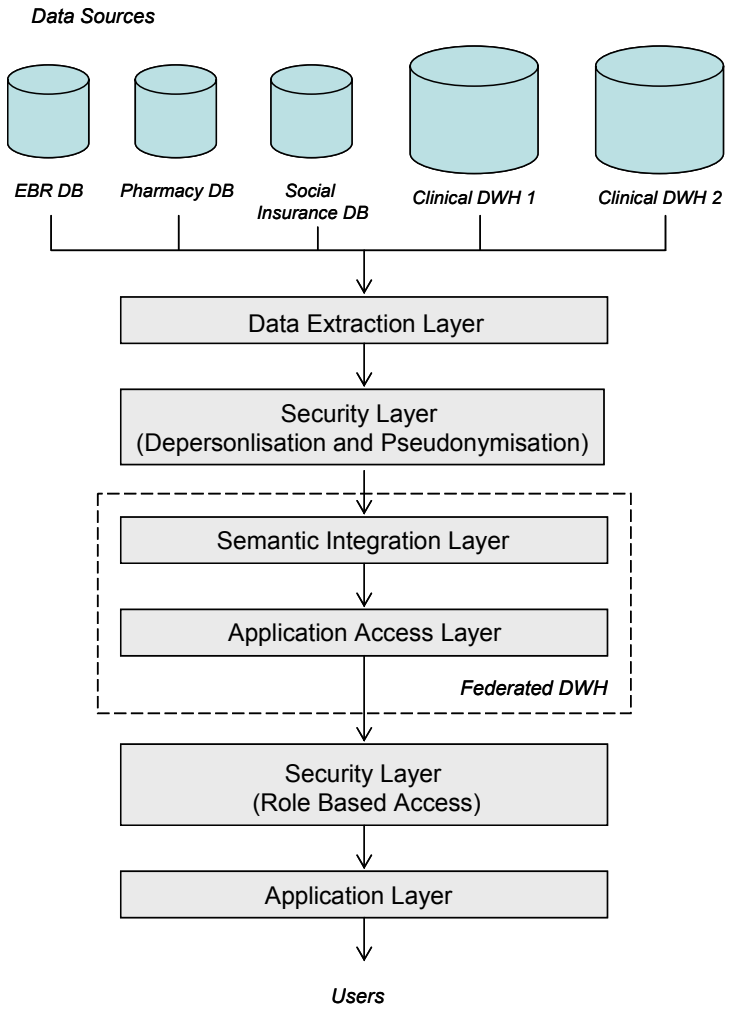


Figure 32: Heterogeneous Data Source Integration Layers

The pharmaceutical sector provides the federation with drug information such as medication description, packaging size, pharma-id number, indication group, pharmaceutical form (pills, juice etc.), and medication fee. In an analogous way to the social insurance company, it uses web services to provide the federation with the necessary data.

The evidence-based rules repository is a collection of most accurate and most efficient research evidence. It provides the federation with the best fitting guidelines "on demand" for a given patient and a given disease through web services.

Since only a "single version of evidence" and a unique interpretation of the joined data should exist, it is necessary to have a unique singular common federated schema, as described in previous chapter. For building the conceptual model of the federated DWHs, just data relevant to further analyses and reporting are considered. In this phase, business users (from the domain of clinical/social insurance

management) have to specify the respective sensitivity levels of data. The data modeller incorporates the specified privacy restrictions into the resulting logical data model.

7.2 Mediated Query System in an Healthcare Environment

The essential part of integration of logical schemas of the underlying DWHs as well as of data structures originating on the diverse participating legacy systems (such as relational or XML databases) is the semantic integration layer (Figure 33). Our model includes wrappers and mediator, which are two main architectural components of a mediated query system.

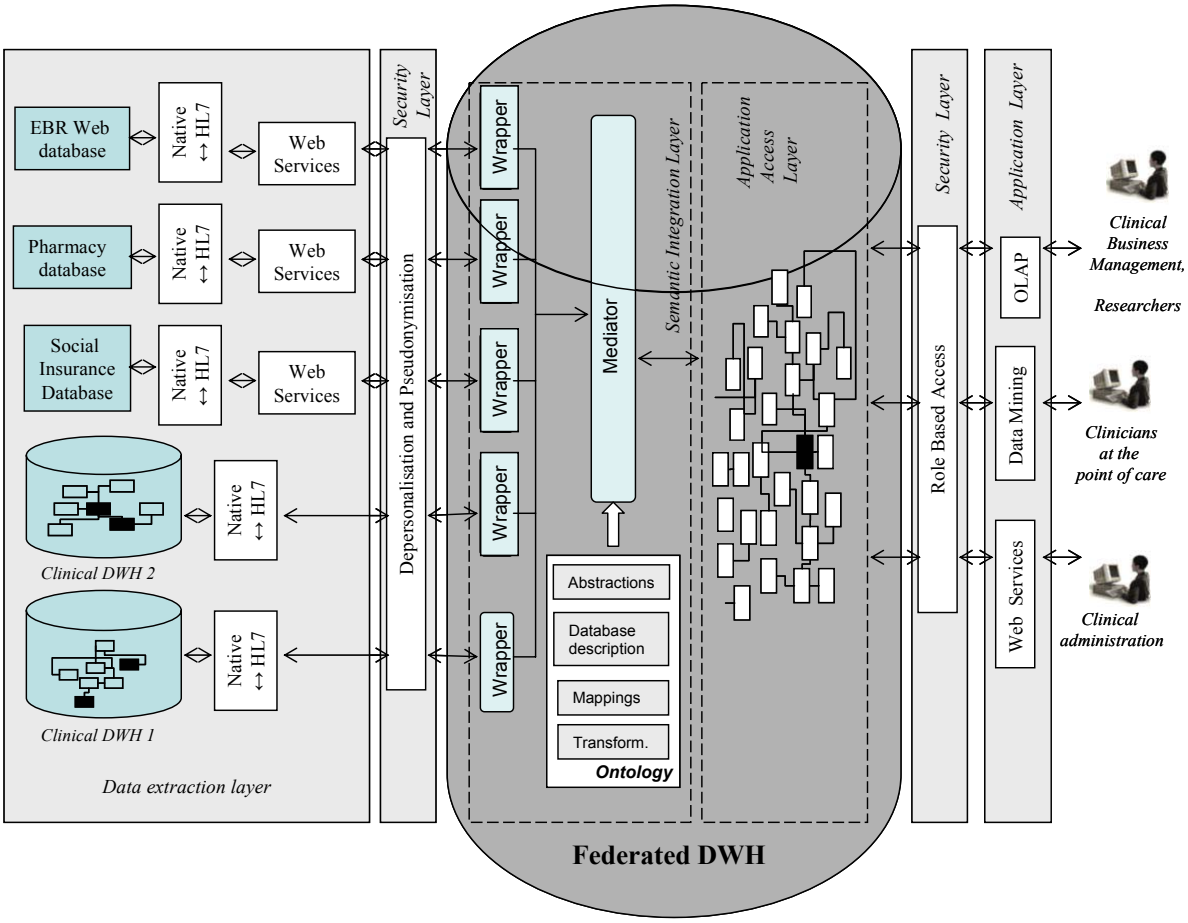


Figure 33: Medical Federated DWH Integration Workflow

7.2.1 Ontology

Ontologies are common methods of representing knowledge for interacting with heterogeneous data sources. In the DataFoundry research project [Critchlow et al., 1998], the authors focus on the use of ontologies as a formal method for storing and using the metadata required to perform automatic mediator generation.

The DataFoundry ontology represents four different concepts for mediator generation:

1. *Abstractions* of domain specific concepts - they represent knowledge about the concepts contained within the data sources being integrated.
2. *Database descriptions*, which consist of language independent class definitions that closely mirror the physical layout of a relational database.
3. *Mappings,* used to identify the correspondence between database and abstraction attributes.
4. *Transformations*, which identify methods for conversion among different representations of the same characteristics.

7.2.2 Wrapper

Wrappers [Haslhofer, 2006] encapsulate local data sources and export their functionality and the metadata stored therein. They accept queries in a certain language and return metadata in a united form. The wrapper keeps the data schema locally for the specific data source it deals with. It translates the semantics of the metadata between the exported schema and the native data descriptions. Depending on the data source, a wrapper can be a default DBMS interface (for relational databases) or custom-built for specific sources.

By integrating wrappers, we can cope with technical heterogeneities among local systems without having to modify them.

7.2.3 Mediator

The mediator [Haslhofer, 2006] handles the global queries from the application layer, unfolds them into sub-queries and disperses these sub-queries to the relevant local data sources via their wrappers. The local results will be returned by wrappers; the mediator finally combines and presents the result to the client. Hence, the mediator will keep the global data schema and the mapping between global and local schemas. To maintain the dynamic mapping between local and global schemas, an semantic-based mediator/wrapper is one of most interesting problem solving approaches.

7.2.4 Query Handling

Figure 34 shows an example of how a user's query is handled by wrappers and a mediator, from the query submission to the presentation of the result. On receiving the SQL query, the system performs the following:

1. DWH-Application invokes the mediator.

Query unfolding:

2. The mediator resolves the submitted query into partial queries, according to the exported schemas previously exposed by the wrappers. It determines which wrappers are relevant to the corresponding sub-queries.
3. The mediator passes the sub-queries to the affected wrappers.

4. A wrapper receives its sub-query and translates it into the format that can be understood by the underlying data source (database, web service etc.).
5. The wrapper forwards the adapted sub-query to the local DBMS or to the responsible Web Service for execution.

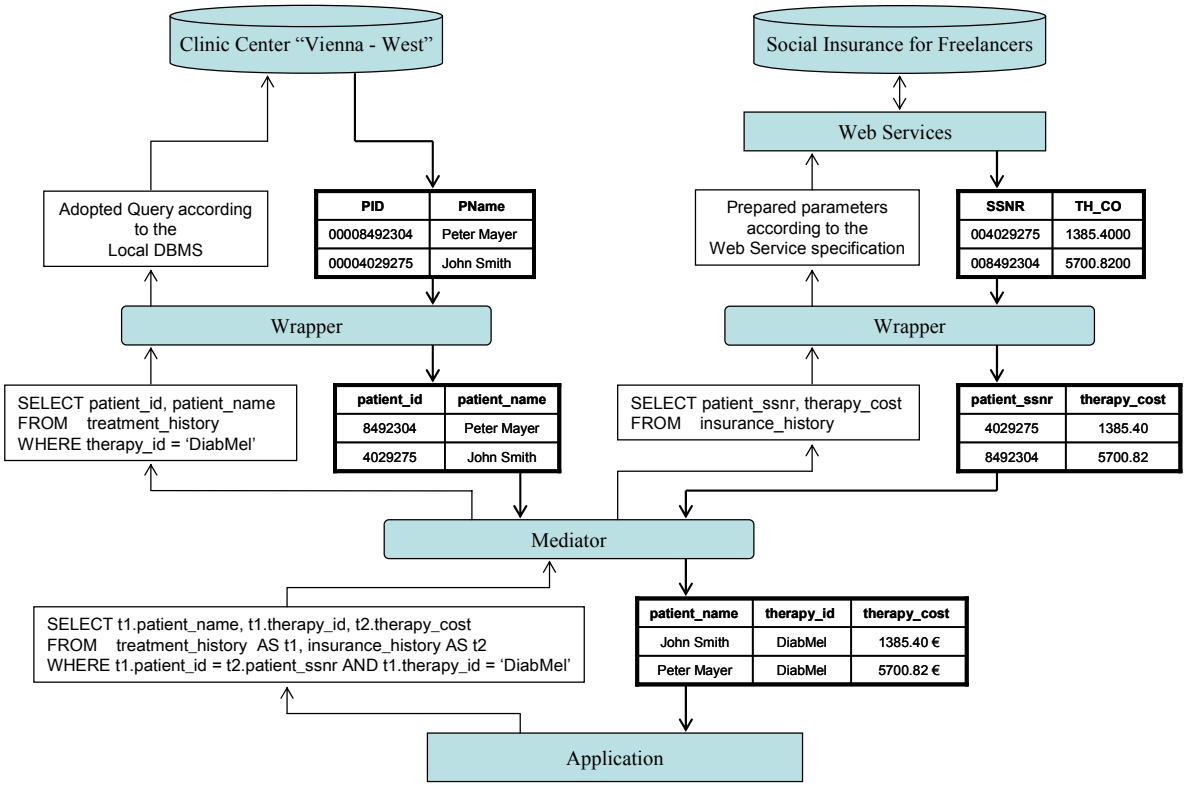


Figure 34: Mediated Query System in Healthcare Environment

Query answering:

6. The wrapper retrieves the answer data set, translates it into its exported schema.
7. The wrapper passes its answer data set to the mediator.
8. The mediator integrates partial query results into one answer set, transforms and formats it so that it can be processed by the application.
9. The mediator passes the answer set to the application so that it can be presented to the user.

The users of a federated DWH are not aware of the fact that the data they are querying may be distributed across the network. Through data mining tools, web services, ad-hoc queries, and predefined reports (OLAP tools), users are able to analyse data as if they were physically stored in a centralized DWH. A role-based access model guarantees that each user gets access only to those data which are necessary for performing his (her) tasks.

7.3 Running Example: Federated DWH Supporting Clinicians at the Point of Care

In order to illustrate the described federated DWH-based approach, we build upon our running example, introduced in previous chapters. Here, we assume that the emergency room clinician is querying the federated DWH while examining the patient. The following scenario depicts the starting point of our example: The Diabetes patient suffering from a progressive liver disease is complaining about itchy rash on his hands. Since the attending physician is not familiar with the patient's medical history and needs to act quickly, he(she) is using federated DWH facilitating EBM to find the most efficient therapy, which does not conflict with the patient's ongoing Diabetes and liver disease treatment.

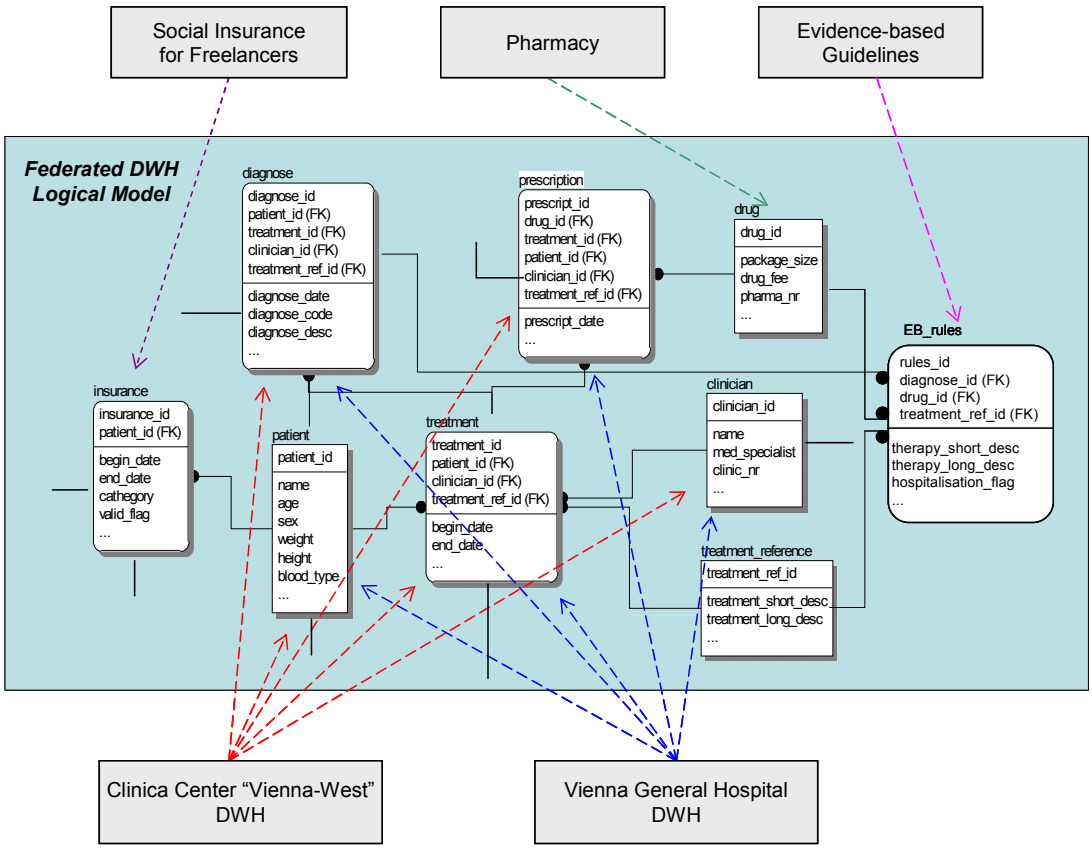


Figure 35: Extract from Federated DWH LDM with Corresponding Data Sources

The clinician is using an OLAP tool which is set up on the federated DWH. One small but representative part of the federated DWH LDM is shown in Figure 35. Dashed arrows show the data flow between underlying data sources and resulting tables.

