The Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine

Nevena Stolba[1], Marko Banek[2], A Min Tjoa[3]

[1] *Women's Postgraduate College for Internet Technologies (WIT)*
Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria
stolba@wit.tuwien.ac.at
[2] *Department of Fundamentals of Electrical Engineering and Measurements*
Faculty of Electrical Engineering and Computing (FER), University of Zagreb, Croatia
marko.banek@fer.hr
[3] *Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria*
amin@ifs.tuwien.ac.at

Abstract

Healthcare organisations practicing evidence-based medicine strive to unite their data assets in order to achieve a wider knowledge base for more sophisticated research as well as to provide a matured decision support service for the care givers. The central point of such an integrated system is a data warehouse, to which all participants have access. Due to the high confidentiality of healthcare data, and the privacy policy of participating organisations, the proposed warehouse is not created physically but as a federated system. Its conceptual model is based on a widely accepted international standard to overwhelm the heterogeneity of the components.

Any disclosure of health data, especially when related to a particular person, could be irreparably harmful, and their protection is even legally prescribed. Depersonalisation and pseudonymisation are used to ensure that personal identities are made secret before sending data to the federation.

In this paper a case study of a federation of health insurance data warehouses (HEWAF) is described. The protection of data privacy and confidentiality in the underlying warehouse is guaranteed through reliable security measures in the federation.

1. Introduction

Evidence-based medicine (EBM) is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients [10]. It complements an existing clinical decision making process with the most accurate and most efficient research evidence. Application of evidence-based medicine concepts speeds up the transfer of clinical research findings into practice, leading to cost reduction and to the improvement of the healthcare process as whole.

A successful application of evidence-based medicine is strongly related to the data it relies on. The central part of an evidence-based medical information system is a large centralised data warehouse that unites all relevant internal healthcare data of an institution with the evidence-based guidelines coming mostly from outside sources.

In order to enhance the productivity of their administration, healthcare organisations practicing evidence-based medicine are striving to a better cooperation with other related organisations. The more data is joined together, the more knowledge can be gathered from it, to the benefit of all participants. However, since most of the organisations have been developing their local information systems independently (using different, mutually incompatible data formats) an integration of the heterogeneous warehouses is needed first.

The goal of this paper is to point out the advantages of consolidating data warehouses for the organisations practicing evidence-based medicine. We also examine the security risks that can arise in such an environment.

The contribution of our work is to present a federated data warehouse model for evidence-based medicine. We show the advantages of such an approach in the area of cost-effectiveness, security and usability. Furthermore, we propose security measures that need to be taken in order to assure patient privacy and to protect the highly sensitive medical data.

The paper is structured as follows. In Section 2 the need of integrating heterogeneous healthcare data is discussed and a federated data warehouse solution is proposed. Measures guaranteeing privacy in healthcare environment are outlined in Section 3. Section 4 introduces HEWAF, a use case of health insurance data warehouse federation and illustrates its conceptual model based on HL7 standards. Section 5 describes the use of depersonalisation and pseudonymisation in order to assure privacy and confidentiality in HEWAF. An outline of the related work is given in Section 6. Conclusions are drawn in Section 7.

2. Consolidation of data warehouses

Since healthcare decisions must be made over large, statistically significant data patterns, healthcare organisations practicing evidence-based medicine need to join their data into a single data warehouse, which becomes the foundation of the knowledge discovery system. Discovering rules in evidence-based medicine requires processing of the detailed data that correspond to the basic (i.e. most detailed) grain level in the component warehouses, so that all data from the components must be copied into the joint warehouse.

When several independent organisations share their data for mutual purposes, they may not allow any physical copy 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 any access to the data. Highly confidential healthcare records are a typical example of such data. In such situations a federated data warehouse is needed.

According to Sheth and Larson [11] a federated database is “a collection of cooperating database systems that are autonomous and possibly heterogeneous”. A federated data warehouse is a functional warehouse, a “big umbrella”. No central, large data warehouse that collects data from smaller component warehouses is created: heterogeneous data warehouses are functionally integrated into a single unit from the conceptual point of view using a unique common conceptual model. 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. Each component unit itself must work independently of the federation.

