Ruprecht-Karls-Universität Heidelberg / Hochschule Heilbronn

Diplomstudiengang Medizinische Informatik

September 12, 2011

DIPLOMA THESIS

Evaluation of the Secondary Use approach from Vanderbilt and its usability for Germany

Justin Doods

Supervisor:	Prof. Dr. Björn Bergh (University Hospital Heidelberg / ZIM)
Co-supervisor:	Assist. Prof. Joshua Denny M.D., M.S. (Vanderbilt University)
Advisor:	Dipl.-Inform. Med. Markus Birkle (University Hospital Heidelberg / ZIM)

Acknowledgment

First and foremost I would like to thank the CIO of the Department of Information Technology and Medical Engineering of the University Hospital Heidelberg Prof. Dr. Björn Bergh for taking over my supervisorship.

I would also like to thank Assist. Prof. Joshua Denny from the University of Vanderbilt for becoming my co-supervisor and for giving me the opportunity to come to Vanderbilt for my research.

My gratitude also goes to my advisor Dipl.-Inform. Med. Markus Birkle for his constructive advice and support for this thesis.

A special thanks goes to Haresh Bhatia who let me stay at his apartment during my stay in Nashville and without whom my visit might not have been possible.

I would also like to thank Johannes Porzelt and David Stein for proofreading my thesis.

Last but not least I would like to thank my parents, who supported me and made it possible for me to study.

Abstract

English This work addresses the analysis of the secondary use systems from Vanderbilt University, Nashville Tennessee (USA), and a followed up comparison to German concepts under the consideration of German legislation. In doing so important processes were modeled based on an on-site visit and later on compared with data privacy concepts by Technologie- und Methodenplattform für die vernetzte medizinische Forschung (TMF). It was also addressed, to what extend an adaption of processes and methods would be compatible with German data privacy.

The assessment of the results illustrates, that a better part of the underlying processes from Vanderbilt could be transferred to Germany, but that certain tasks would have to be implemented differently. It's also emphasized, that in spite of preventive measures and mechanisms risks for the privacy exist.

German Diese Arbeit befasst sich mit der Analyse der Secondary Use Systeme der Vanderbilt Universität, Nashville Tennessee (USA), und einem anknüpfenden Vergleich mit deutschen Konzepten und unter Berücksichtigung der deutschen Gesetzeslage. Dabei wurden, auf Basis einer vor Ort durchgeführten Analyse, wichtige Prozesse modelliert und im Anschluss mit den Datenschutzkonzepten der Technologie- und Methodenplattform für die vernetzte medizinische Forschung (TMF) verglichen. Weiterhin wurde darauf eingegangen, inwiefern eine Übertragung der Prozesse und Methoden mit dem deutschen Datenschutz vereinbar wäre.

Die Bewertung der Ergebnisse zeigt, dass ein Großteil der zugrundeliegenden Prozesse in Vanderbilt auf Deutschland übertragen werden können, jedoch bei gewissen Methoden andere Ansätze gewählt werden müssen. Es wird ebenfalls hervorgehoben, dass es trotz Schutzmaßnahmen und -mechanismen Risiken für die Privatsphäre gibt.

Contents

List of Abbreviations

BPMN Business Process Model and Notation

CPOE Care Provider Order Entry

CPT Current Procedural Terminology

DB Database

DOB Date Of Birth

DUA Data Use Agreement

EMR Electronic Medical Record

ETL Extract Transform Load

FDPA Federal Data Protection Act

GANI_MED Greifswald Approach to Individualized Medicine

GGDA German Genetic Diagnostics Act

GUI Graphical User Interface

HIPAA Health Insurance Portability and Accountability Act

HIS Hospital Information System

ICD International Statistical Classification of Diseases and Related Health Problems

IDAT Identity Data

IRB Institutional Review Board

LIS Laboratory Information System

MARS Medical Archival and Retrieval System

MDAT Medical Data

MIST MITRE Identification Scrubber Toolkit

MRN Medical Record Number

MMO Multimedia Objects

NCBI National Center for Biotechnology Information

NIH National Institute of Health

NLP Natural Language Processing

PID Patient Identifier

PHI Protected Health Information

PSN Pseudonym

PPV Positive Predictive Value

RIS Radiology Information System

SD Synthetic Derivative

SDPA State Data Protection Act

SHA State Hospital Act

TMF Technologie- und Methodenplattform für die vernetzte medizinische Forschung

TTP Trusted Third Party

UML Unified Modeling Language

UMLS Unified Medical Language System

RUI Research Unique Identifier

VUMC Vanderbilt University Medical Center

ZIM Department of Information Technology and Medical Engineering

List of Figures

List of Tables

1 Introduction

1.1 Aim and Motivation

Laws and insurance policy reasons obligate clinical staff to document every step of the patients treatment path and in this process sizable amounts of data accumulate. In the past the documentation was paper-based, but with compulsory periods of record-keeping reaching from 10 to 30 years in Germany, huge amounts of paper cumulated and recovery of documents was laborious. With technological advances in IT and storage media, Electronic Medical Record (EMR)'s were a logical step forward to make information retrievable and manageable again and nowadays every major hospital stores their documentation digital.