The clinician is querying the patient's healthcare record, containing all existing anamnesis/diagnostic data and all of patient's past treatments. Further, he(she) is interested in the patient's overall health condition (allergies and medication incompatibilities) as well as the personal data (age, weight, family predisposition to some diseases etc.). In addition, the clinician has to find out what kind of therapies

will be covered by the patient's social insurance. Finally, he(she) is aiming to find the treatment which proved to be the most effective under given conditions.
The answering procedure takes place in two phases:

1. The user (the clinician) is querying the federated DWH by providing only the patient's name and the corresponding social insurance number. In the first step, the mediator sends queries containing these two parameters to all relevant participating sources (in this case study, these are: clinical DWHs, pharmacy, and social insurance company). The aim of this step is to retrieve all patient data, which might be interesting as input for the second step, namely querying EBG database.
   Responsible wrappers return the corresponding data to the mediator, which in sequel joins them and produces a result data set.

2. In the second step, the resulting data set (containing patient's age, blood pressure, blood type, diseases history, list of received treatments and medications, social insurance categorisation etc.) is used as the input parameter for querying of the evidence-based guidelines database. The mediator forwards a new query to the EBG- database wrapper and retrieves the final data set.

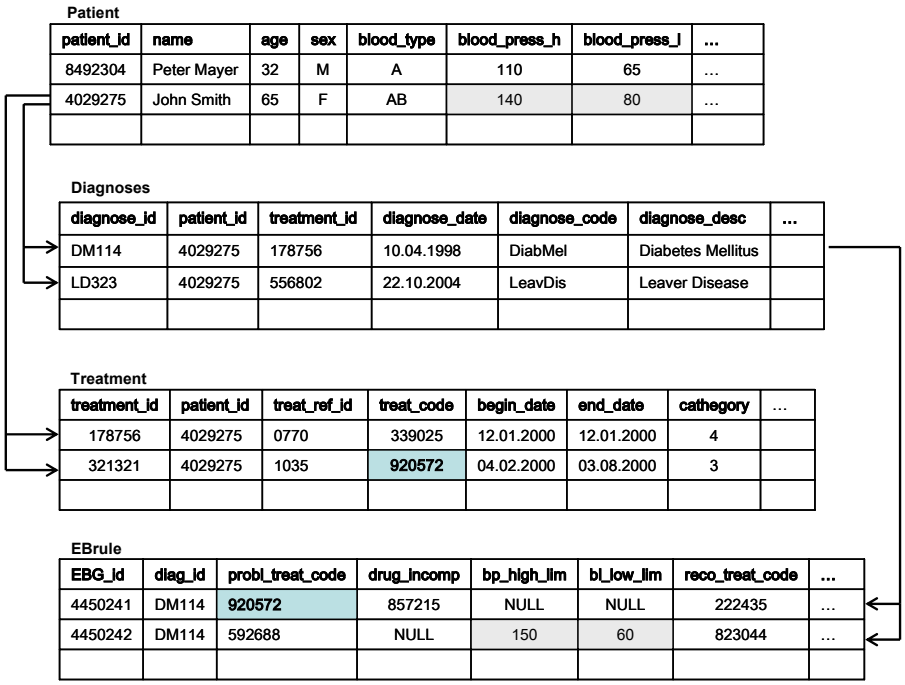


Figure 36: A Group of Tables Involved in Finding the Best Fitting Evidence-Based Rule

As shown in Figure 36, some of the existing treatment rules are disqualified due to medication or treatment incompatibilities. Since the proposed treatments must be adjusted to the patient's health risks, parameters like blood pressure may play the determining role in the treatment verification

process. Nevertheless, the scope of a patient's social insurance contract (refers to treatment categorization) is significant for the determination of applicable treatment.

In our example, the evidence-based guideline with identification: EBG_id = '4450241' is not applicable for reasons of treatment incompatibility. Namely, in the past, Mr. Smith has received a treatment with treatment_code = '920572', which is listed as a problem treatment for this guideline.

The evidence-based guideline with identification: EBG_id = '4450242' is the best fitting guideline for Mr. Smith's medical condition, so it is forwarded to the federated DWH.

In the last step, an OLAP tool presents the result to the clinician in an understandable and illustrative way.

8 Rule 4: Security Requirements for the DWH Federation in the Field of Evidence-Based Healthcare

Due to the innovative character of the field of evidence-based medicine, no explicit common security regulations and standards exist yet. Defining these measures is a process in which both the patient's individual rights (patient's privacy and data protection) and the collective, societal demands (scientific progress and development of new technologies) need to be considered. This chapter discusses the security measures to be enforced in order to protect the extraordinarily sensitive nature of health data while using clinical data sources for knowledge discovery and development of evidence based guidelines. Our approach consists of depersonalization, pseudonymization, and role-based access.

Ensuring information system security by avoiding unauthorized access, preventing intrusions, information disclosure, etc. is an important and demanding issue that organisations across all industries are dealing with. Over the past few years, there has been an increased awareness of data security in the healthcare sector. Healthcare decision support systems are comprised of large volumes of *sensitive* data and therefore must guarantee a high degree of their protection. Protection of confidential patient data is a subject of international regulations. Issues of privacy, software regulation, and ethical and legal aspects of data processing in healthcare may become main sources of conflicts.

Healthcare providers and governmental authorities are focused on adopting secure information systems which would both drive their business success and protect the confidentiality of patient data. Privacy considerations are the main arguments against collecting of data from different, remote sources into a centralized DWH system. DWH and ad-hoc OLAP analysis produce a security conflict: on one hand, they strive to make all available data as easily accessible as possible, while on the other hand they must incorporate security measures to protect sensitive, low-level data stored in the DWH.

When data mining is deployed on top of the medical DWH, it is usually required to apply the mining algorithms without observing the confidential data values. There are many challenging issues that need

further investigation in the context of federated DWH in healthcare sector from both privacy and security perspectives.

8.1 Guaranteeing Security in the Healthcare Environment

Loading delicate medical data into diverse unsecured databases or putting it on-line increases the risks of data disclosure by unauthorised users. Apart of containing basic demographic information about the patient, such as height, weight, gender, and blood type, electronic medical record may also contain some information about psychiatric treatments, HIV status, genetic predispositions, physical abuse, and so on. Unlike other businesses, where security gaps are reparable, every exposure of medical information causes unrecoverable privacy losses. For example, a credit card fraud caused by security deficiency of a bank can be repaired, while disclosed medical information of a person cannot be made secret again [Rindfleisch, 1997].

Detecting, controlling, and combating disease outbreaks early are the keys to preventing life-threatening infectious diseases. Outbreak of diseases such ad bird flu, SARS or even threats of bio-terrorism have made disease surveillance into a national priority. For example, the Real-Time Outbreak Detection System (RODS) at the University of Pittsburgh Medical Center [Tsui et al., 2003], uses data collected from regional healthcare providers and purchase records of over-the-counter drugs to determine outbreak patterns. This system forwards all regional data to a central DWH for evaluation purposes. Although data is de-identified in accordance with HIPAA safe-harbour rules by removing 19 kinds of identifiers, privacy concerns remain about both patient privacy and organizational privacy [Clifton et al., 2004].

The need to improve their business and the quality of healthcare will encourage medical organisations to release their highly sensitive data for loading into federated DWH. The consequence is that the security measures in such a system need to be extremely reliable.

The usability of the federated DWH systems is of extreme importance and therefore one has to prevent the danger that the medical staff is overwhelmed by non-user-friendly security procedures.

8.1.1 Confidentiality Threats to Patient Data

Due to the growing competition in the healthcare market, health providers tend to use the patient medical information coming from the federated DWH for purposes other than those initially intended. Such consolidated data represent valuable asset for marketing of new drugs (in medical products domain), for targeting new customers (in health insurance domain), and so on. These uses go beyond the confidentiality concept of the patients signing the consent forms and belong to the long list of confidentiality threats in the healthcare environment.

The most important threats to healthcare information confidentiality may be grouped into three categories:
(1) from inside the patient care institution, (2) from within secondary settings, and *(3) from outsider intrusion into medical information systems [National_Research_Council, 1997].* Table 11 gives an overview of these threats.

Threat Group	Threat	Description
From inside the patient care institution	Accidental disclosures	Medical personnel make innocent mistakes and cause unintentional disclosures.
	Insider curiosity	Medical personnel abuse their record access privileges out of curiosity or for their own purposes.
	Insider subornation	Medical personnel knowingly access information and release it to outsiders for spite, revenge or profit.
From within secondary settings	Uncontrolled secondary usage	Those who have access rights to patient information for the purpose of support of primary care may exploit that access for other purposes not envisioned in patient consent forms (i.e. data mining)
Outsider intrusion into medical IS	Unauthorized access	Outsiders may steal information, damage systems, or disrupt operations.

Table 11: Confidentiality Threats [National_Research_Council, 1997]

Data mining represents another privacy threat, which is expected to grow, but can be successfully eliminated through the implementation of depersonalization technique.

8.1.2 The Basic Security Concepts for the Healthcare Environment

The main ethical concern of federated DWH for evidence-based purposes is to provide mechanisms and policies to preserve patient privacy while delivering a large decision support system for research purposes and supporting caregivers in their clinical practice. [Rindfleisch, 1997] describes the three well-known concepts for protecting healthcare information as follows:

- ***Privacy***: The right and desire of a person to control the disclosure of personal health information.
- ***Confidentiality***: The controlled release of personal health information to a care provider or information custodian under an agreement that limits the extent and conditions under which that information may be used or released further.
- ***Security****:* A collection of policies, procedures, and safeguards that help maintain the integrity and availability of information system and control access to their contents.

8.1.3 Security Measures for Protecting Patient Data Privacy

Security risks need to be eliminated by implementing suitable technical and organisational measures [Eckhardt, 2000]. The following is the list of the most important security measures which need to be considered to protect data privacy in decision support systems, in order to facilitate evidence-based medicine:

1. All users of a healthcare decision support system must identify themselves through a password
2. Any data modification must bear a digital signature
3. Data access needs to be logged, and log files are preserved for a certain period of time in order to enable later tracking of any data manipulation
4. Confidential health data should only be stored in a coded or encrypted form on a mobile medium
5. Transportation security must be assured through Public Key Infrastructure
6. Data used for evidence based medicine purposes needs to be depersonalized and pseudonymized
7. A role-based access model has to be implemented

The research goals of evidence-based medicine - recognizing the disease symptoms and treatment patterns from the clinical data - can be reached by analyzing unidentifiable patient data. Depersonalization and pseudonymization procedures are used to protect patient privacy and confidentiality. We consider the last two security measures to be of a very special interest in the area of evidence-based medicine and we will therefore discuss them in more detail in the next subchapters.

8.1.4 The Role of Patient Consent

The final point in making a decision about which data will be released for sharing has to be the patient. Through the patients' portal, the patient should be enabled to specify his (her) privacy preferences for sharing personal and medical data for research, marketing, insurance coverage purposes etc. For example, the patient can allow the dissemination of his(her) medical data for research use but forbid the use of personal data for marketing purposes.

The consequence of the patient's uncertainty about the protection of his (her) medical data might be the avoidance of the needed treatments or specifying incorrect data for his (her) medical record. This carries the risk of use of incorrect data for research studies.

8.2 International regulations

Most privacy laws balance benefit vs. risk: access is allowed when there is adequate benefit resulting from access [Clifton et al., 2004]. In this subchapter, we give an overview of some significant international privacy and data protection regulations.

8.2.1 USA: HIPAA

Health Insurance Portability and Accountability Act of 1996, [HIPPA, 1996], is a set of regulations developed by various government agencies and members of the health industry. Its purpose is to simplify the information exchange between healthcare institutions and at the same time to protect the patient rights. HIPAA regulations have a mandatory character. All healthcare organisations in the USA have to comply with the HIPAA provisions, which are developed by the US Department of Health and Human Services (HHS). The administrative simplification provisions are organized in following four groups, as described by [Duftschmid et al., 2003]:

1. Electronic Health Transaction Standards

A set of ANSI formats, which have to be uniformly applied within all electronic transactions related to healthcare claims, health plan eligibility, referral certification and authorization, healthcare claim status enrolment and disenrollment in a health plan, healthcare payment and remittance advice, health plan premium payments, coordination of benefits, and related transactions are described.

2. Unique Identifiers

Distinctive IDs for healthcare providers, employers, health plans, and patients are defined. This should fix the problems resulting from the usage of multiple different ID numbers, which are allowed by the current healthcare system.

3. Security and Health Information and Electronic Signature Standards

Uniform level of protection for all electronically transferred or stored health information that can be linked to an individual is described. Apart from technical specifications regarding, among other things, transmission security and the implementation of access control and audit control mechanism, the installation of various administrative and physical safeguards is also prescribed.

4. Privacy and Confidentiality Standards

This part is concerned with the question of who has the right to access personally identifiable health information. It includes specifications that:

- Limit the non-consensual use and release of private health information.
- Give patients new rights to access their medical records and to know who else has accessed them.
- Restrict most disclosure of health information to the minimum needed for the intended purpose.
- Establish new criminal and civil sanctions for improper use or disclosure.
- Establish new requirements for access to records by researchers and others.

8.2.2 Canada: PIPEDA

PIPEDA (the *Personal Information Protection and Electronic Documents Act*) is a Canadian law governing how private sector organizations collect, use and disclose personal information in the course

of commercial business. PIPEDA was passed in the late 1990s to promote consumer trust in electronic commerce. The act was intended to reassure the European Union that Canadian privacy laws were adequate to protect the information of European Citizens [Wikipedia, 2007c].

As of January 1, 2004, all Canadian businesses are required to comply with the privacy principles set out by PIPEDA. The Act covers both traditional paper-based and on-line business.

Under PIPEDA, personal information must be [Strategis, 2005].:

- collected with consent and for a reasonable purpose
- used and disclosed for the limited purpose for which it was collected
- accurate
- accessible for inspection and correction
- stored securely

PIPEDA defines *personal information* as "information about an identifiable individual" that includes any factual or subjective information, recorded or not, in any form. The legislation also covers *sensitive personal information*, which may include health or medical history, racial or ethnic origin, political opinions, religious beliefs, trade union membership, financial information and sexual preferences.

8.2.3 European Community: Directive on Protection of Personal Data

Directive 95/46/EC on the protection of personal data is a part of the "Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data" [Wikipedia, 2007d]. This directive regulates the processing of personal data, regardless of whether the processing is automated or not. Its principle is that personal data should not be processed at all, except when certain conditions are met. These conditions fall into three categories: transparency, legitimate purpose, and proportionality [Wikipedia, 2007d]:

Transparency

The data subject has the right to be informed when his(her) personal data are being processed. The controller must provide their name and address, the purpose of processing, the recipients of data and all other information required to ensure the processing is fair. Data may be processed only under the following circumstances:

- when the data subject has given his (her)consent
- when the processing is necessary for the performance of or the entering into a contract
- when processing is necessary for compliance with a legal obligation
- when processing is necessary in order to protect the vital interests of the data subject
- when processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed

- when processing is necessary for the purposes of legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject .
- The data subject has the right to access all data processed about him. The data subject even has the right to demand the rectification, deletion or blocking of data that is incomplete, inaccurate or isn't being processed in compliance with the data protection rules.

Legitimate purpose

Personal data can only be processed for specified explicit and legitimate purposes and may not be processed further in a way incompatible with those purposes.

Proportionality

Personal data may be processed only insofar as it is adequate, relevant, and not excessive in relation to the purposes for which they are collected and/or further processed. The data must be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, with regard to purposes for which they were collected or for which they are further processed, are erased or rectified. The data should not be kept in a form which permits identification of data subjects for longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.

8.2.4 United Kingdom: The Data Protection Act

The Data Protection Act (DPA) is a United Kingdom Act of Parliament [Wikipedia, 2007b]. It defines a legal basis for the handling of information relating to living people in the UK. It is the main piece of legislation that governs protection of personal data in the UK. Although the Act does not mention privacy, in practice it provides a way in which individuals can enforce the control of information about them.

Compliance with the Act is overseen by an independent government authority, the Office of the Information Commissioner. The act defines the principles of information-handling practice, listed below [Wikipedia, 2007b]:

- Data may only be used for the specific purposes for which it was collected.
- Data must not be disclosed to other parties without the consent of the individual whom it is about, unless there is legislation or other overriding legitimate reason to share the information.
- Individuals have a right of access to the information held about them, subject to certain exceptions.
- Personal information may be kept for no longer than is necessary.

- Personal information may not be transmitted outside the European Economic Area (EEA) unless the individual whom it is about has consented or adequate protection is in place, for example by the use of a prescribed form of contract to govern the transmission of the data.
- Subject to some exceptions for organisations that only do very simple processing, and for domestic use, all entities that process personal information must register with the Information Commissioner.
- Entities holding personal information are required to have adequate security measures in place. Those include technical measures (such as firewalls) and organisational measures (such as staff training).

8.2.5 The Netherlands: NEN 7510

In addition to the Dutch Data Protection Act (Wbp) the *Dutch standard for data protection*, NEN 7510, is a relevant directive for the care sector [Dutch_DPA, 2006]. NEN 7510 provides guidelines and starting points to determine, implement, and maintain data security measures for care institutions. At present there is no statutory requirement for care institutions to comply with the standard. It is expected that this will be the case after 2006, as a result of the introduction of the Citizen's Service Number (CSN) in the care sector. With respect to the CSN, the Dutch Data Protection Authority (Dutch DPA) issued the "Advies over Wet gebruik BSN in de zorg" (Advisory regarding the Act relating to the use of the Citizen's Service Number in the care sector).