2.1. Requirements on a federation of health insurance databases

Health insurance data warehouses store information about patient encounters, treatments, therapies and drug prescriptions that are supported by insurance companies and institutions. These data have to be joined in a federated data warehouse. Although different means of protecting such confidential data exist (described in Section 3) insurance companies generally reject the idea of creating any physical copy of their internal data. Legislation aimed at protecting personal data, like HIPPA in the USA, PIPEDA in Canada or EU Data Protection directive [9], which took effect during the last decade, gives additional support to such a decision.

Since only a “single version of truth” and a unique interpretation of the joined data should exist, only a singular federation schema (the common conceptual model) is needed. The federation administrator has to take care of the federation’s privacy policy.

The collaboration project for evidence-based medicine described in this paper merges data warehouses of different health insurance organisations. Other organisations in the health insurance domain are expected to join the project. A universal, simple and flexible common conceptual model must enable potential future integrations to be done seamlessly and with a minimum effort.

2.2. Healthcare standards

The highest level of generalisation and portability for the conceptual model can be achieved if an international standard that covers all areas of healthcare is adopted. Current international standards used for healthcare information: HL7, ENV 13606 and openEHR have been developed in parallel since the early 1990s, some of them adopting some concepts of the others.

HL7 (Health Level Seven, [14]) is an ANSI-accredited standard developing organisation (SDO) for healthcare data. Version 3 of HL7 standards defines the object-oriented Reference Information Model (RIM), the starting point for all HL7 standards. RIM introduces six backbone *foundation classes*. HL7 Clinical Document Architecture (CDA) is an XML-based document markup standard that specifies the structure and semantics of clinical documents for the purpose of their exchange. A CDA document consists of a body (containing medical data) and a header (containing data about people and organisations connected to the data i.e. patient, clinician and hospital). Clinical data within the body of the document can be nested recursively. HL7 is used in many countries, primarily in hospitals.

openEHR and ENV 13606 also introduce object-oriented reference models and a modular structure of healthcare documents. The general information model of openEHR [17] 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 a substantial revision due to its unnecessary complexity, which even led to some ambiguity and non-interoperability [13]. xDT [15] is a de-facto standard in Germany, used by health insurance organisations, pharmacists and primary healthcare ordinations. Meanwhile, German hospitals have adopted HL7 standards. A comprehensive integration of xDT and HL7 standard has been performed by Sciphox [18]. 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 to implement the conceptual model of our federated data warehouse. 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. Meanwhile, it is general enough to allow a great degree of flexibility, expressed by using the semi-structured CDA constructions. 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).

3. Guaranteeing security in healthcare environment

Electronic processing of medical information is a major research issue, since this information is highly sensitive and private. Loading sensitive data into diverse unsecured data bases or putting it online increases the risks of data disclosure by unauthorised users. 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 [8].

The need to improve their business and the quality of healthcare has forced medical organisations to release their highly sensitive data for loading into decision-support systems such as data warehouses. The nature of data warehouse and ad-hoc OLAP analysis makes all available data as easily accessible as possible. The consequence of opening sensitive data for processing is that the security measures in a data warehouse need to be extremely reliable.

The main ethical concern of federated data warehouses for evidence-based purposes is to provide mechanisms and policies to preserve patient privacy while delivering a huge decision support system for research purposes and for supporting caregivers in their clinical practicing.

Rindfleisch [8] describes the well-known concept 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.

Evidence-based medicine research goals of recognizing the disease symptoms and treatment patterns from the clinical data can be reached by analyzing unidentifiable patient data. Depersonalisation and pseudonymisation procedures are used to protect patient privacy and confidentiality (in our work, we assume that the transportation security is guaranteed through Public Key Infrastructure, so that is not the subject of this paper).

3.1. Depersonalisation

The goal of depersonalisation 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 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 birth day, precise address, nick names, name of relatives etc.) that can be used for patient identification are removed.

3.2. Pseudonymisation

A patient is uniquely identified by her (his) social security number, which is made secret by the means of pseudonymisation Pseudonymity is 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) [20].

Pseudonymisation is used in the area of evidence-based medicine, because a consolidation of different patients' data needs to be carried out, whereas patients' identities must stay unknown.

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

1. one-way pseudonymisation
2. reversible pseudonymisation

As there is no need for re-identification of a person, the simplest kind of pseudonymisation - one-way pseudonymisation - is used. 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 [4]), more complex pseudonymisation forms need to be applied [22]. This procedure is called reversible pseudonymisation. Later re-identification can only be done by the data owner.