Comprehensive information about the protection of data can be found in the Dutch DPA study entitled "The Security of Personal Data". This report assumes three levels of protection, subdivided into exclusivity classes. Because of the use of special data, the highest level of protection is required. Some requirements for the protection of medical data are described in [Dutch_DPA, 2006]:

- the implementation of access control with respect to electronic files;
- the encryption (application of encryption techniques) of health-related data on the Internet;
- the storing of information carriers, such as paper files and computer disks, in a burglary-resistant environment;
- outsourcing of data to another organisation only if this organisation has implemented the required security measures;
- deletion of data so that reconstruction of the original data is no longer possible.

The aforementioned study lists a number of security regulations. It is important that clear agreements are made within an organisation with respect to measures to be implemented. These agreements must also be documented and must be known to employees.

8.2.6 Austria: MAGDA-LENA

MAGDA-LENA framework which guides healthcare data exchange in Austria has been developed by the STRING (Standards and Guidelines for the Employment of the Informatics in Austrian Healthcare) commission in 1998. It outlines both the technical and organizational aspects governing the development of Austrian healthcare information network, which will allow patient-related multimedia information to be exchanged between both healthcare and social security facilities. [Duftschmid et al., 2003] summarize its key sections as follows:

Message content, models, and standards

The goal of this section is to promote the use of standard formats for exchange of healthcare information. Message formats should not be defined by the parties themselves on a bilateral basis, but should be based on global rules.

Identification variables

For secure data transmission in the healthcare arena, both the communicating parties and the transmitted data must be unambiguously identified.

Data privacy and security

MAGDA-LENA recommends the participants to implement specific measures to achieve data privacy and security in the participants' internal domain as well as in their electronic communications with others.

Network providers and nodes

Minimum standards to be followed by network providers participating in the Austrian healthcare network are defined.

Unlike HIPPA, for the time being MAGDA-LENA guidelines have only a recommendatory status. We will discuss MAGDA-LENA references in more detail in the next chapter.

8.3 Preserving Privacy in an Federated Environment

Individual healthcare providers define their own security policies to protect the patient data stored in local DWH or local databases. When they participate in DWH federation, the privacy problem gets more complex. Even revealing the presence of sensitive patient data in an underlying DWH may endanger their privacy. In our approach, we propose depersonalization and pseudonymization techniques as well as role-based access mechanism to ensure the security of patient data.

8.3.1 Depersonalization

Depersonalization is a removal of any residual information that might risk identification – e.g. names of relatives, nick names, place names, unusual occupations, etc. [Taweel et al., 2004].

The HIPAA Privacy Rule allows healthcare institutions to disclose depersonalized health information without restriction. This means that if an institution de-identifies patient data to a statistically and scientifically acceptable level, as determined by a qualified individual, it may use and disclose such data for research [Agrawal et al., 2007].

The goal of depersonalization is to assure that re-identification of a person by means of other personal data available is not possible. This may be done by:

- ***Grouping data*** – hiding sensitive data by grouping (for example: patient's age is not shown precisely but in the age areas of 20-30, 30-40, 40-50, etc.). The definition of data groups is made according to the research or business goals.
- ***Hiding data*** – all personal data interesting for detailed data mining (occupation, hobbies), which can potentially be used for patient identification are concealed.
- ***Removing data*** – key identifying data unnecessary for the research (e.g. name, exact birthdate, precise address, nick names, name of relatives etc.) that can be used for patient identification are removed.

The growing amount of personal data available on the internet, the existence of fast computers with large store capacities, as well as the rush in developing data mining techniques, increase the risk of re-identification of personal data which is believed to be anonymous. Even when the most obvious identifiers are deleted or hidden when individual data is disclosed, the combination of other identifiers may still lead to re-identification of the patient. For instance, if the set of patient records has been depersonalized, no personal identifiers such as name, birthdate or social security number is available. Although no single identifiable attribute value exists, it is very likely that some combination of values would allow de-identification of the person. For example, for a patient who is the only female older than 80 living in a certain district and having occupation: Member of Parliament, it is not difficult to obtain her name and subsequently her medical history.

In our approach, we recommend the use of strong depersonalization measures, which in combination with pseudonymization and role-based access ensure the privacy of patient data.

8.3.2 Pseudonymization

Pseudonymity is one of the innovative privacy enhancing techniques, which can unlock valuable data sources. It is defined as a state of disguised identity resulting from the use of a pseudonym. The pseudonym identifies a *holder*, that is, one or more human beings who possess but do not disclose their true names (legal identities) [Wikipedia, 2007e].

Pseudonymization is suitable for use in the area of evidence-based medicine, because a consolidation of different patient data needs to be carried out, whereas patient identity must stay unknown.

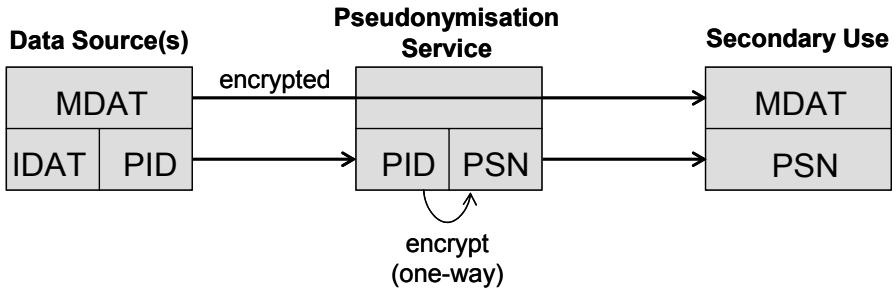
Depending on the requirements, two kinds of pseudonymization can be used:

1. one-way pseudonymization
2. reversible pseudonymization

For the purposes of EBM, there is no need for re-identification of a person, so we propose the use of one-way pseudonymization.

Our approach is based on a one-way pseudonymization procedure described by [Pommerening and Reng, 2004]:

In this procedure, it is assumed that a unique patient identifier (PID) is available in the electronic health record and in other data sources. The pseudonymization procedure than consists of a one-way encryption of the PID and should be implemented as a *Trusted Third Party* (TTP) service. A typical feature of this service is the use of *asymmetric encryption*: The data source encrypts the medical data with the key of the secondary user and sends the PID (not the identity data) as well as the encrypted medical data to the pseudonymization service. This service then encrypts the PID and sends it to the secondary user, together with the encrypted medical data (Figure 37). TTP cannot read the medical data, only the secondary user can decrypt them. The secondary user cannot decrypt the pseudonym.



MDAT – medical data **IDAT** – identity data **PID** – patient identifier **PSN** – pseudonym

Figure 37: Data Flow for One-Time Secondary Use [Pommerening and Reng, 2004]

Because the pseudonymization service does not store the association between PIDs and pseudonyms, and cannot reverse the encryption, there is no need to treat the PSN as secret, as long as the TTP implements an effective sender authentication and authorisation.

In case when the need for re-identification exists (for example, to warn patients of risks uncovered by research or in order to recruit patients for clinical trials [Kalra et al., 2004]), more complex pseudonymization forms need to be applied [GVG©, 2004]. This procedure is called reversible pseudonymization. Later re-identification can only be done by the data owner.

In order to recognize the patterns in medical data and to derive proper guidelines, EBM implies a long-term observation of patient data. For this reason, it is essential to assure that one patient always gets the same pseudonym, even if she (he) changes her (his) name, address or social insurance company.

8.3.3 Database Security Models

In the following, we will summarize the most common database security models – *discretionary security model, mandatory security model* and the one which we will apply in our approach: *role-based access model*.

8.3.3.1 Discretionary Security Model

Discretionary security specifies the rules under which subjects can create and delete objects, grant and revoke authorizations for object access to others. They are fundamental to operating systems and DBMS. Discretionary access controls (DAC) are based on the concept of a set of security objects O, a set of security subjects S, a set of access privileges T defining the kind of access a subject has to a certain object, and, in order to represent the content-based access rules, a set of predicates P. Applied to relational databases O is a finite set of values representing relational schemas. S is a finite set of potential subjects representing users, groups of users or transactions operating on behalf of users. Access type privileges are the set of database operations such as select, insert, delete, update, execute, grant or revoke. The tuple $<o,s,t,p>$ is called access rule and a function f is defined to determine if an authorization $f(o,s,t,p)$ is valid or not:

$$f: O \times S \times T \times P \rightarrow \{True, False\}$$

For any $<o,s,t,p>$, if $f(o,s,t,p)$ evaluates into True, subject s has authorization t to access object o within the range defined by predicate p.

Most systems supporting DAC store access rules in an access control matrix. In its simplest form the rows of the matrix represent subjects, the columns represent objects and the intersection of a row and a column contains the access type that the subject has authorization for with respect to the object.

Discretionary security is enforced in most commercial DBMS products and is based on the concept of database views. Instead of authorizing a user to the base relations of a system, the information of the access control matrix is used to restrict the user to a particular subset of the data available [Kirkgoze et al., 1997].

8.3.3.2 Mandatory Security Model

While discretionary models are concerned with defining, modelling, and enforcing access to information, mandatory security models are in addition concerned with the flow of information within a system [Kirkgoze et al., 1997]. Mandatory security requires that security objects and subjects are assigned to certain security levels represented with a label. The label for an object o is called its classification (*class(o)*) and a label for a subject s is called its clearance (*clear(s)*). The classification represents the sensitivity of the labelled data, while the clearance of a subject represents its trustworthiness to not disclose sensitive information to others. A security label consists of two components: a level from a hierarchical list of sensitivity level or access classes (for example: *top secret* > *secret* > *confidential* > *unclassified*) and a member of a non-hierarchical set of categories representing classes of object types of the universe of discourse. Mandatory access control (MAC) requirements are formalized by two rules. The first one protects the information in the database from unauthorized disclosure and the second one protects data from contamination or unauthorized modification by restricting the information flow from high to low.

1. Subject s is allowed to read data item d ***if clear(s) >= class(d).***
2. Subject s is allowed to write data item d ***if clear(s) =< class(d)***.

8.3.3.3 Role Based Access Control Model

Access control models in DWH environment are responsible for determining which (group of) users are allowed to access what data portions. DBMS apply access control to regulate access to tables and views.

There are three abstractions of control in access control system context: *access control policy*, *access control model*, and *access control mechanism* [Ferraiolo et al., 2003]. Policies are high-level requirements that specify how access is managed and who, under what circumstances, may have access to what information. At a high level, those policies are enforced through a mechanism that translates user's access request, often in terms of a simple table lookup, to grant or deny access. Access control models are written to describe the security properties of an access control system.

Adopting mandatory access controls to better fit into general purpose data processing practice and offering a design framework for database containing sensitive information are the main goals of the Role Based Access Control (RBAC) model. RBAC model is traditionally used to manually assign users to the eligible roles or, in case of a large number of users, to perform role assignment automatically. It is used for DWH in order to ensure that in evidence-based medicine users can only access those data which is granted to the role they have. The roles are given according to the set of authorisation rules defined by the institution's security policy.

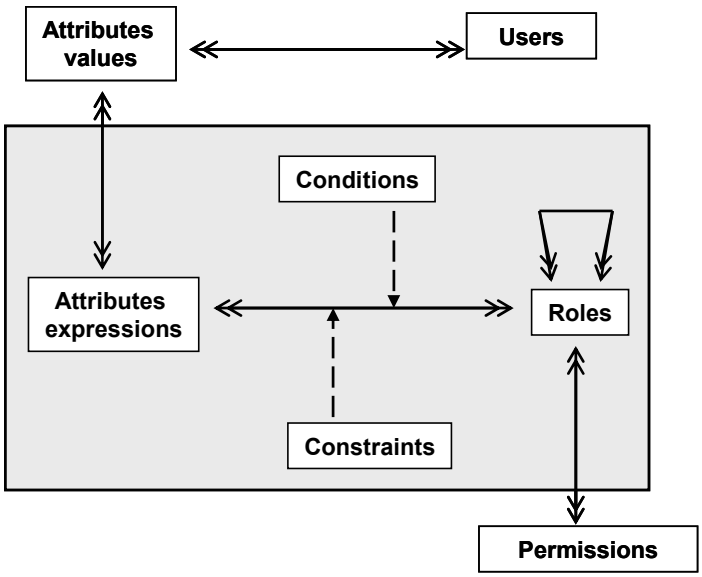


Figure 38: RB-RBAC Model [Al-Kahtani and Sandhu, 2000]

The main concept of the RBAC model is the role, which can be viewed as a semantic construct around which access control policy is formulated. Permissions are associated with the role to which users are

assigned based on factors such as their responsibilities and qualifications. RBAC model has emerged as a proven and superior alternative to traditional discretionary and mandatory access controls [Al-Kahtani and Sandhu, 2003] .

The role should be regarded as a job description regardless of the actor performing it. Roles should be assigned with exactly those authorisations that are needed to fulfil the duties of the job. Each user in the DWH should be assigned to at least one role, though multiple roles are allowed. A user can play only one role at a time and the roles can easily be re-assigned. This policy prevents authorisation conflicts among the roles of a user and it does not represent a limitation in real-life situations, as long as users can easily change their role according to the tasks to be performed. Since DWH offers read-only access, the role-based security model is also limited to this.

In [Al-Kahtani and Sandhu, 2000], authors modify RBAC such that it becomes rule-based and thus they refer to it as Rule-Based RBAC or RB-RBAC. In this model, an enterprise defines the set of rules that are triggered to automatically assign users to roles. These rules take into account:

- The attributes of the client that are expressed using attribute expression as defined by the language provided by the model.
- Any constraints on using roles.

Figure 38 shows that the users have many-to-many explicit relation with attribute values and thus, with attribute expressions. One user can have one or more attribute expressions depending on the information he provides. Further, two or more users may provide identical attribute expressions. A specific attribute expression corresponds to one or more roles. The roles may be hierarchically related to one or more roles.

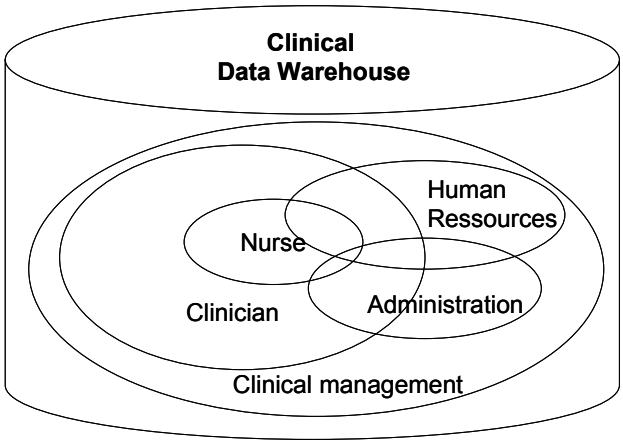


Figure 39: Role Based Access Rights in a Clinical DWH

A diagram representation of the data access rights for five different roles defined in a clinical DWH is given in Figure 39. Each one of the roles (nurses, clinicians, human resources manager, administrative clerk, and clinical management) has different data needs. Clinical management acts as a super-user

and has access to all data stored in the DWH. None of the users can identify the patient, since his (her) social security number has been pseudonymized and the rest of the data depersonalized.

8.4 Running Example: Security in Healthcare DWH Federation

In this subchapter, we will give an example to illustrate how the security measures described above should be realized in order to protect sensitive patient data stored in a medical federated DWH. Here, we will focus on the federation, built by the social insurance institutions and thus reuse our running example.

Social insurance companies are willing to deliver their data into the federation only if they trust the installed security measures. One of the main organisational concerns of such federation is to guarantee privacy and confidentiality of the insurant data and adequate reliable security measures for data access.

Companies participating in this federation come from the following lines of business:

- Health insurance
- Retirement pension insurance
- Contribution calculation

In our example, we will concentrate on the health insurance line of business. It concerns all active insurants and all retirees. Depending on their income, they are divided into two groups: the ones who are allowed to visit the physician of their choice and those who are supposed to visit the contracted physician.

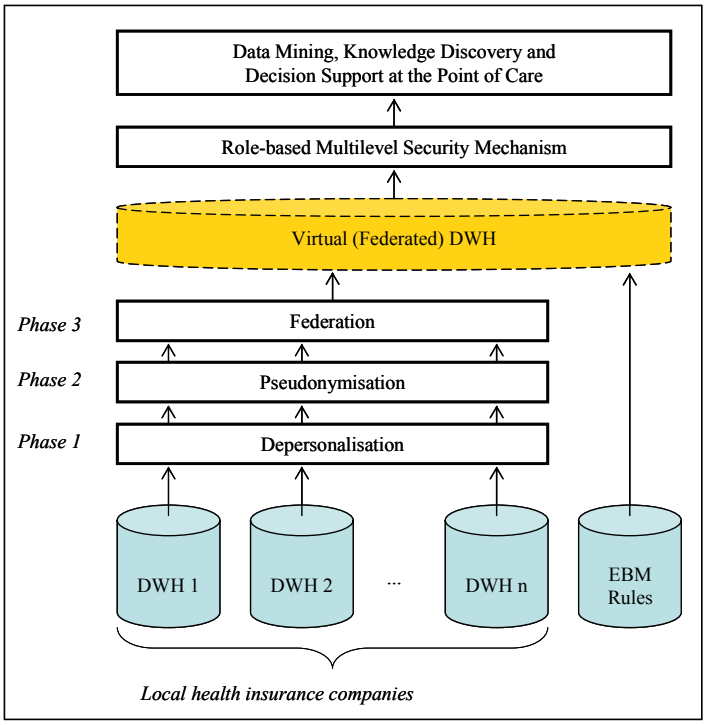


Figure 40: Phases of DWH Consolidation

As we already emphasized, the purpose of our federation is to unite data assets of local DWHs in order to gain a broader base for knowledge discovery and data mining. The final goal of the federation is to find the most effective therapies and treatments for given diseases, and thereby increase the insurant recovery rate and reduce health insurance costs.

In order to guarantee the security and confidentiality of the highly sensitive data in the new environment of federated DWH, the social insurance companies apply a three-phase consolidation process, as presented in Figure 40. After the depersonalization and pseudonymization procedures have been applied, the virtual (federated) DWH can be built. Subsequently, the role-based access model is implemented. It assures that only authorised users have access to the sensitive data. After putting the security mechanism in place, the use of a federated DWH facilitating EBM is possible.

8.4.1 Depersonalization in Healthcare DWH Federation

In the depersonalization phase, we first recognise sensitive data, which need to be grouped, hidden or removed. Experience has shown that it is much better to hide or remove sensitive data than to provide the users with null-values, since that motivates users' curiosity and presents security challenge.

While creating the conceptual data model of the federated DWH, business users (clinical management) specify sensitivity levels of data. Table 12 shows a part of such a sensitivity level list.

Table	*Attribute*	*Sensitivity level*	*Depersonalization measure*
Patient	Name	Very High	Remove: very sensitive data, not supposed to be seen by anyone.
Patient	Date of birth	Middle	Group: Create new attribute "age" and group patients into following groups: 0-10, 11-20, 21-30,…
Patient	Gender	Low	None, accessible by all users
Patient	Profession	Middle	Group professions into: employee, artist, manufacturer, health professional etc.
Patient	Degree	High	Hide: highly sensitive data, may be seen only by authorised users
Address	Street	Very High	Remove
Address	Postal code	Very High	Remove
Address	City	Middle	Group: Create new attribute "region" and group cities geographically (i.e. Baden, St.Pölten, Wr. Neustadt = Lower Austria)
Tax Data	Tax number	Very High	Remove
Tax Data	Local tax office	High	Hide

Table 12: Depersonalization of Sensitive Data

In this stage of the federated DWH design, only data relevant for further analyses and reporting are considered. Irrelevant data (i.e. entities: Bank_Details, Account_Details) are not extracted from the source DWH.

The data modeller incorporates the specified privacy restrictions into the resulting logical data model. Figure 41 shows the logical data model of the depersonalized data. Key attributes, social security number, and relative's social security number, will undergo a pseudonymization process in the next step. They have no relevance for querying tasks in the federated DWH and are hence crossed out in the model. Attributes with the sensitivity level *very high* (i.e.: name, street, postal code, tax number) can easily be misused to identify the patient and are therefore removed from the federated DWH model. *Highly sensitive* data (i.e.: degree, local tax office) are hidden and can only be seen by authorised users. In Figure 41, these attributes are written inside curly braces.

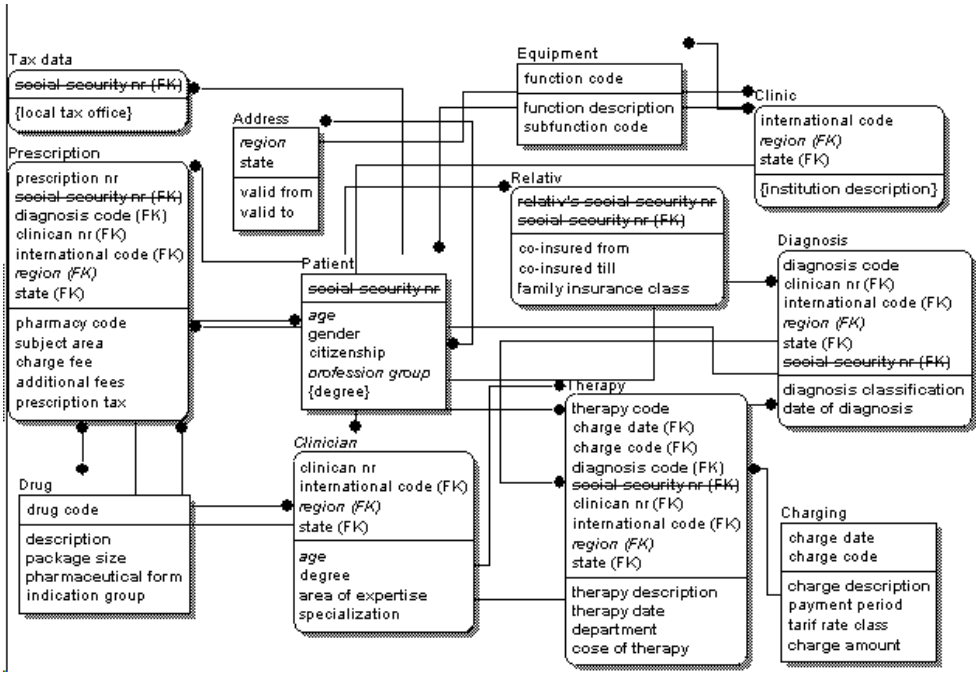


Figure 41: Logical Data Model of Depersonalized Data in a Component Warehouse

The attributes with sensitivity level *middle* (i.e.: age, profession group, region) are transformed in order to protect their sensitiveness. The attributes with *low sensitivity* present no security risk, and can be accessed by all users.

8.4.2 Pseudonymization in Healthcare DWH Federation

The pseudonymization procedure can be performed after completing the depersonalization process. Since no re-identification of a patient is needed to reach the intended goals, one-way pseudonymization procedure is applied. According to the Austrian law [Bundesgesetz, 2004], pseudonymization must be performed by a trusted third-party organisation.

Figure 42 represents the privacy preserving measures during query processing in the federated DWH. A user query is submitted to the federated DWH and then reassembled into particular sub-queries for each of the underlying DWHs, according to its local logical data model. The query mapping is realised in the federated DWH. Each underlying local DWH receives and subsequently processes its corresponding query. The resulting answer to a partial query consists of three main data parts:

- **SSN** – **S**ocial **S**ecurity **N**umber, which is used as unique patient identifier
- **PD** – sensitive **P**ersonal **D**ata, to which the user access is restricted
- **HCD** – **H**ealth **C**are **D**ata, which is non-sensitive medical data

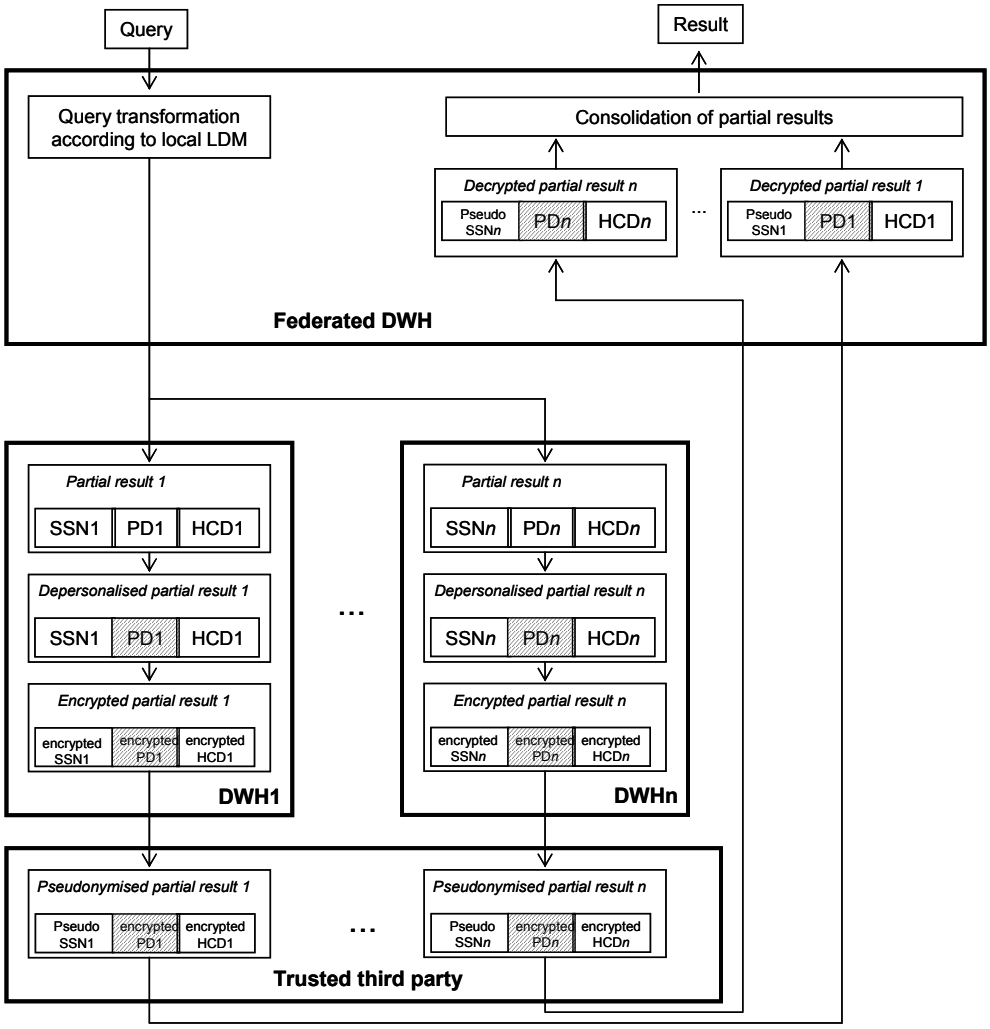


Figure 42: Privacy Preserving for Query Processing in Federated DWH

The query result undergoes a data depersonalization process in its originating DWH. Here, all sensitive personal data (PD) are grouped or concealed, so that they cannot be used for patient identification.

The result of a depersonalized partial query is encrypted to enable a secured transportation to the pseudonymization service.

Pseudonymization is performed by a trusted third party. Since a different pseudonymization key is used for each pseudonymization, it is not possible to re-identify individual patients.

Once received by federated DWH management system, all partial query results are decrypted and consolidated into a single result query, which is delivered to the user of the federated DWH.

8.4.3 Role Based Access Control Mechanism in Healthcare DWH federation

We will use an access control matrix to describe the access rights of users accessing our federated DWH. Since "read" is the only kind of data access in this environment, the represented matrix is modified Lampson's access control matrix [Lampson, 1974]. Each row represents a specific role whereas each column contains a federation table or view (Table 13).

Subject/ Object	Patient	Clinician	Therapy	Charging	Diagnosis	Drug	Prescription	Tax data
Clinical managem.	x	x	x	x	x	x	x	x
Clinician	x	x	x	--	x	x	--	--
Nurse	x	x	x	--	--	x	--	--
HR	--	x	x	--	--	--	--	--
Administra-tion	x	--	--	x	--	--	x	x

Table 13: Access Control Matrix for the Healthcare DWH Federation

According to the access control matrix, clinical management is a super-user who is allowed to see all data contained in the federation. Clinicians can access data related to the patients, therapies and drugs received, clinicians who where in charge, and their diagnoses. All financial data are inaccessible to the clinicians. Nurses have similar authorities, with the restriction that they can not see the diagnosis data. Human resources (HR) are assigned the read rights for the tables containing data about the clinicians and the therapy they prescribed, whereas clinical administration, which is in charge of the financial tasks, has insight into the patient, charging, prescription, and tax data tables and views.

In system with large number of users, tables, and views, like the federated DWH is expected to be, the access control matrix could become very large and sparsely populated. As such, an access control matrix is rarely implemented as a matrix but is almost always implemented as a representation of the matrix, for instance ACLs and capability lists [Ferraiolo et al., 2003].

Part III

Application Scenarios

9 Federated DWH to Support the Interoperability of National Healthcare IS in Austria

A concisely defined information network for exchange of sensitive patient's data is a significant step towards enabling national and international cooperation. In this chapter, we present technical and organizational aspects for creation of Austrian healthcare information network. We offer an application scenario for the federation of Austrian health insurance DWHs in compliance with the national governing framework for electronic exchange of patient-related data. This example will illustrate the benefits of standardized, identifiable and secure consolidation of the local DWH and promote innovation in the public sector.

In Austria, both commercial institutions and governmental bodies are interested in building effective and sustainable healthcare system. Increasing patient expectations and rising budgetary pressure are forcing governmental health authorities to rapidly improve the quality of communication between individual health system participants. Since proper management of public health on both national and international level can be executed only on the basis of comprehensive clinical data, building of exhaustive, standardized, electronic patient lifetime records is a necessary step. More comparable health data can help national health authorities to collaborate, for example to gain productivity by supporting reimbursement procedures or to tackle communicable diseases.

In this chapter, we give an application scenario which represents MAGDA-LENA guided electronic data exchange between the Austrian social insurance DWH on one side and the DWH federation on the other. This scenario incorporates the themes which were handled in previous chapters: federated DWH model, semantic integration, deployment of internationally adopted communication standards for message exchange, and legal regulations for the creation of a healthcare information network.

9.1 Organizational Aspects for Creation of Austrian Healthcare Information Network

The Commission of the European Communities (2004) announced in their Action plan for a European e-Health Area that achieving a seamless exchange of health information across Europe requires common structures and ontologies of information transferred between health information systems. Member states, as well as other countries concerned with raising the quality of healthcare through integration of medical information systems, support initiatives aimed at building of electronic healthcare networks.

The Austrian healthcare system is highly fragmented. Clinics, social insurances, and other medical institutions store their patient data in non-standardized, system-specific formats. In order to make information exchange between healthcare service providers possible, the reformation of the information system is a necessary step.

Austrian social insurance service providers are well aware of the benefits they can gain from an integrated healthcare information system. A large data base, resulting from this integration, would represent a precious knowledge source. A decision support system and the deployment of evidence-based medicine would enable social insurance institutions to locate efficient ways of enhancing the quality of health care provided to the insurants, with the possibility of reducing treatment costs at the same time.

The participants in the application scenario described in this chapter are Austrian social insurance providers. Other organisations in the healthcare domain can join the federation and contribute to the development of the integrated healthcare network, which would be the foundation for e-prescription, drug cost control, patient's electronic lifetime record and further e-Health projects for enhancing the quality of medical care. This scenario is guided by the regulations provided by the commission designated by the Austrian ministry of health. A short overview of the major recommendations from this framework is given in the next subchapter.

9.2 MAGDA-LENA Framework for Electronic Exchange of Patient Data

In 1995, the Austrian Ministry of Health and Woman Affairs created the STRING commission to advise the minister on all strategic healthcare issues. Three years later, this commission developed the MAGDA-LENA framework as the governing framework for electronic exchange of patient-related data in Austria. Unlike some other frameworks, which are formulated as legally binding rules (for example HIPAA in USA), MAGDA-LENA [STRING-Kommission, 2000] has solely a recommendatory status. MAGDA-LENA outlines the technical and organizational aspects for the

creation of an Austrian healthcare information network, which facilitates the exchange of patient-related data between healthcare providers and social security institutions.
The MAGDA-LENA framework contains the following four sections:

1. Message contents, models, standards
2. Identification variables
3. Data privacy and security
4. Network providers and nodes

In the following subsections, we present the concise overview of the main recommendations as outlined by [Duftschmid et al., 2003].

9.2.1 Message Contents, Models, and Standards

The standard formats for exchange of health information are specified. Messages exchanged between communicating parties should not be defined by parties themselves, but are based on internationally adopted standard message formats. The development of a healthcare information model as a basis for standard messages remains as a subject of international efforts. Since for the time being such a system is still in a planning phase, MAGDA-LENA prescribes the use of existing international and national standards related to medical informatics, which satisfy the particular requirements of a communication process. In case that such standards do not exist for a given area of application, messages are developed along the procedural model defined in MAGDA-LENA to optimize their uniformity.

9.2.2 Identification Variables

In order to eliminate any kind of data abuse during the transmission process, the communicating parties, as well as the transmitted data, must be uniquely identified. The authentication procedure for the communicating parties is briefly described here. MAGDA-LENA requires communication parties to be identifiable via registered directories within the organization they are assigned to. Furthermore, the roles of the communication parties as well as the identification of transmitted data are reviewed here.

9.2.3 Data Privacy and Security

Communicating parties are required to implement certain security measures in order to guard data privacy and security in their internal domain as well as during electronic communication with others. For this purpose, a list of obligatory implementation specifications (e.g. content encryption algorithms and protocols, requirements for electronic signature, security tokens, and password systems) and guidelines for their proper implementation is provided.

9.2.4 Network Providers and Nodes

Minimum standards that have to be obeyed by network providers participating in the Austrian healthcare network are defined. They include guidelines regarding contracts with clients and guidelines concerning network interfaces.