In the case of long-term observation, it must be assured that one patient always gets the same pseudonym, even if she (he) changes her (his) name, address or health insurance company.

4. HEWAF: a use case of health insurance data warehouse federation

Data warehouses that facilitate evidence-based medicine deal with huge amounts of delicate patient data. In the following sections we present HEWAF (Healthcare Warehouse Federation), a use case study of a federated data warehouse for the Austrian health insurance organisations. The purpose of the warehouse development project is to apply the principles of evidence-based medicine in the health insurance area, which finally leads to new, improved and more efficient algorithms in clinical practice as well as reducing costs. A basic overview of the common federated conceptual model of HEWAF is given in the rest of this section. Section 5 describes mechanisms for assuring privacy and confidentiality of healthcare data in the federation.

4.1. 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 as of June 30, 2005) is given in Table 1. The UML class diagram containing relationships between them is shown in Figure 1.

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.

Table 1. The six backbone classes of RIM

Class name	**Definition**	**Example**	**sub-classes**
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
Partici-pation	association between an Act and a Role, with Entity playing that Role	Dr. Smith prescribes a therapy for patient Doe	no
Act-Relati-onship	directed association between a source Act and a target Act.	a biopsy procedure as a result of an observation	no
Role-Link	connection between two roles expressing a dependency between those roles	clinician (employee)–hospital (employer)	no

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.

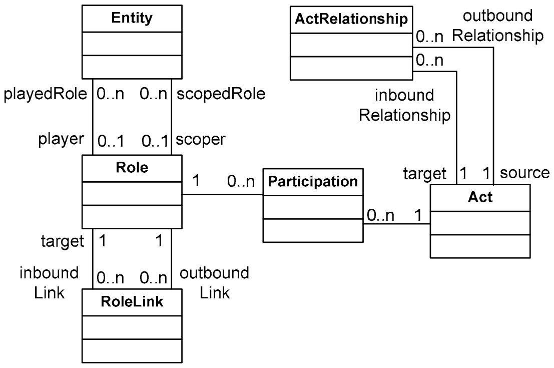


Figure 1. Class diagram showing the backbone classes of HL7 RIM

CDA focuses on Acts, Entities and Roles. Their interaction is defined in a more flexible manner than in RIM, in correspondence to the great structural freedom of XML. CDA data types describing RIM Acts, Entities and Roles conform to the RIM model with a still incomplete attribute matching. The body of the document contains CDA representation of RIM Acts, as shown in Figure 2 (example taken from [14]). It is a set of recursively organised *components* and *sections*. Each *section* consists of a pair of human readable *text* (that can be converted to HTML using XSLT stylesheets) and a number of *entry* elements wrapping descriptions of Acts (possibly containing a nested description of other Acts, to which they form an ActRelationship). The *entry* in Figure 2 describes the process of body temperature measurement.

```
<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>
```

Figure 2. Body of a CDA document

The topic of the Observation is uniquely stated by the *code* sub-element of *observation*. HL7 uses LOINC [16] and SNOMED CT [19] encoding.

4.2. Logical models of component warehouses

Multidimensional conceptual model is generally used for designing data warehouses as they are subject-oriented and give particular, user-perspective view of the information. They are most often implemented in relational database management systems (DBMSs), but, especially recently, other implementations have also been used (multi-dimensional DBMSs, databases for XML storage). Relational implementations do not necessarily use the star-schema solution. Figure 3 shows a non-star-schema logical warehouse model of a health insurance company participating in the federation project.

The owners of the warehouse pointed to therapy charging and drug prescriptions as the most interesting issues, which include crucial financial parameters. Therapy and prescriptions actually represent facts (with financial attributes as measures) while patient, clinician and clinic are typical dimensions for both facts (the prescriptions fact has an additional drug dimension). The necessary time dimension is also present. 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 a part of a patient encounter).