9.3 Building a Federated DWH According to the MAGDA-LENA Recommendations

In previous chapters, we handled the relevant security issues for federated DWH in the area of EBM introducing healthcare standards and in particular HL7 RIM Foundation Classes. The goal of our investigation was focused on the technical issues of gathering local DWHs into a federation. In this chapter, we intend to adopt the proposed federated DWH model corresponding to the MAGDA-LENA recommendations by broadening our focus to accommodate the necessary legal restrictions.

Austrian social insurance institutions operate heterogeneous DWHs, which are functionally integrated into a single unit. Each local DWH is providing the federation with the designated data, according to the MAGDA-LENA recommendations.

Figure 43 shows an extract from the multidimensional logical data model as it is stored in one of the social insurance DWHs. It contains data about the patients' demographic data (Patient), address (Location), diagnosis, prescribed drugs, social insurance benefits, and Patient Dimension, Time, hospitalisation length, and costs (Facts), and is created according to ADAPT™ [Symmetry_Corporation, 2007] notation.

9.3.1 HL 7 RIM for Standardized Data Transmission

Since each local DWH stores its data in its own system-specific format, large portions of different data contents are semantically closely related. The *insurant* is uniquely identified by his (her) social security number. Information about insurant's *status* (self-insured, co-insured), scope of insurance, as well as various descriptive fields are stored here. Additionally, these DWHs hold data about *consumed services* (treatments, therapies, drugs) covered by the social insurance. For the purposes of compensation calculation for the insurant, the whole range of *accounting data* is loaded into the DWH. Data about drug *dosage*, as well as *charging data* for pharmacy contribution calculation, is another important part of it.

According to the MAGDA-LENA recommendation about message content, models and standards, underlying DWHs have to use internationally adopted standard message formats for transmission of their health information to the federation. In chapter 6 we have explained why messaging standards are particularly important for defining how information is packaged and communicated from one party to

another. For our application scenario we choose the HL7 V3 communication standard since it is general enough to allow for the necessary level of flexibility.

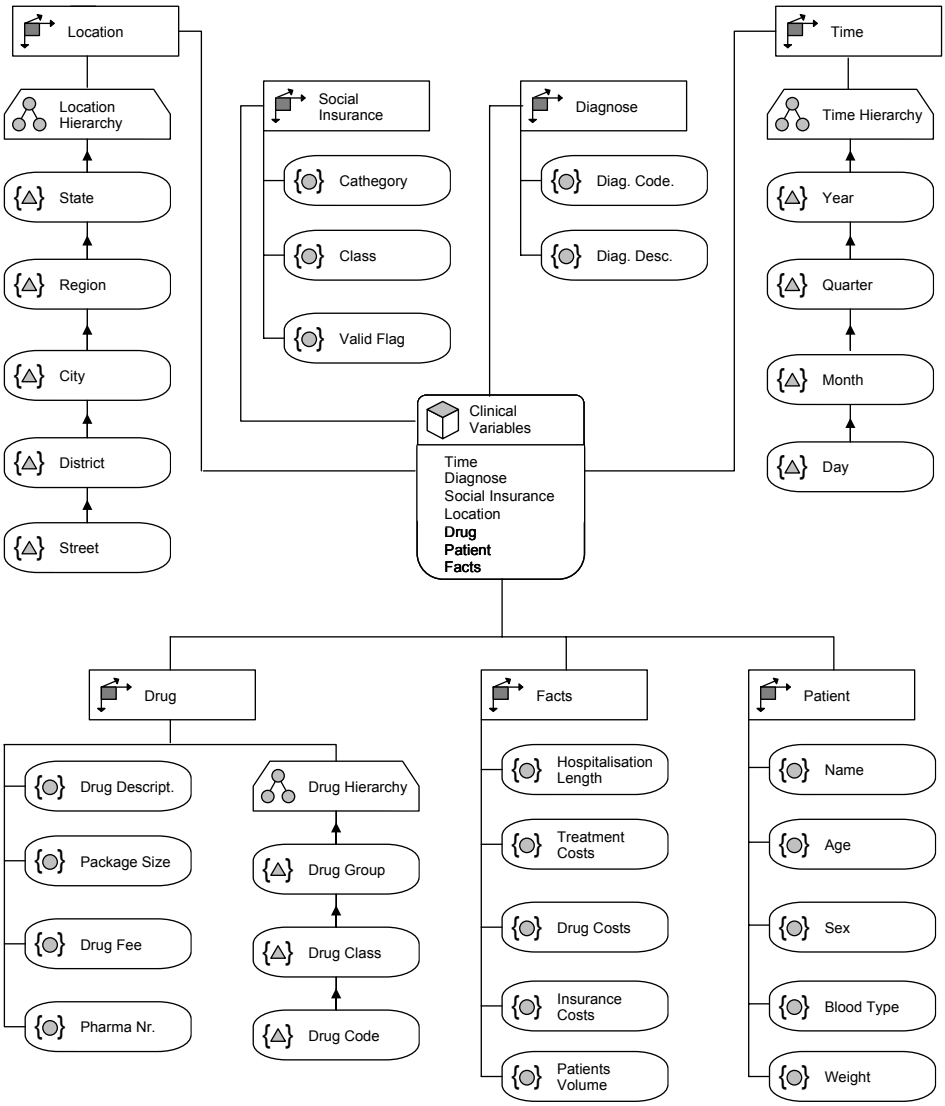


Figure 43: Multidimensional Social Insurance LDM

Figure 44 shows an example representing HL7 RIM model for the treatment charging process taking place in a social insurance company. On one side, we see two person e*ntities* in the *roles* of clinician and patient. They p*articipate* in the *activity* treatment as, respectively, performer and subject. On the other side is the entity person, in the role of a payer, subject of the billing activity. Another person entity participates in the billing activity as a performer in the role of an accountant. This role is scoped by the entity controlling. These two activities (treatment and billing) are related to each other, which is represented through the *act relationship* treatment charging.

Using HL7 RIM message formatting for transmission of healthcare data between local DWH and the federation fulfils the first requirement of the MAGDA-LENA framework, as described in subchapter 9.2.1.

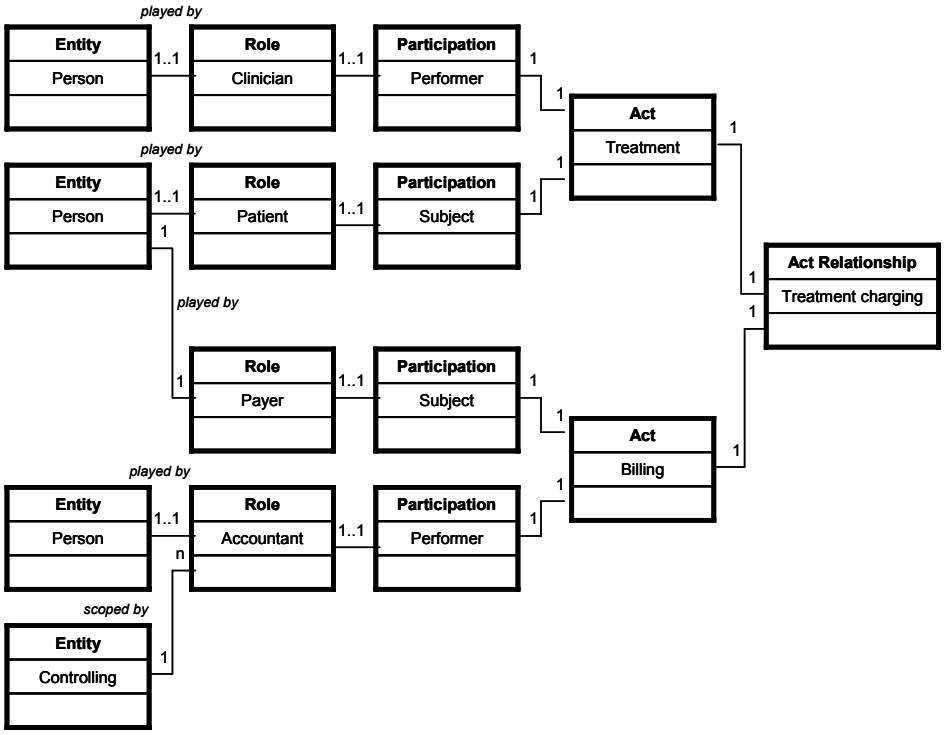


Figure 44: HL7 RIM Class Diagram for Treatment Charging Process

9.3.2 Ensuring the uniqueness of instance-identifiers

According to the MAGDA-LENA framework, communicating parties, as well as the transmitted data, must be uniquely identified in order to prevent any kind of data abuse during transmission process. Each participant of the healthcare network must apply for access at the designated authority. Additionally, a participant is assigned a pre-defined role. He (she) is then accountable for addressing the messages to authorized receivers only. Whenever patient's data is transmitted, it needs to be identified by a unique patient-id. Finally, transmitted documents have to bear a unique document identifier.

Table 14 represents the items that need to be identified and the corresponding identification method. In our case study, *communication parties* are social insurance institutions participating in the federation. For the time being, their identification method is still undefined, but we will assume that there is a specific authority responsible for assignment of the communicating party's ID. The *role* of the communicating party is specified by the Ministry of Health. At present, social insurance number (SVNR) is the only suitable candidate for the purpose of the unique *patient's identifier*. Unique

identification of transmitted *documents* still needs to be specified. We will explain this issue in detail next.

Object of identification		Identification method
Communicating party	Physician	Contracting party number provided by Hauptverband or physician number provided by the Medical Association (Ärztekammer)
	Pharmacy	Pharmacy operating number
	Hospital	Hospital number
	Social Insurance Institution	Unknown
Role of Communicating Party		Assigned by Ministry of health
Patient		Social insurance number
Document		Document ID

Table 14: Identification List

In our application scenario, we follow the HL7 V3 RIM approach for identification and classification as it has been described by [Heitmann, 2006]. Heitmann illustrates these procedures through an example, which presents the *Observation* class. (Figure 45 represents the XML statement of the observation.) Here, a *code* is used to state the art of the examination and *value* is used for storing the actual result. *Effective time* contains the exact time of the observation and the observation is uniquely identified through *id*. All attributes have pre-assigned and well defined data types.

```
<Observation>
   <id          extension="38e4748"
                root="1.2.276.0.76.3.67.982"/>
   <code
                code="8459-0"
                codeSystem="2.16.840.1.113883.6.1"/>
   <effectiveTime value="20060214"/>
   <value value=120" unit="mm[Hg]"/>
</Observation>
```

Figure 45: XML Representation of the Observation [Heitmann, 2006]

In this example, the specific observation (e.g. systolic blood pressure) is characterized as an instance of an observation (id is an instance identifier.). For such purposes, the application system must create a globally unique instance-identification. This can be assured by an HL7 procedure. Element *code* bears the classification of the observation. HL7 uses common medicine coding system LOINC [Regenstrief_Institute, 2007]. The actual code is stored in the XML-attribute *code* and *codeSystem* contains the unique identifier of the coding system. When electronic data exchange takes place, the concept of HL7 requires a world-wide unique identification of all medical observations. HL7 provides a procedure to ensure that, but the hosting DWH must be upgraded in order to support the production of these references.

9.3.3 Securing confidential patient data during message exchange

Following security precautions are recommended by MAGDA-LENA:

- Password identification for the user of the federated DWH (authentication)
- Any data modification must bear a digital signature
- Tracking of data manipulation through log files
- Confidential health data must be transmitted in encrypted form
- Transmission confirmation upon receiving of confidential data
- A role-based access model has to be implemented

The last section of MAGDA-LENA recommendations concerns network providers and nodes. Since this subject is not DWH specific, we will not handle it here.

9.4 Overall Federated DWH Model for Austrian Social Insurance Providers

Figure 46 represents the model of the DWH federation process according to MAGDA-LENA recommendations. Underlying social insurance DWHs are providing the federation with requested data in their native XML format. Using XSLT transformation, retrieved data is transformed from native into HL7 V3 XML schema. Further, identification and coding procedures round off the desired HL7 V3 XML structure. For security reasons, the confidential patient data is encrypted for transmission.

The acceptance of the transmitted data is acknowledged by a confirmation message. The essential part of integration of logical schemas of the underlying DWHs is the semantic integration layer. Our model includes wrappers and mediator, which are two main architectural components of a mediated query system. Wrappers encapsulate local data sources and export their functionalities and the metadata stored therein. They accept queries in a certain language and return metadata in a united form. The mediator handles global queries from the application layer, unfolds them into sub-queries and disperses these sub-queries to the relevant local data sources via their wrappers [Haslhofer, 2006]. The local results will be returned from wrappers; the mediator finally combines and presents the result to the client. The result (consolidated HL7 V3 message) is subsequently converted into XML or directly into SQL.

Only authorized users may access the federated DWH. The Ministry of health is responsible for access and role assignment for all participants of this healthcare network. According to the role-based access

control model, each user may see only those portions of data necessary for performing the tasks of his (her) role.

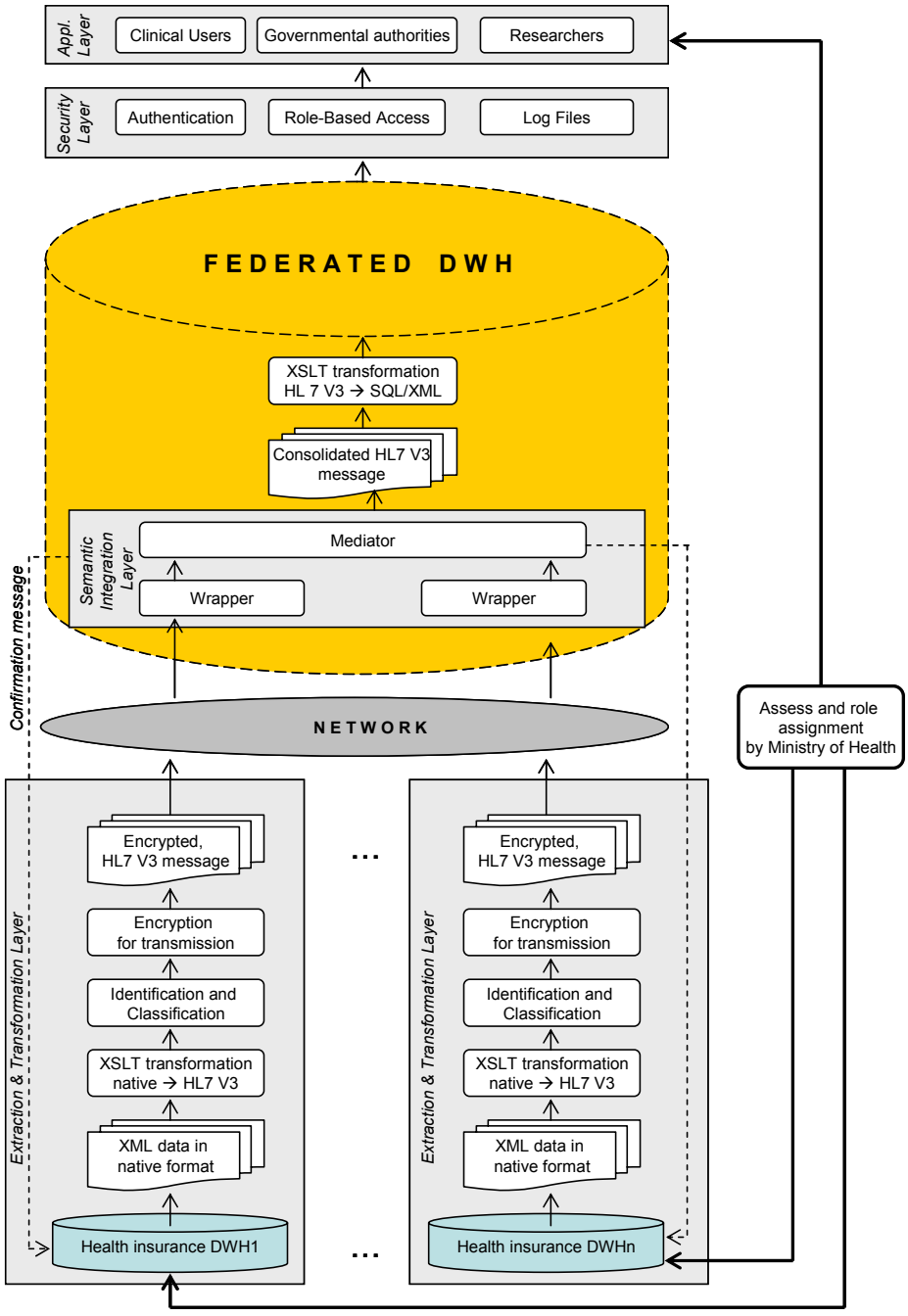


Figure 46: DWH Federation According to MAGDA-LENA Recommendations

Authentication procedures are put in place and all data manipulation is logged for later analysis. Users are querying federated DWH through data mining and OLAP-Tools by applying predefined reports or performing ad-hoc analysis.

10 Federated DWH Support of Real-Time Sense and Response for Effective Healthcare

In this chapter, we explore the idea of conceptualisation of DWH federation as a foundation for sustainable, appropriate, and legitimate healthcare DSS, which includes sense and response mechanism and facilitates EBM in order to primarily support the patient monitoring process. This mechanism includes a new diagnostic technology into decision-making process and aims to improve the healthcare business process through enhanced automation of decision-making. The role of highly trained practitioner is to oversee this automated system, as no manual interaction is needed. A scenario illustrates a possible application field in the area of emergency and intensive care.