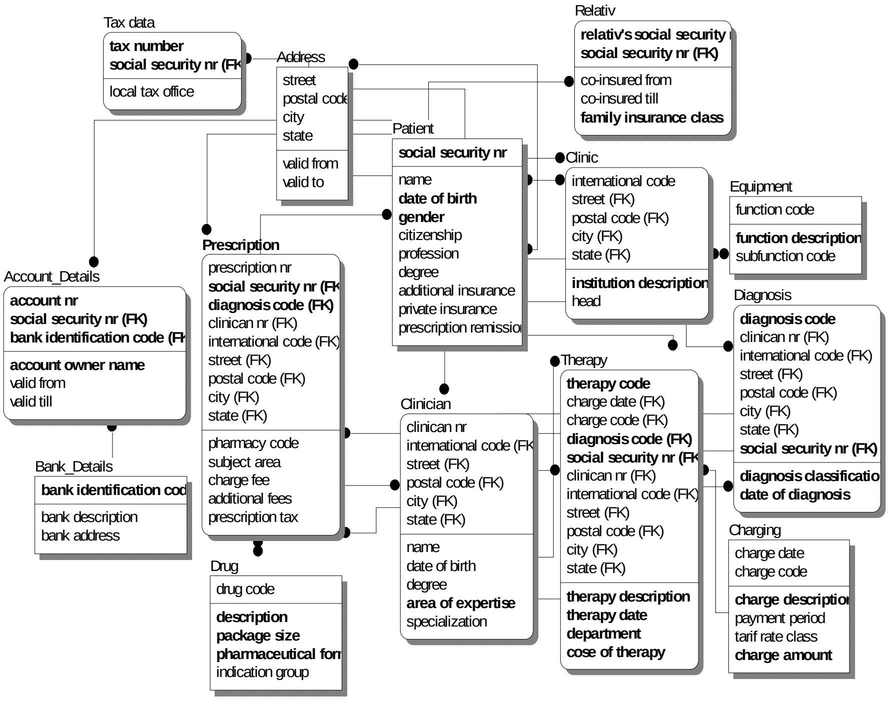


Figure 3. Logical data model for the data warehouse of a health insurance company

Such a 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 the DimDiagnosisDesc dimension, from those illustrating the encounter, which are presented as the FactDiagnosisEnc fact. DimDiagnosisDesc will also be a dimension for FactPrescription and FactTherapy facts. An illustration of the multidimensional model for the warehouse is given in the class diagram in Figure 4.

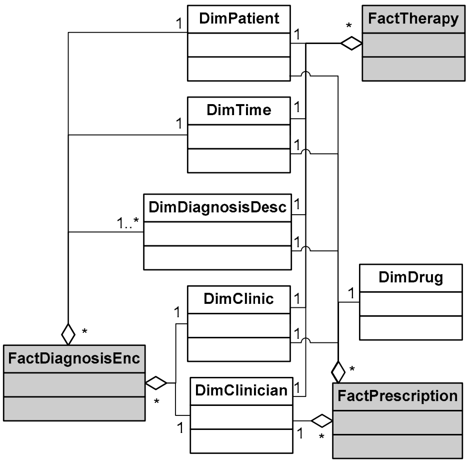


Figure 4. Multidimensional model of the sample health insurance data warehouse

4.3. Conceptual model of HEWAF

HEWAF introduces a common conceptual model for a federation of health insurance data warehouses that is based on HL7 RIM model and CDA format. The 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. Evidence-based medicine, the main purpose for creating the federation, adds diagnoses (disorders and diseases) as another crucial topic. Complete patients' medical history will be examined and compared (using various data mining techniques). Facts and dimensions in HEWAF are shown in Figure 4.

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).

Among many existing multidimensional models we chose the UML-based object-oriented conceptual model approach for data warehouses as presented by Trujillo and colleagues [5, 12] because of its strong formalism and the 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 4 to RIM classes with CDA notation, as shown in Table 2.

Table 2. Facts and dimensions in HEWAF conceptual model

fact/ dimen-sion	**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`
prescript-tion	Substance-Administration	Act	`substanceAdministration`
therapy	Procedure	Act	`procedure`
encounter	PatientEncounter	Act	`encounter`
drug	Material+Role (no sub-class)	Entity	`manufacturedProduct`

5. Security in HEWAF

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 insured patient's data and the adequate reliable security measures for data access.