In the coming decades, the costs of treatment of chronic diseases of an aging population are going to grow massively. By contrast, the working population is not going to develop proportionally, hence the number of people consuming the healthcare will be much greater than the number of those who will be paying for it. One efficient way of tackling the enormous hospitalisation and treatment costs will be monitoring of patients at their homes as long as possible, rather than admitting them into hospitals or nursing homes.

At this time, there are several monitoring and telemonitoring solutions available. However, they are not taking advantage of the newest medical findings and statistically proven, evidence-based knowledge combined with patient's health record in almost real-time manner. In this chapter, we will propose a novel approach in which a sense and response mechanism will be deployed to monitor patient's condition, using the sensored data as input when querying the federated DWH (for accurate patient's health record) including evidence-based rules, for prompt finding of an appropriate treatment under given circumstances.

10.1 Remote Monitoring Systems

The early beginnings of remote telecare started years ago with two-way video as a means of providing remote visits. Caused by growing need for disease management in recent years, patients monitoring has achieved growing acceptance. Modern remote monitoring systems interact more substantially with patients and support each patient's specific needs. The application of such a system enables early intervention through monitoring and thus improves clinical outcomes.

Homecare monitoring systems have been researched for several years and there are already a few commercial solutions and business models. The existing solutions, however, often rely on medical call centres, transmit their data through telephones and are based on proprietary technology. One of the first established homecare monitoring system is well@home [Patient Care Technologies®], which is actually dubbed a "clinical management system" used by patients at home. Embedded measurements of blood pressure, pulse, oxygen saturation, temperature, ECG, and respiration rate are taken through wired sensors and transmitted using a built-in modem. A physician reviews the data and returns a treatment plan that patients can see on a large touch screen that is part of the well@home device [Hein et al., 2006] .

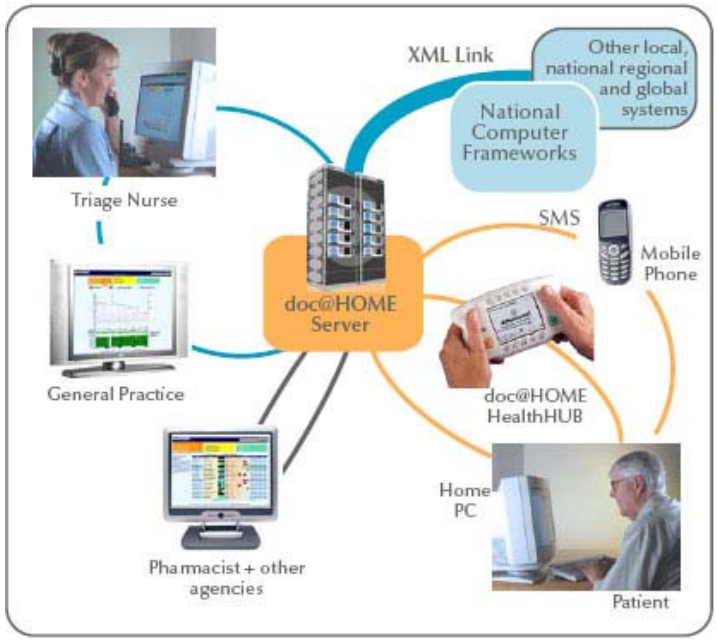


Figure 47: doc@HOME Service [Docobo®, 2006]

Another integrated telehealth solution for remote management of patients with a range of Long Term Conditions is the doc@Home (Figure 47). The doc@HOME service is designed to provide a systematic approach to management of chronic disease at home and other locations remote from the clinician's office. Patient/Clinician interaction is typically via the Docobo HealthHub, a fit for purpose, robust, handheld data collection unit which connects through a standard telephone line at the

patient's home to a secure server. Healthcare Professional interaction with the doc@HOME service is via secure Web access using standard browsers, enabling patient management at a range of locations. Changes in patient trends can be identified and the authorised user notified. An authorised clinician can access the patient record on demand and send messages directly to the patient, for example, a request to visit the surgery or to change the frequency and/ or volume of medication [Docobo®, 2006]. As stated in [eHealth, 2006b], improving "Management of Health Risks" and Patient Safety is one of the cornerstones to assure policy support for the European Commission FP7 eHealth research topics. Emergency response systems are one of the significant factors of safe healthcare management. There are therefore several proposed prototypes and developed projects on this topic. Authors in [Hashmi et al., 2005] propose a scalable emergency medical response system that couples the efficient data collection of sensor networks with the flexibility and interoperability of a web services architecture. [Moron et al., 2005] present a wireless medical monitoring system which is capable of receiving and processing the pulse-oximetry signals from one or several monitored patients. SAPHIRE [Hein et al., 2006] is an Intelligent Healthcare Monitoring for the Homecare Scenario using multi-service architecture. Akogrimo E-Health [Loos, 2006] deployed Heart Monitoring and Emergency Scenario on the Mobile Grids. CodeBlue [Malan et al., 2004] proposes a wireless infrastructure intended for deployment in emergency medical care, integrating low-power, wireless vital sign sensors, PDAs, and PC-class systems.

In general, the existing approaches focus on protocols for data collection from sensor network and the communication between components in order to process data, mostly using agent technology, and to deploy the presentation data in mobility devices. These approaches thus could not deal with the ad-hoc analysis to recognize the pattern-based data trends. With the use of federated DWH facilitating EBM, combined with the Sense & Response loop, our approach could process the ad-hoc query in the acceptable response time.

10.2 Sense and Response Mechanism in Healthcare

The Sense & Response mechanism [Nguyen et al., 2005] was initially developed to help organizations to monitor the IT events in business processes in order to proactively respond to business situations or exceptions with minimal latency. The data processing is controlled by the "Sense & Response loops" which are able to continuously receive, process, and augment events from various source systems, and transform these events in near real-time into performance indicators and intelligent business actions. In this subchapter, we apply the "Sense & Response" mechanism in healthcare systems with the aim of real-time monitoring of the heath condition of high-risk patients to react instantly in case of emergency. The "Sense & Response loops" contain 5 stages as described in Figure 48.

In the *sense stage*, events (i.e. sensors data from the patients, blood examination results etc) are continuously captured and transmitted in a unified format. The *interpret stage* transforms raw event data into a heath care indicator (such as heart vibration, percentage of oxygen in blood and so on). In the a*nalyse stage*, the heath information will be analysed to determine the root cause and predict the next states. This process acquires the historical data from the federated DWH to analytically conduct a pattern. Depending on the analytical result, the *decide stage* proposes the best option for improving the current patient's health condition and determines the most appropriate action for a response to the situation. Finally, the real actions (i.e. a call to the doctor or ambulance, sending an alarm or messages, etc.) are conducted in the *response stage*.

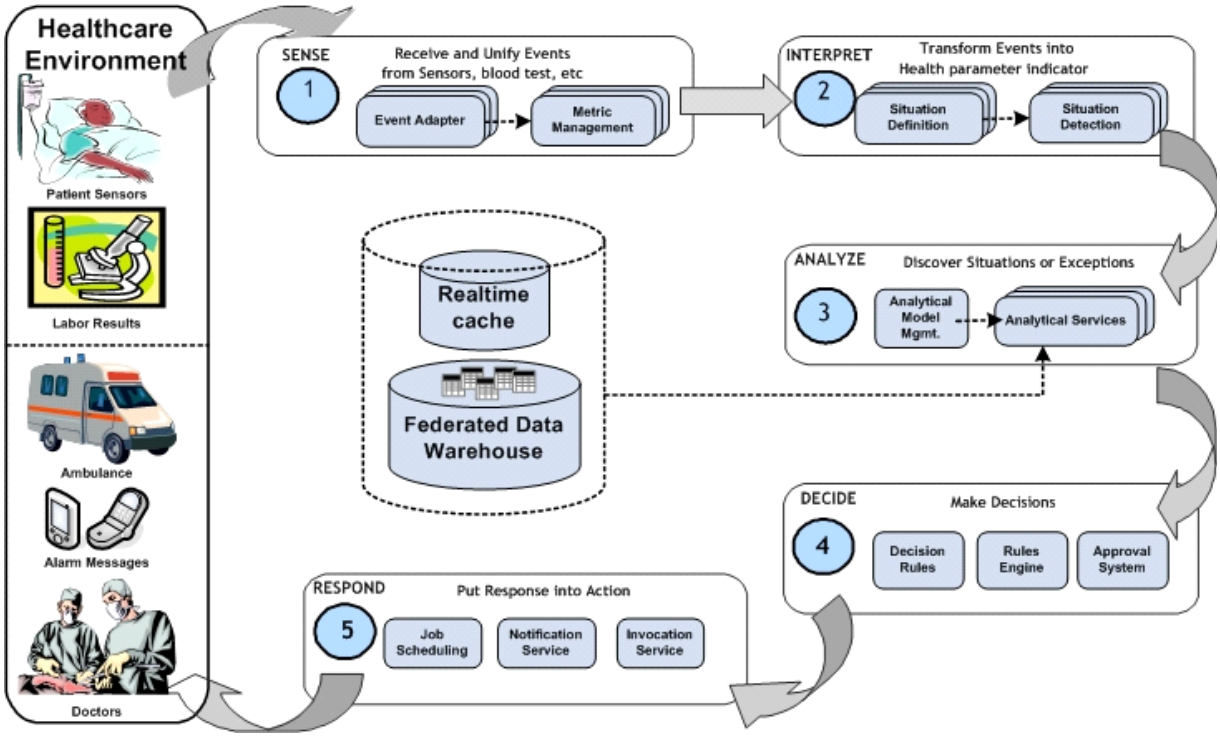


Figure 48: Sense & Response Mechanism in Healthcare Environment

The Sense & Response architecture was realized as a pool of services (system services and Sense & Response services) and the underlying infrastructure enabling a robust communication and interaction between them. Each stage of the Sense & Response loop is supported by special services which can flexibly interact with each other via the Service Bus. The Sense stage as an example may include the Event Transformation services containing the Event Adapters which receive events from the sensors and transform them into a standardized format (e.g. XML or HL7). With this open architecture we can include user defined services for various tasks such as discovering situations, a third-party analysis tool as an analytical service or an external rule engine for making automated decisions in the Sense & Response loops.

The strength of our approach lies in the ability of federated DWH system to receive the monitored patient data, gather all historical data about the patient's condition and combine them with the guidelines coming from evidence-based medicine in near real-time. Only this way it is possible to offer the best-fitting treatment to the patient, reduce treatment costs and preserve data privacy.
In the next subchapter, we will illustrate our approach through an example, where we apply the federated DWH facilitating EBM in a real-time sense and response environment of an emergency room. (Similar approach is applicable in the home-care environment.)

10.3 DWH Application in Real-Time Sense and Response Environment for Support of Emergency Room Decision-Making

Recent advances in treatment of acute coronary syndromes have raised awareness that the prompt reaction to chest pain presentation may be life saving. Most patients presenting with chest discomfort have a non-ischemic electrocardiogram on presentation, but are routinely admitted to hospital because of diagnostic uncertainty for occult myocardial infarction (MI) or ischemia [Gorenberg et al., 2005]. We will show how the federated DWH facilitating EBM and integrating the newly developed device for non-invasive detection of ischemia can support decision-making process at the point of care to rule out the heart attack.

10.3.1 Non-Invasive Device for Measurement of Central Aortic Pressure Changes

The decision to discharge patients with chest pain from non-cardiac origin from the emergency room (ER) is a challenging task. There are many uncertainties and legal aspects that make this decision difficult. One of the problems with chest pain patients arriving to ER is that diagnosis is deferred for at least 6 hours when the enzymatic markers become diagnostic. An early immediate diagnosis of chest pain of non-cardiac origin may be of great help for the clinician to avoid unnecessary admissions. In [Gorenberg et al., 2005], the authors describe their new non-invasive device for measurement of central aortic pressure changes. The ascending limb of the aortic pressure (dP/dtejc) was measured with a newly designed computer controlled device (*CardioWatch Ltd., Matam Advanced Technology Center, Haifa, Israel*), shown in Figure 49.

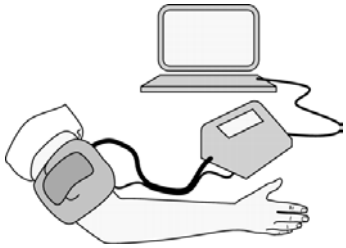


Figure 49: non-invasive device for measurement of central aortic pressure changes [Gorenberg et al., 2005]

This device consists of three components: (1) a sphygmonometric arm cuff attached to an air pressure unit, (2) an array of proprietary sensors attached to the arm at the antecubital space over the brachial artery, and (3) a computerised monitoring system. Figure 50 presents the detailed view of the new sensor system and artery with the collapsed region under the cuff. Upstream pressure is systolic.

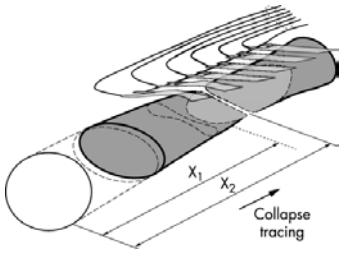


Figure 50: Detailed View of New Sensor System and Artery [Gorenberg et al., 2005]

According to [Gorenberg et al., 2005], all patients with dP/dtejc ≥ 150 were discharged after at least 1 day of admission and MI was not identified in any patient. Using this criterion, nearly 40% of patients presenting with acute chest pain could be spared the risks and costs of unnecessary admission to hospital and more invasive cardiac testing by simply adding an easy to use, immediately obtained, non-invasive test to the diagnostic protocol, and using a threshold of $dP/dt_{ejc} \geq 150$ to rule out heart attack. However, this method has limitations concerning patients with previous MI and severe arrhythmias and patients with severe hypotension.

At this stage of the examination process, when the physicians can not rely completely on the results of the examination devices, we suggest the application of the federated DWH facilitating EBM and incorporating sense-and-response mechanism. For the patients with previous MI and severe arrhythmias and patients with severe hypotension, patient's EHR needs to be checked every time the observation is carried out, in order to prevent erroneous diagnoses.

10.3.2 DWH Support for Diagnosis

At the admission into the ER, the device described above will be applied first. This examination takes only 10 minutes and it can be used for first quick exclusion test. According to [Gorenberg et al., 2005], patients with $dP/dt_{ejc} < 150$ presumably had suffered a heart attack and will be hospitalized for further examinations. For all other patients, their medical history needs to be checked, before they can be discharged. Here, DWH will be checked for the first time: The Sense & Response Loop 1 queries the federated DWH to verify if the particular patient has suffered from MI, severe arrhythmias or severe hypotension before. Patients whose EHR contains such entries are susceptible to heart attack and will also be hospitalized for more detailed observations. All other patients can be released from the ER.

10.3.3 DWH Support for Building EHR from Heterogeneous Data Sources

In order to create EHR for a particular patient, medical history data is gathered into the federated DWH from all available sources (point 2 in Figure 51). This record contains data about the patient's health condition, possible chronic diseases or allergies, diseases he (she) has suffered from in the past, treatments and drugs prescribed, as well as social insurance data. Since data collection from distributed data sources can be time-consuming and decision making in such urgent cases is time critical mission, caching mechanisms for speeding up the querying process can be considered.