Companies participating in this federation have the following business lines:

- Health insurance
- Retirement pension insurance
- Contribution calculation

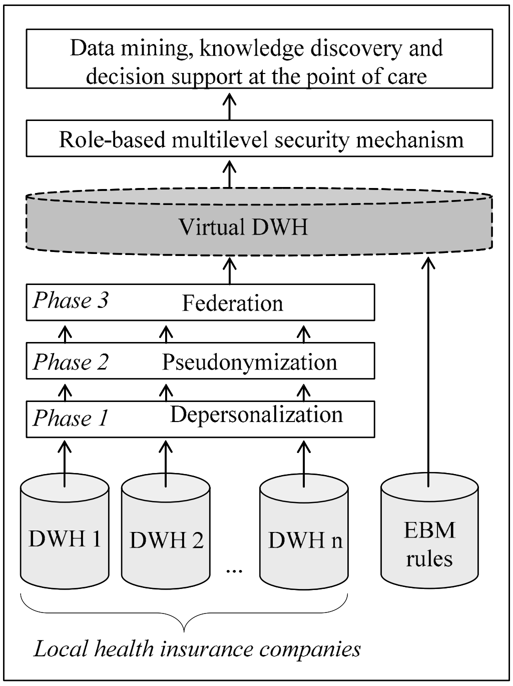
In our research we focus on the health insurance business line. It concerns all active insurants and all retirees. Depending on their income, they are divided into two groups: the ones who are allowed to visit the physician of their choice and those who are supposed to visit the contracted physician.

The purpose of the federation is to unite the data assets of local data warehouses in order to gain a broader base for knowledge discovery and data mining. Data mining heavily relies on evidence-based rules. 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 data warehouses, the social insurance companies apply a three-phase consolidation process, as presented in Figure 5:

1. Depersonalisation
2. Pseudonymisation
3. Federation

After the virtual federated data warehouse has been built, a *role-based multilevel security mechanism* 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 data warehouse facilitating evidence-based medicine is possible. An implementation of the security mechanism and an inquiry about the usage of the federated data warehouse is not the subject of this paper.



Data mining, knowledge discovery and decision support at the point of care
Role-based multilevel security mechanism
Virtual DWH
Phase 3
Federation
Phase 2
Pseudonymization
Phase 1
Depersonalization
DWH 1
DWH 2
...
DWH n
EBM rules
Local health insurance companies

Figure 5. Phases of data warehouse consolidation

5.1. Depersonalisation

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

While creating the conceptual data model of the federated data warehouse, business users (clinical management) specify sensitivity levels of data. Table 3 shows a part of such a sensitivity level list.

Table 3. Depersonalisation of sensitive data

Table	*Attrib.*	*Sensitivity level*	*Depersonalisation 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	Professi on	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

When building the conceptual model of the federated data warehouse, only data relevant for further analyses and reporting are considered. This data undergo a security check as well as a depersonalisation and pseudonymisation process. Irrelevant data (i.e. entities: Bank_Details, Account_Details) are not extracted from the source data warehouse.

The data modeller incorporates the specified privacy restrictions into the resulting logical data model. Figure 6 shows the logical data model of the depersonalised data. Key attributes, social security number and relative's social security number, will undergo a pseudonymisation process in the next step. They have no relevance for querying tasks in the federated data warehouse 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 data warehouse model. *Highly sensitive data* (i.e.: degree, local tax office) are hidden and can only be seen by authorised users. In Figure 6, these attributes are written inside curly braces.

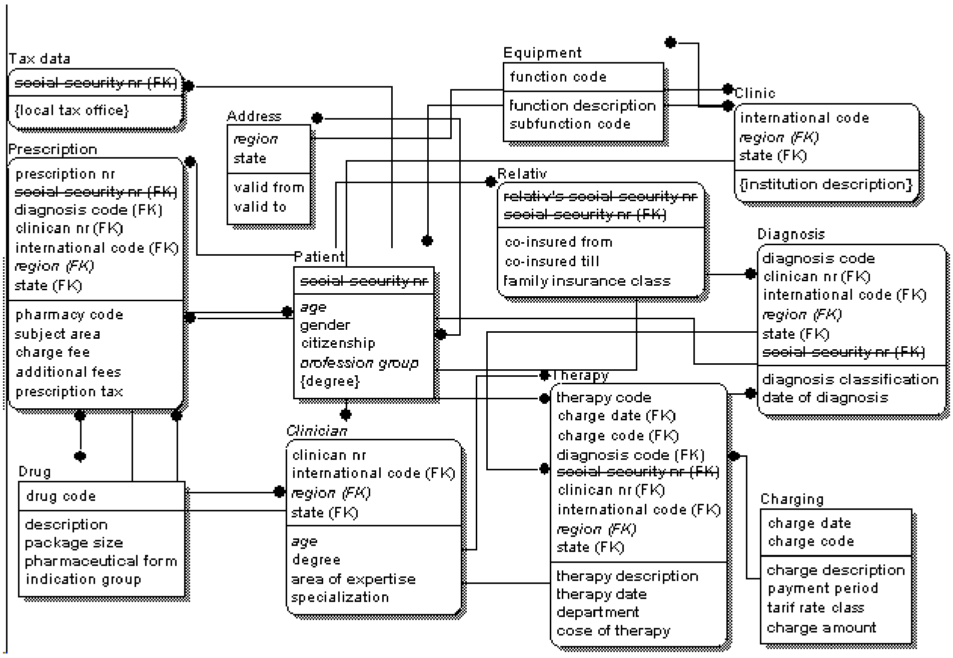


Figure 6. Logical data model of depersonalised data in a component warehouse

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.

5.2. Pseudonymisation

The pseudonymisation procedure can be performed after completing the depersonalisation process. In our case no re-identification of a patient is needed to reach the intended goals. Therefore, a one-way pseudonymisation procedure is applied. According to the Austrian law [21], pseudonymisation must be performed by a trusted third-party organisation.

Figure 7 represents the privacy preserving measures during query processing in the federated data warehouse. A user query is submitted to the federated data warehouse and then reassembled into particular sub-queries for each of the underlying data warehouses, according to its local logical data model. The query mapping is realised in the federated data warehouse. Each underlying local data warehouse receives and in the sequel processes its corresponding query. The resulting answer to a partial query consists of three main data parts:

- SSN – Social Security Number, which is used as unique patient identifier
- PD – sensitive Personal Data, to which the user access is restricted
- HCD – Health Care Data, which is non-sensitive medical data

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

The result of a depersonalised partial query is encrypted to enable a secured transportation to the pseudonymisation service.

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

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

6. Related work

Mapping of the access schemes between relational data models in OLAP systems is the subject of work of Priebe and Pernul [7]. They approach to this issue from the application side by introducing a methodology and a language for conceptual OLAP security design.

Rindfleisch [8] addresses the hard access to the clinical medical information. He points out that the move to the EPR needs to be done because of the many requirements and obligations (i.e. need to improve healthcare through decision support aids, meeting the needs of highly mobile patients, need for better clinical research and growing use of telemedicine and telecare).