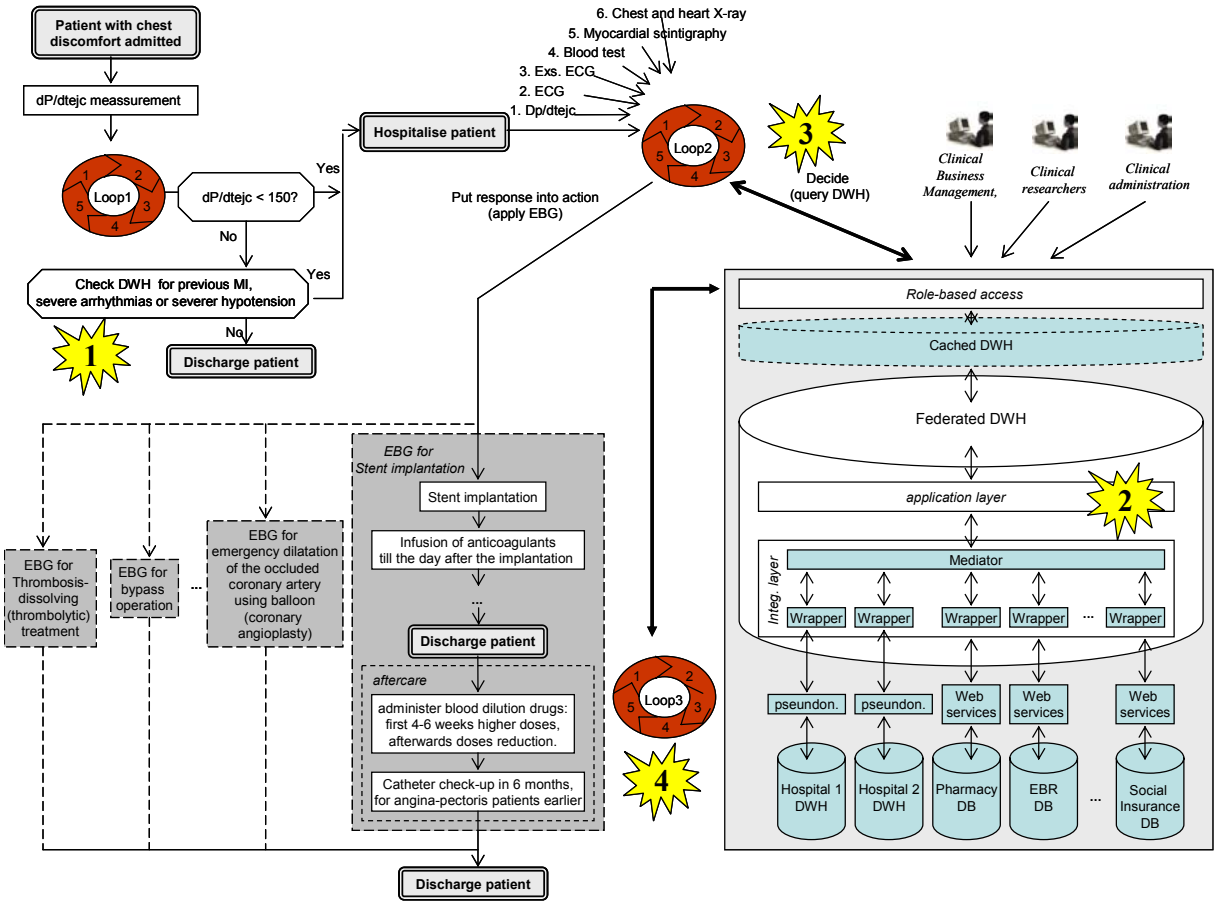


Figure 51: Decision Making Process for Chest-Pain Patients

10.3.4 DWH Support for Finding the Appropriate EBG

Hospitalized patients consequently undergo more detailed examinations like electrocardiogram (ECG), exercise ECG, blood test, and in some cases even myocardial scintigraphy or chest and heart X-ray. The results of these examinations are input parameters for the second Sense & Response Loop. At this point (Figure 51, point 3) EHR (old data) is merged with the examination findings (new data). The result data is used as input for finding the best fitting treatment for a given patient and a given disease.

The statistical excellence of DWH can be fully exploited in this stage. Predefined analytical procedures automatically search through the EBG database, drill-down and roll-up as well as slice and dice functions are deployed in order to detect the most effective treatments. Usually, more than one treatment is offered to the physician in charge. In our scenario, these are bypass operation, coronary angioplasty, stent implementation, thrombolytic treatment etc. At this stage, physician has to interact with the system in order to decide which treatment will be the best one to apply.

10.3.5 DWH Support for Tracing the Recovery Trend

In the next step, the chosen treatment (in our scenario, it is the stent implementation) is applied. According to the selected EBG, he (she) will receive the anticoagulants intravenously till the day after the implementation. After that, diverse tests are made before the patient can be released from the hospital.

According to the EBG, in the first 4-6 weeks patient receives higher doses of blood dilution drugs. Afterwards, the dose is reduced. A catheter check-up is designated 6 months after the intervention, except for the angina-pectoris patients, who are scheduled for an earlier examination. During the aftercare time, DWH is deployed for checking of recovery process (Figure 51, point 4). The health condition of the high-risk patients especially is regularly checked against the available experience (history) values stored inside the federated DWH (Sense & Response Loop 3). With the support of large volumes of historical data residing in the DWH and powerful statistical tools available, an alert is automatically issued whenever a deviation of recovery data is noticed.

11 Conclusion

In this chapter, we give a summarized overview of the work which has been done within this thesis. Furthermore, we share our experiences regarding the problems that can be expected when designing and implementing a large healthcare IT project. We conclude the thesis with the outlook into the future research directions.

11.1 Summary

In this thesis we have pointed out the potentially highly transformative role that the federated DWH technology could play within the healthcare sector. We argue that this technology can be deployed to enable efficient interoperability between heterogeneous medical information systems, which is the prerequisite for building sustainable healthcare decision-support systems. On the other hand, federated DWH represent the ideal basis for the integration of the EBM into clinical decision-making processes. Applying the DWH technology and data mining techniques can shorten the development time for creation of evidence-based rules and increase the certainty of the recommendations. Evidence-based rules become blueprints for the treatment team and road maps for the patients. The patients get access to a clearly written outline of the expected health progress over time, which contributes to a better understanding, acceptance and communication with the treatment team. In this thesis, we have identified the major payoffs of applying EBM in the decision-making process:

- Reduction of medical errors
- Greater operating efficiency
- Improvement of patient care
- Reduction of treatment costs

The connection of EBM to the clinical decision-support systems merges two powerful methods for enhancing the quality of care. To realise this potential, DWH technology is deployed to provide decision makers with the literature-based and practice-oriented evidence in both detailed and aggregated form. Decision-support tools help view this data in any requested way and data mining techniques allow the insight into unknown data patterns. The consequence of not applying EBM is the

time loss in the patient treatment procedure. However, even when EBM is applied, it needs to be presented to the decision makers in a proper and usable manner. Purely external, evidence-based knowledge is not sufficient for efficient treatment of the individual patient. This knowledge needs to always be adjusted to the patient's health condition and preferences. It takes a very powerful system to accomplish this task. We consider the federated DWH to be the suitable solution for integration of evidence-based rules into the clinical decision support.

Federated DWH facilitating EBM has an obligation to provide a mature security policy and to assure reliable security measures to guarantee the privacy of the highly sensitive healthcare data. Since the disclosure of patient identification is not required for EBM, pseudonymized data make the ideal foundation for pattern recognition and creation of evidence-based guidelines. The federated approach is a step towards decentralisation of security assurance; each of the underlying DWHs provides the federation with depersonalized and pseudonymized patient data.

In this thesis, we have shown that the federated DWH technology can meet two major requirements: the local need for flexibility and performance and the global need for consistency and control. When deployed for consolidation of heterogeneous medical information systems and for integration of evidence-based knowledge into clinical decision-support systems, it enables strategic decision making for both clinical business management and for the caregivers at the point of care, which results in a better service for the patient, the medical personnel and administrative staff.

11.2 Lessons Learned

In the course of the research work on this thesis, we have learned much about the obstacles which could face a large IT project focused at the cooperation among major participants of the healthcare system. In this subchapter, we summarize the major obstacles we have experienced.

- **The role of the investor:** If the project's goal is the modernisation of the medical information system, (for instance, in our case, building a sustainable healthcare DSS), it is funded by the government and the funding is tied to the accomplishment of specific implementation milestones, than it is easier to achieve the goals placed upon it [Infoway, 2007b]. In the case of a largely privatised healthcare system, where there are many private electronic medical records, any change in the medical information system, aimed at building a unique patient healthcare record, needs to be agreed upon by all care providers involved. This makes the changes in such a system a challenging and a long lasting task.
- **Attainment of the physicians' cooperation:** The success of the project is highly dependable on the physicians' willingness to cooperate. For this reason, it is important to make it clear that the purpose of the improved decision support system is not to boost the caregivers' efficiency but to enhance the quality of care.

- **Data ownership:** Data ownership is a delicate issue in the context of sharing sensitive patient data. The risk of weakening of the ownership position and co-determination rights in the healthcare arena is one of the major reasons for the healthcare providers to refuse the release of their own data for general use. They tend to retain the full control of their data assets even when the data is to be exchanged with their peers. Federated DWH solution, where data remains at its origin and is only retrieved if required, is a preferable answer to address such concerns.
- **Costs:** Building a sustainable healthcare network is a highly expensive task [NHS, 2005]. Successive approach, where such network is first built and proven to be efficient on a smaller scale (for example in one region only) is a good start for broader, nation-wide projects [Schanner, 2006].
- **Data quality:** The insufficient quality of data stored in databases and data marts within one organisation complicates the automated data processing. Patient and treatment data are often redundantly stored in different tables or the tables are poorly populated. While doing the research work on this thesis, we have experienced problems when trying to identify Diabetes Mellitus patients in one social insurance database. The diagnosis tables were insufficiently populated and we had to use indirect ways (by analysing prescribed drugs and therapies) to detect the Diabetes patients. In such cases, powerful analytical tools (like OLAP and data mining tools) and close cooperation with the clinician are critical for the project success.

11.3 Outlook

Several issues that were covered in this thesis shall be further developed in future work. The most relevant ones are:

- **Strengthening the interoperability at the international level:** Cooperation among international healthcare providers is needed in order to facilitate global healthcare for moving population. This concern encompasses development and adoption of international healthcare standards as well as creation of the unique patient identifier.
- **Full automation of EBM updates:** DWH and data mining technology needs to be further developed in order to be able to follow and integrate the changes coming out of the clinical research. It is an open research issue to create machine-interpretable sources of evidence and thus to automate the whole process of clinical evidence integration into a CDSS.
- **Improvement of security measures for guaranteeing patient data privacy:** In coming years, the security regulations and standards specifying the protection of patient data confidentiality will be continually extended. Depersonalization and pseudonymization techniques need to be further developed in order to meet the coming requirements for development of a mature security system. In our further work, we will handle the secure access models suitable for healthcare decision support systems based on a role-based approach in more detail.

Abbreviations

ACL	Access Control List
CDA	Clinical Document Architecture
CDSS	Clinical Decision Support System
CfH	Connecting for Health
CSN	Citizen Service Number
DAC	Discretionary Access Control
DB	Database
DBA	Database Administrator
DBMS	Database Management System
DPA	Data Protection Act
DSS	Decision Support System
DWH	Data Warehouse
EBG	Evidence-Based Guidelines
EBM	Evidence-Based Medicine
ECG	Electrocardiogram
ECTL	Extract, Correction, Transformation and Load
EEG	Electronencephalogram
EER	Enhanced Entity Relationship
EHR	Electronic Health Record
EPA-I	Electronic Patient Record Index
ER	Entity Relationship
HIPPA	Health Insurance Portability and Accountability Act
HL7	Health Level Seven
IHE	Integrating the Healthcare Enterprise
IT	Information Technology
ITI TF	Infrastructure Technical Framework
LDM	Logical Data Model
MAC	Mandatory Access Control
MAGDA-LENA	Medizinisch-Administrativer Gesundheitsdatenaustausch-. Logisches & Elektronisches Netzwerk
MI	Myocardial Infarction
MPI	Patient Master Index
NHS	National Health Service

OLAP	On-Line Analytical Processing
OLTP	On-Line Transactional Processing
PID	Patient Identifier
PIPEDA	Personal Information Protection and Electronic Documents Act
RBAC	Role Based Access Control
RB-RBAC	Rule-Based Role Based Access Control
RCT	Randomized Controlled Trial
RDF	Ressource Description Framework
RIM	Reference Information Model
SQL	Structured Query Language
SSN	Social Security Number
STRING	Standards und Richtlinien für den Informatikeinsatz im österreichischen Gesundheitswesen
SVNR	Sozialversicherungsnr. (Social Security Number)
TTP	Truster Third Party
UML	Unified Modeling Language
XDS	Cross-Enterprise Document Sharing
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

Bibliography

Abidi S.S.R., A. S. R. (2001). A Case of Supplementing Evidence Base Medicine with Inductive Clinical Knowledge: Towards Technology Enriched Integrated Clinical Evidence System. 14th IEEE Symposium on Computer-Based Medical Systems, (CBMS'2001), Bethesda (USA), 26-27 July 2001, IEEE.

AERZTE_ZEITUNG (2005). " Briten stellen auf elektronische Patientenakten um."

Agrawal, R., T. Grandison, C. Johnson and J. Kiernan (2007). "Enabling the 21st Century Health Care Information Technology Revolution." Communications of the ACM: 50(2):35-42.

Al-Kahtani, M. A. and R. Sandhu (2000). Access control meets public Key infrastructure, or: Assigning roles to strangers. 2000 Symposium on Security and Privacy.

Al-Kahtani, M. A. and R. Sandhu (2003). Induced Role Hierarchies with Attribute-Based RBAC. SACMAT'03, Como, Italy, ACM Press.

Audiimoolam, S., M. Nair, R. Gaikwad and C. Qing. (2005, February 7, 2005). "The Role of Clinical Pathways in Improving Patient Outcomes." Retrieved 18.7.2007.

Beyer, M., K. A. Kuhn, C. Meiler, S. Jablonski, et al. (2004). Towards a Flexible, ProcessOriented IT Architecture for an Integrated Healthcare Network. 2004 ACM Symposium on Applied Computing, Special Track on Computer Applications in Health Care (CAHC), Nicosia, Cyprus, ACM Press 264-271.

BIOPATTERN. (2007). "BIOPATTERN Network of Excellence." Retrieved 20.8.2007, from http://www.biopattern.org.

BMJ. (2007). "The British Medical Journal." Retrieved 19.9.2007, from http://www.bmj.com/.

Briggs, B. (2005). "Decision Support Software Matures." Health Data Management.

Britt, P. (2005, November 1, 2005). "A Warehouse of Opportunity." Finance Tech Retrieved 24.9.2007, from http://www.financetech.com/showArticle.jhtml?articleID=173402279.

Brobst, S. and J. Rarey (2001). "The five stages of an Active Data Warehouse evolution." Teradata Magazine.

Bundesgesetz (2004). "179. Bundesgesetz, Gesundheitsreformgesetz 2005, 7. Unterabschnitt, § 84a. (5)." Bundesgesetztblatt für die Republik Österreich, Teil I: 12.

Carner®. (2007). "An Introduction to the Evidence Based Decision Support Center (EBDSC)." Retrieved 26.9.2007, from www.cerner.com/public/filedownload.asp?LibraryID=12130.

CEBM. (2004). "Why the sudden interest in EBM?" Retrieved 18.7.2007, from http://www.cebm.utoronto.ca/intro/interest.htm.

Clifton, C., A. Doan, A. Elmagarmid, M. Kantarcioglu, et al. (2004). Privacy-Preserving Data Integration and Sharing. DMKD'04, Paris, France, June 13, 2004, ACM.

Cochrane_Collaboration. (2007). "Cochrane Centres." Retrieved 18.7.2007, from http://www.cochrane.org/contact/entities.htm.

Cochrane_Library. (2007). "Evidence for healthcare decision-making." Retrieved 18.7.2007, from www.theCochraneLibrary.com.

Codd, E. F. (1985). "How Relational is Your Database Management System?" Computerworld.

Craig, J. C., L. M. Irwig and M. R. Stockler (2001). "Evidence-based medicine: useful tools for decision making." The Medical Journal of Australia, MJA: 174: 248-253.

Critchlow, T., M. Ganesh and R. Musick (1998). Automatic Generation of Warehouse Mediators Using an Ontology Engine. 5th INternational Workshop on Knowledge Represenation Meets Databases (KRDB '98), Seattle, Washington, USA, May 31, 1998, CEUR-WS.org 10:8.1-8.8.

dbMotion (2005). "The dbMotion™ Solution A White Paper."

Docobo®. (2006). "doc@HOME." Retrieved 16.8.2007., from http://www.docobo.co.uk/products_docathome.htm.

Donovan, M. (2005). "Real-time connections." Retrieved 25.9.2007., from http://www.teradata.com/t/page/131974/index.html.

Dowsey, M. M., K. M.L. and S. N.M. (1999). "A clinical pathway reduced length of stay, time to ambulation, and complications after hip and knee arthroplasty." Med J Aust: 170:59-62.

Duftschmid, G., W. Dorda and W. Gall (2003). MAGDA-LENA versus HIPAA: Two national frameworks for healthcare data exchange. a-telemed, 25-33.

Dutch_DPA. (2006). "Handling of Your Medical Data." Retrieved 29.8.2007, from http://www.dutchdpa.nl/documenten/en_inf_subj_Handling_Medical_Data.shtml.

Eckhardt, A. (2000). "Data Abuse and What Can be Done to Prevent it, The Patient in the Data Network." TA 36A/2000.

EHCRSupA. (2007). "Introduction to ENV 13606 Version 1.0." Retrieved 27.9.2007, from http://www.chime.ucl.ac.uk/work-areas/ehrs/EHCR-SupA/13606v1_8/sld003.htm.

eHealth. (2006b). "Impact of Emerging ICT on Patient Safety." Retrieved 27.9.2007, from http://www.ehealth-for-safety.org/workshops/bio_medical%202006/bio_medical_index.html.

Eichelberg, M., T. Aden, J. Riesmeier, A. Dogac, et al. (2005). "A Survey and Analysis of Electronic Health Care Record Standards." ACM Computer Surveys: 37(4):277-315.

Elsevier. (2007). "The Lancet."19.9.2007, from http://www.thelancet.com/.

Elsmari, R. and S. Navathe (2004). Fundamentals of Database Systems, Addison Wesley.

EU (1995). "EU Data Protection directive." Official Journal of the European Communities: 31.

Ferraiolo, D. F., D. R. Kuhn and R. Chandramouli (2003). Access Control Policy, Models, and Mechanisms - Concepts and Examples. Role-Based Access Control. Norwood, MA, Artech House**:** 27-49.

Ferrell, K. (2007). "Continental Airlines - Landing the right data." Retrieved 25.9.2007, from http://www.teradata.com/t/page/133723/index.html.