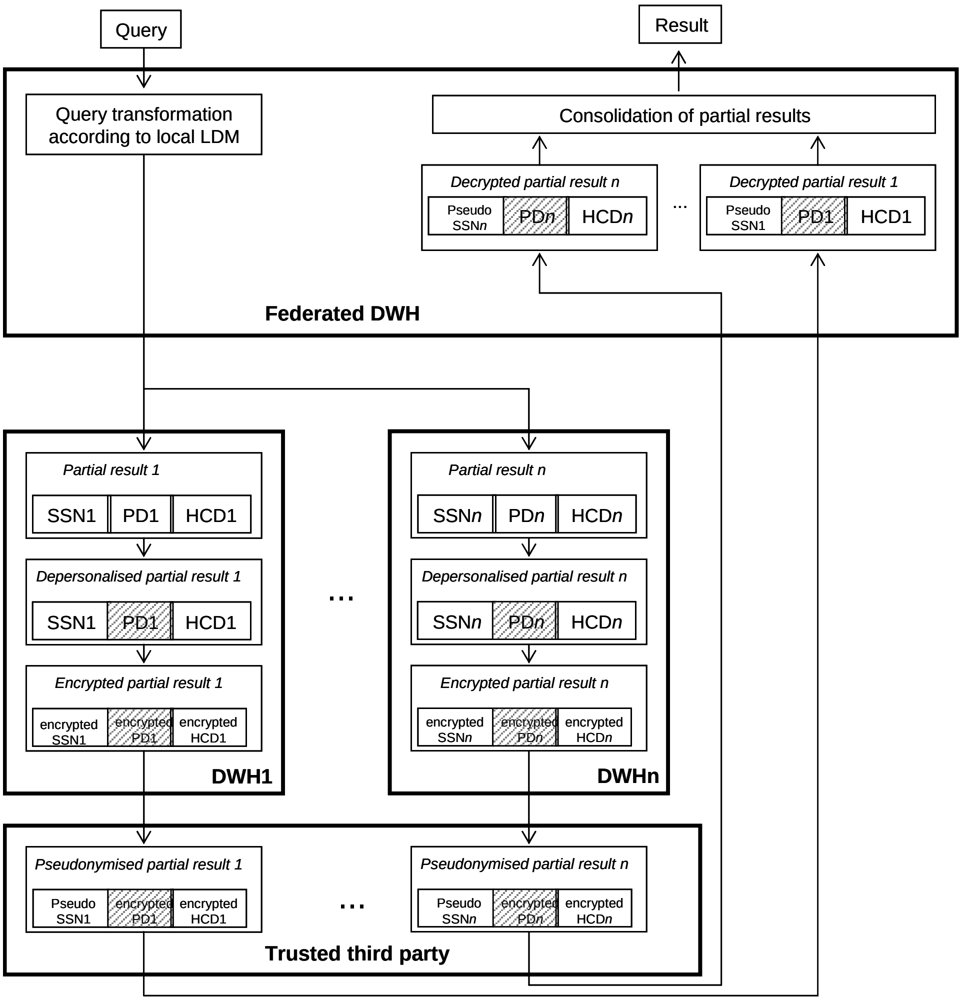


Figure 7. Privacy preserving for query processing in federated DWH

Further, Rindfleisch specifies the threats caused by the confidentiality of healthcare information and lists the general technological interventions to improve the system security.

Clifton and colleagues [2] propose a privacy framework for data integration and sharing for mining purposes, concerning healthcare in particular. The administrator of each component database defines which data are private by specifying a set of privacy views, and which users can be allowed to view private data by specifying privacy policies.

Pseudonymisation process has not been explored much so far. Some authors were engaged with that topic though: Pommerening and Reng [6] give an overview of pseudonymisation models. They distinguish between the straightforward pseudonymisation for one-time use of data and more sophisticated methods for long-term data accumulation. A description of pseudonymity and anonymity models for healthcare purposes is given by ATG [22]. They distinguish between centralised and decentralised data store and for each of them, different pseudonymity and anonymity models are proposed. The models presented by ATG are examined for their security, administrational and organisational effort. CLEF project described by Kalra et al. [4] aims to provide a Pseudonymisation repository of cancer patient history that can be accessed by researchers. They outline the necessity of robust mechanisms and policies to ensure that patient privacy and confidentiality are preserved while delivering a repository of medically rich information for the purposes of scientific research.

Basic principles for creating database federations are given by Sheth and Larson [11], who introduce a five-level architecture for federated databases (an extension of the well-known three-level ANSI/X3/SPARC approach for standard, non-federated database systems). A bottom-up integration approach is proposed for already existing databases, while a top-down concept of integration based on the common schema is recommended when all components are developed from scratch. An application of federation concepts to data warehouses is outlined in [3].

A grid-based system for federating heterogeneous health care information systems in Canada at the national level is described in [1]. Privacy and security services are provided using Public Key infrastructure (PKI). HL7 RIM and HL7 CDA are used as the common model of the federation. However, while this project is primarily aimed at transactional use of databases (very often access to small amounts of data, generally personal), our intention is to federated data warehouses for purpose of statistical analysis and discovering patterns.

7. Conclusion and future work

This paper presents a comprehensive and flexible conceptual model for health insurance organisations practicing evidence-based medicine, which federate their data warehouses in a single federation warehouse. It is based on the object-oriented RIM model and XML-based CDA messaging concept of the widely adopted international standard HL7.

Data warehouses facilitating evidence-based medicine are obliged to provide a matured security policy and to assure reliable security measures to guarantee the privacy of the highly sensitive health care data. The federated approach is a step towards decentralisation of security assurance; each of the underlying data warehouses provides the federation with depersonalised and pseudonymised patient data. Since for evidence-based medicine the disclosure of patient identification is not needed, pseudonymised data build the ideal foundation for pattern recognition and creation of evidence-based guidelines.

In our further work, we will detail the secure access model suitable for healthcare decision support systems based on a role-based approach.

8. References

[1] I. Bilykh, Y. Bychkov, D. Dahlem, J.H. Jahnke, G. McCallum, C. Obry, A. Anabajo, C. Kuziemsky: "Can GRID Services Provide Answers to the Challenges of National Health Information Sharing?", *Proceedings of the 2003 Conference of the Centre for advanced Studies on Collaborative Research*, IBM Press, 2003, pp. 39-53

[2] C. Clifton, A.H. Doan, A. Elmagarmid, M. Kantarcioğlu, G. Schadow, D. Suciu. J. Vaidya: "Privacy-Preserving Data Integration and Sharing", *Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2004)*, ACM Press, New York, USA, 2004, pp. 19-26

[3] R. Jindal, A. Acharya: "Federated Data warehouse Architecture", Wipro Technologies white paper, 2004 http://hosteddocs.ittoolbox.com/Federated%20data%20Warehouse%20Architecture.pdf

[4] D. Kalra, P. Singleton, D. Ingram, J. Milan, J. MacKay, D. Detmer, A. Rector, "Security and Confidentiality Approach for the Clinical E-Science Framework (CLEF)", HealthGrid 2004 Conference, *Methods of Information in Medicine*, Vol. 44 (2), 2005, pp. 193-197

[5] Lujan-Mora, S., Trujillo, J., Song, I.Y.: "Multidimensional Modeling with UML Package Diagrams", *Proceedings of the 21st International Conference on Conceptual Modelling (ER 2002)*, Lecture Notes in Computer Science, Vol. 2503, Springer Verlag, Berlin-Heidelberg, Germany, 2002, pp. 199-213

[6] K. Pommerening, M. Reng, "Secondary Use of the Electronic Health Record via Pseudonymisation", *Medical and Care Compunetics 1*, Studies in Health Technology and Informatics, Vol. 103, IOS Press, Amsterdam, The Netherlands, 2004; pp. 441 - 446.

[7] T. Priebe, G. Pernul, "A Pragmatic Approach to Conceptual Modeling of OLAP Security", *Proceedings of the 20th International Conference on Conceptual Modeling (ER 2001)*, Lecture Notes in Computer Science, Vol. 2224, Springer Verlag, Berlin-Heidelberg, Germany, pp. 311-324

[8] T. Rindfleich, "Privacy, Information Technology and Healthcare*", Communications of the ACM,* Vol 40 (8), ACM Press, New York, USA,1997, pp. 93-100

[9] V.W. Romney, G.W. Romney**: "**Neglect of Information Privacy Instruction-A Case of Educational Malpractise?**",** *Proceedings of the 5th Conference on Information Technology Education (SIGITE 2004)*, ACM Press, New York, USA, 2004, pp. 79-82

[10] D.L. Sackett, W.M.C. Rosenberg, J.A.M. Gray, R.B. Haynes, W.S. Richardson, "Evidence-Based Medicine: what it is and what it isn't". *British Medical Journal (BMJ)*, Vol. 312 (7032), 1996, pp. 71-72

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

[12] Trujillo, J., Palomar M., Gomez G., Song I.Y.: “Designing Data Warehouses with OO Conceptual Models”, *IEEE Computer, special issue on Data Warehouses*, Vol. 34 (12), 2001, pp. 66-75

[13] ENV 13606 -1 Reference, 2nd Working Draft, http://www.centc251.org/TCMeet/doclist/TCdoc04/N04-012prEN13606-1_2WD.pdf

[14] Health Level Seven (HL7), www.hl7.org

[15] KVB (Kassenärtzliche Bundesvereinigung Deutschland). xDT - Synonym für elektronischen Datenaustausch in der Arztpraxis. http://www.kbv.de/ita/4274.html

[16] Logical Observation Identifiers Names and Codes (LOINC), http://www.regenstrief.org/loinc/

[17] openEHR, www.openehr.org

[18] SCIPHOX, Arbeitsgemeischaft Sciphox GbR mbH, www.sciphox.org

[19] Systematized Nomenclature of Medicine (SNOMED), www.snomed.org

[20] Wikipedia, http://en.wikipedia.org/

[21] 179. Bundesgesetz, Gesundheitsreformgesetz 2005, 7. Unterabschnitt, § 84a. (5), Bundesgesetztblatt für die Republik Österreich, 30.12.2004, Teil I, Seite 12.

[22] GVG©, Gesellschaft für Versicherungswissenschaft und –gestaltung, Aktionsforum Telematik im Gesundheits-wesen, Management-Papier “Pseudonymisierung/ Anonymisierung“, Köln, 2004