Gorenberg, M., A. Marmor and H. Rotstein (2005). "Detection of chest pain of non-cardiac origin at the emergency room by a new non-invasive device avoiding unnecessary admission to hospital." Emerg Med J 22: 486-489.

Gregoratos, G., M. D. Cheitlin, A. Conill and e. al. (1998). "ACC/AHA Guidelines for Implantation of Cardiac Pacemakers and Antiarrhythmia Devices: Executive Summary." Circulation: 97:1325-1335.

GVG© (2004). "Pseudonymisierung/ Anonymisierung." Aktionsforum Telematik im Gesundheitswesen.

Hashmi, N., D. Myung, M. Gaynor and S. Moulton (2005). A Sensor-based,Web Service-enabled, Emergency Medical Response System. EESR '05: Workshop on End-to-End, Sense-and-Respond Systems, Applications and Services, Seattle, WA, ACM.

Haslhofer, B. (2006). A Service Oriented Approach for Integrating Metadata from Heterogeneous Digital Libraries. The 1st International Workshop "Semantic Information Integration on Knowledge Discovery" (SIIK 2006), Yogyakarta, Indonesia, December 4-6, 2006, Austrian Computer Society.

Hauptverband. (2007). "Österreichische Sozialversicherung." Retrieved 27.9.2007, from http://www.sozialversicherung.at.

Hein, A., O. Nee, D. Willemsen, T. Scheffold, et al. (2006). SAPHIRE - Intelligent Healthcare Monitoring based on Semantic Interoperability Platform - The Homecare Scenario. European Conference on eHealth 2006, Fribourg, Switzerland, October 12-13, 2006, GI 191-202.

Heitmann, K. U. (2006). "Kommunikation mit HL7 Version 3 – Aspekte der Interoperabilität im Gesundheitswesen." Datenbank Spektrum(17/2006): 12-16.

HIPPA. (1996). "HIPPA.ORG." Retrieved 27.9.2007, from http://www.hipaa.org/.

HIR-group. (2006). "Health Informatics Research." Retrieved 26.9.2007, from http://www.cit.uws.edu.au/hir/.

HL7. (2007). "Health Level Seven." Retrieved 26.9.2007, from http://www.hl7.org.

IHE. (2006, June 2006). "IHE Cardiology Technical Framework Year 2: 2005-2006 Volume I Integration Profiles." Retrieved 26.8.2007, from http://www.ihe.net/Technical_Framework.

IHE. (2007a). "IHE changing the way healhtcare connects." Retrieved 27.9.2007, from http://www.ihe.net/.

IHE. (2007b). "IHE Connectathons Present a Unique Testing Opportunity." Retrieved 26.9.2007, from http://www.ihe.net/Connectathon/.

IHE. (2007c). "IHE in Europe 2007." Retrieved 27.9.2007, from http://ihe.univ-rennes1.fr/europe2007/.

Infoway, C. H. (2007a). "Transforming Health Care." Retrieved 20.9.2007, from http://www.infoway-inforoute.ca/.

Infoway, C. H. (2007b). "Advancing Canada's next generation of healthcare." Retrieved 11.10.2007, from http://www.infoway-inforoute.ca/en/pdf/Vision_2015_Advancing_Canadas_next_generation_of_healthcare.pdf.

Inmon, W. H. (1992). Building the Data Warehouse, John Wiley.

InterfaceWare. (2007). "HL7 Standard." from http://www.interfaceware.com/manual/hl7.html.

JAMA. (2007). "The Journal of the American Medical Association." Retrieved 19.9.2007, from http://jama.ama-assn.org/.

Jindal, R. and A. Acharya. (2004). "Federated Data warehouse Architecture, Wipro Technologies white paper." Retrieved 19.9.2007, from http://hosteddocs.ittoolbox.com/Federated%20data%20Warehouse%20Architecture.pdf.

Kalra, D., P. Singleton, D. Ingram, J. Milan, et al. (2004). Security and Confidentiality Approach for the Clinical E-Science Framework (CLEF). HealthGrid 2004, 44 (2): 193-197.

KBV. (2006). "xDT - Synonym für elektronischen Datenaustausch in der Arztpraxis." Retrieved 27.9.2007, from http://www.kbv.de/ita/4274.html.

KFF. (2006). "Comparing Projected Growth in Health Care Expenditures and the Economy." Snapshots: Health Care Costs Retrieved 22.10.2007, from http://www.kff.org/insurance/snapshot/chcm050206oth2.cfm.

Kim, K. (2005). "Clinical Data Standards in Health Care: Five Case Studies." ihealh reports.

Kimball, R. (1996). The Data Warehouse Toolkit, John Wiley.

Kirkgoze, R., N. Katic, M. Stolba and A. M. Tjoa (1997). A Security Concept for OLAP. 8th International Workshop on Database and Expert System Application 1997, September 1-2, 1997, IEEE Computer Press 619-626.

Krenn, F. (2006). "IT-Platforms help to improve efficiency in Health Care." Retrieved 27.9.2007, from www.ehealth-bg.org/images/Siemens_Franz%20Krenn_1.ppt.

Lampson, B. W. (1974). "Protection." 8(1):18-24.

Lin, F., S. Chou, S. Pan and Y. Chen (2000). Mining Time Dependency Patterns in Clinical Pathways. 33rd Hawaii International Conference on System Sciences, January 4-7, 2000, IEEE CNF 1:8.

Loos, C. (2006). "E-Health with Mobile Grids: The Akogrimo Heart Monitoring and Emergency Scenario."

Lujan-Mora, S., J. Trujillo and I. Y. Song (2002). Multidimensional Modeling with UML Package Diagrams. 21st International Conference on Conceptual Modeling (ER 2002), Berlin-Heidelberg, Germany, Springer Verlag 2503:199-213.

Malan, D., T. Jones, M. Welsh and S. Moulton (2004). CodeBlue: An Ad Hoc Sensor Network Infrastructure for Emergency Medical Care. International Workshop on Wearable and Implantable Body Sensor Networks, BSN 04.

Moron, M. J., E. Casilari, R. Luque and J. A. Gazquez (2005). A wireless monitoring system for pulse-oximetry sensors. 2005 Systems Communications (ICW'05), August, 14-17 2005 79-84.

National_Research_Council. (1997). "For the Record: Protecting Electronic Health Information." Retrieved 20.7.2007.

Nguyen, T. M., J. Schiefer and A. M. Tjoa (2005). Sense & Response Service Architecture (SARESA): An Approach Towards a Real-time Business Intelligence Solution and Its Use for a Fraud Detection Application. 8th ACM international workshop on Data warehousing and OLAP (DOLAP '05), Bremen, Germany, ACM Press 77-86.

NHS. (2005, June 2005). "The Spine." Retrieved 15.9.2007, from http://www.connectingforhealth.nhs.uk/resources/comms_tkjune05/spine_factsheet.pdf.

NHS. (2007). "Connecting for Health." Retrieved 15.9.2007, from http://www.connectingforhealth.nhs.uk/.

Open_Clinical. (2006). "Knowledge Management for Medical Care." Retrieved 27.9.2007, from http://www.openclinical.org/clinicalpathways.html.

openEHR. (2007). "The openEHR Foundation." Retrieved 27.9.2007, from www.openehr.org.

Patient Care Technologies®, I. "well@home Device." Retrieved 16.8.2007., from http://www.ptct.com/wah_device.html.

Perreault, L. E. and J. B. Metzger (1999). "A Pragmatic Framework for Understanding Clinical Decision-support." Journal of Healthcare Information Management: 13(2):5-21.

PIPEDA. (2006). "PipedaInfo." Retrieved 10.2.2006, from http://www.pipedainfo.com/.

Pommerening, K. and M. Reng (2004). "Secondary use of the EHR via pseudonymisation." Studies in Health Technology and Infromatics: 103:441-446.

Regenstrief_Institute. (2007). "Logical Observation Identifiers Names and Codes (LOINC®)." Retrieved 20.9.2007, from http://www.regenstrief.org/medinformatics/loinc/.

RESolutionLtd. (2007). "resolution xCase™." Retrieved 25.9.2007, from http://www.xcase.com/.

Rindfleisch, T. (1997). "Privacy, Information Technology and Healthcare." 40:92-100.

Sackett, D. L., W. M. Rosenberg, J. A. Gray, R. B. Haynes, et al. (1996). "Evidence-Based Medicine: what it is and what it isn't. [editorial]." BMJ 196: 312 (7023):71-72.

Sartipi, K., M. H. Yarmand and D. G. Down (2007). Mined-Knowledge and Decision Support Services in Electronic Health. International Workshop on Systems Development in SOA Environments, Washington, DC, USA, IEEE Computer Society.

Schanner, A. (2006, September 2006). "NÖ am Weg zur elektronischen Gesundheitsakte." Retrieved 20.9.2007, from http://www.lknoe.at/.

Schuerenberg, B. K. (2003). "Clearing the Hurdles to Decision Support." Health Data Management.

SCIPHOX. (2006). "Dokumenten-Kommunikation im Gesundheitswesen." Retrieved 27.9.2007, from http://sciphox.hl7.de/ueber_uns/flyerallgemein.pdf.

Sheth, A. P. and J. A. Larson (1990). Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Computing Surveys. New York, USA, ACM Press. 22 (3)**:** 183-236.

SNOMED. (2007). "Systematized Nomenclature of Medicine (SNOMED)." Retrieved 20.9.2007, from http://www.snomed.org/.

Spronk, R. (2007, 19.3.2007). "The Spine, an English national programme." Retrieved 15.9.2007, from http://www.ringholm.de/docs/00970_en.htm.

Stolba, N., M. Banek and A. M. Tjoa (2006c). The Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine. First International Conference on Availability, Reliability and Security (ARES 2006), Vienna, Austria, April 20-22, 2006, IEEE Computer Society Press 329-339.

Stolba, N., T. M. Nguyen and A. M. Tjoa (2007a). Towards a Data Warehouse Based Approach to Support Healthcare Knowledge Development and Sharing. 2007 Information Resources Management Association (IRMA) International Conference, Vancouver, Canada, May 19-23, 2007, IGI Publishing.

Stolba, N., T. M. Nguyen and A. M. Tjoa (2007c). Towards sustainable decision-support system facilitating EBM. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2007), Lyon, France, August 23-26, 2007.

Stolba, N. and A. Schanner (2007). eHealth Integrator - Clinical Data Integration in Lower Austria. Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED 2007), Plymouth, England, June 25-27, 2007.

Stolba, N. and A. M. Tjoa (2006a). An Approach towards the Fulfilment of Security Requirements for Decision Support Systems in the Field of Evidence-Based Healthcare. Knowledge Rights - Legal, Societal and Related Technological Aspects (KnowRight2006), Vienna; Austria, February 16-17, 2006, Austrian Computer Society 51-59.

Stolba, N. and A. M. Tjoa (2006b). The Relevance of Data Warehousing and Data Mining in the Field of Evidence-Based Medicine to Support Healthcare Decision Making. International Conference on Computer Science (ICCS 2006), Prague, Czech Republic, February 2006, Enformatika.

Stolba, N., A. M. Tjoa, T. Mueck and M. Banek (2007b). Federated Data Warehouse Approach to Support the National and International Interoperability of Healthcare Information Systems. 15th European Conference on Information Systems (ECIS 2007), St. Gallen, Switzerland, June 7-9, 2007.

Strategis. (2005). "PIPEDA Overview." Retrieved 29.8.2007, from http://privacyforbusiness.ic.gc.ca/epic/site/pfb-cee.nsf/en/hc00005e.html.

STRING-Kommission (2000). "Rahmenbedingungen für ein logisches österreichisches Gesundheitsdatennetz ("MAGDA-LENA"), Teil 2: Hauptteil, Version 2.0."

Sun, L., P. Hu, C. Goh, B. Hamadicharef, et al. (2006). Bioprofiling over Grid for Early Detection of Dementia. 1st international conference on Scalable information systems (INFOSCALE '06), Hong Kong, ACM Press.

SYBASE. (2007a). "Elisa Mobiilsideteenused." Customer Success Story Retrieved 25.9.2007., from http://www.sybase.com/detail?id=1040227.

SYBASE. (2007b). "China Telecom." Customer Success Story Retrieved 25.9.2007., from http://www.sybase.com/detail?id=1036771.

Symmetry_Corporation. (2007). "ADAPT." Retrieved 19.9.2007, from http://www.symcorp.com/.

T-Systems. (2007). "Innovatives Design." Retrieved 20.9.2007, from http://www.t-systems.at/.

Taweel, A., A. Rector, D. Kalra, J. Rogers, et al. (2004). CLEF – joining up healthcare with clinical and post-genomic research. HEALTHCARE COMPUTING, 22-24 March 2004, British Computer Society 203-212.

Teradata. (2003, October 2003). "ONE GmbH." Industry Solutions > Communications Retrieved 25.9.2007, from http://www.teradata.com/t/pdf.aspx?a=83673&b=107494.

Teradata. (2004, July 2004). "Office Depot." Industry Solutions > Retail Retrieved 25.9.2007, from http://www.teradata.com/t/pdf.aspx?a=83673&b=166244.

Teradata. (2006, April 2006). "Teradata Solutions Methodology." Retrieved 24.9.2007, from http://www.teradata.com/t/pdf.aspx?a=83673&b=92291.

Tsui, F.-C., J. U. Espino, V. M. Dato, P. H. Gestelnad, et al. (2003). "Technical Description of RODS: A Real-time Public Health Surveillance System." The Journal of the American Medical Informatics Association (JAMIA): 10(5):399-408.

Uni_Münster. (2001). "Medizincontrolling/DRG Research Group." Retrieved 20.9.2007, from http://drg.uni-muenster.de/de/behandlungspfade/cpathways/clinicalpathways_reisebericht.php.

W3C. (2004a, Febrauary 10, 2004). "OWL Web Ontology Language Overview." Retrieved 18.7.2007, from http://www.w3.org/TR/owl-features/.

W3C. (2004b). "Resource Description Framework (RDF)." Retrieved 18.7.2007, from http://www.w3.org/RDF/.

W3C. (2007c, June 14, 2007). "SPARQL Query Language for RDF." Retrieved 18.7.2007, from http://www.w3.org/TR/rdf-sparql-query/.

Walters, T. (2003). "TELE.RING Dialing up customized services." Retrieved 25.9.2007., from http://www.teradata.com/t/page/114954/index.html.

White, B. (2004). "Making Evidence-Based Medicine Doable in Everyday Practice." American Academy of Family Physicians, Familiy Practiced Management.

Wikipedia. (2007a). "Randomized controlled trial." Retrieved 18.7.2007, from http://en.wikipedia.org/wiki/Randomized_controlled_trial.

Wikipedia. (2007b). "Data Protection Act." Retrieved 29.8.2007, from http://en.wikipedia.org/wiki/Data_Protection_Act.

Wikipedia. (2007c). "Personal Information Protection and Electronic Documents Act." Retrieved 29.8.2007, from http://en.wikipedia.org/wiki/Personal_Information_Protection_and_Electronic_Documents_Act.

Wikipedia. (2007d). "Directive 95/46/EC on the protection of personal data." Retrieved 29.8.2007, from http://en.wikipedia.org/wiki/Directive_95/46/EC_on_the_protection_of_personal_data#Scope.

Wikipedia. (2007e). "Pseudonymity." Retrieved 30.8.2007, from http://en.wikipedia.org/wiki/Pseudonymity.

Wikipedia. (2007f). "Dementia." Retrieved 20.9.2007, from http://en.wikipedia.org/wiki/Dementia.

Wyatt, J. C. and D. J. Spiegelhalter (1991). Field trials of medical decision-aids: potential problems and solutions. Annu Symp Comput Appl Med Care, 3-7.

Curriculum Vitae

Nevena Stolba

Born: 21.10.1971, Sarajevo, Bosnia-Herzegovina

Education

2003 – 2007	Ph.D. studies of Computer Science at Vienna University of Technology
09/2000	Teradata Certified Professional
01/1998	Master of Science (Computer Science) at Vienna University of Technology
1993 – 1998	Master Studies of Computer Science at Vienna University of Technology
1992 – 1993	Master Studies of Electrical Engineering at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
1990 – 1992	Master Studies of Electrical Engineering at the Faculty of Electrical Engineering, University of Sarajevo, Bosnia-Herzegovina
1986 – 1990	Mathematics Gymnasium, Sarajevo, Bosnia-Herzegovina Graduated with honours.

Work Experience

Since 2003	Research Assistant at Women's Postgraduate Collage for Internet Technologies, Institute of Software Technology and Interactive Systems, Vienna University of Technology
2001 – 2003	Business Consultant, CRM / Data Warehouse Solutions, Teradata division of NCR, Vienna, Austria
1998 – 2001	Data Warehouse Consultant, NCR Teradata Solution Group, Vienna, Austria

Peer-Reviewed Publications

Eleven scientific papers presented at international conferences.

Awards

05/2007	Best Paper Award for the paper "Towards a Data Warehouse Based Approach to Support Healthcare Knowledge Development and Sharing" at the 2007 Information Resources Management Association (IRMA) International Conference, Vancouver, Canada
04/2000	Consulting Excellence Award at the Global Recognition Event, Teradata Solution Group. Nassau, Bahamas.