The "natural" accumulation of data from routine patient care and the tremendous information contained within this data is highly interesting for researchers. Different projects address the utilization of these latent information pools for medical research[1].

The expectation is, that systems which allow medical research by utilizing research databases[2] save a lot of time and money. Instead of going through patient files manually a research database can make all the patients and their documentation searchable and can present the information in a standardized and clearly arranged way. Finding sufficient patients for rare diseases or risk factors can be simplified and accelerated when searching through data of whole populations and consequently saving time that can be spend on actual research.

C. Safran states that "Secondary use of health data can enhance health care experience for individuals, expand knowledge about disease and appropriate treatments, strengthen

[1] GANI_MED: http://www.medizin.uni-greifswald.de/gani_med/; visited: 06.09.2011
EHR4CR: http://www.ehr4cr.eu/; visited: 06.09.2011
[2] See Fundamentals Chapter 2.2

understanding about effectiveness and efficiency of health care systems, support public health and security goals, and aid businesses in meeting customers' needs" [Safran et al., 2007, Page 1].

In contrast, patient data can't be shared and used for research purposes at will, because the use of this sensitive data is directly conflicting with a persons privacy rights. Certain laws and regulations in every country, like the *Bundesdatenschutzgesetz*[3], *Landesdatenschutz-*[4] and *Krankenhausgesetze*[5] in Germany and the *Health Insurance Portability and Accountability Act (HIPAA)* in the United States, constrain hospitals to ensure patient privacy when sharing information. These laws define what can be shared or used for research and in which form. At the same time, in the US, the National Institute of Health (NIH) has a Data Sharing Policy for projects that receive $500.000 or more in annual funding [National Institute of Health, 2003] that was "designed to increase access to data collected through, or studied with, federal funding" [Malin, 2010a, Page 3] .

Nevertheless, even when complying to national laws secondary use systems, with non-identifiable patient data, can be powerful tools for medical research.

At the beginning of this thesis, the secondary use concept propagated in the US, using the example of Vanderbilt's secondary use systems (*Synthetic Derivative (SD)* and BioVU) will be analyzed and the applicability of its processes and methods for Germany will be evaluated. This includes legal and technical standpoints, as well as ascertaining if Vanderbilt's processes and methods can be combined with secondary use approaches such as TMF's concepts.

1.2 Purpose of the Thesis

Secondary use systems are well known in the USA and used for a few years now. Today in Germany secondary use systems are implemented as well but there is no long-standing experience with such systems. To the knowledge of the author, no evaluation and comparison of legal bases and data sharing concepts between the US and Germany have been made so far. This thesis tries to answer the questions of compatibility and portability of the US and German concepts.

[3]Federal Data Protection Act
[4]State Data Protection Acts
[5]State Hospital Acts

1.3 Objectives

From the above stated purposes, the following objectives can be derived:

1. Analyzing Vanderbilt's secondary use processes including alpha numeric data and Multimedia Objects (MMO)'s.
    a) Analyzing Vanderbilt's secondary use of genomic data and the integration with clinical data.
    b) Advantages and disadvantages of the approach, process and implementation.
2. Identifying threats to patients privacy and evaluating protective methods.
3. Review if the processes from Vanderbilt would be compatible with German regulations and secondary use concepts.
    a) Analyzing German privacy laws and regulations.
    b) Compare analyzed processes from Vanderbilt with concepts in Germany.
4. Suggestions for processes/models suitable in Germany.

1.4 Issues to be addressed

From the in Chapter 1.3 mentioned objectives the following issues can be deduced:

1. Getting an overview of the laws and regulations in the US and Germany.
    a) Bundesdatenschutzgesetz, Landesdatenschutzgesetz / Krankenhausgesetz (Baden-Württemberg, Bavaria, Hesse, Mecklenburg-West Pomerania and Rhineland-Palatinate).
    b) Health Insurance Portability and Accountability Act (HIPAA)
2. Analyzing the systems and processes involved in and
3. Creating Use Cases of the secondary use process from Vanderbilt.
4. Comparison of the TMF data privacy protection concept and Vanderbilt's processes and methods.

5. Analyzing re-identification threads and evaluating protection methods.
6. Assess if Vanderbilt's process and methods are applicable with German law and regulations.
7. If possible, adapt processes and methods from Vanderbilt for Germany.

1.5 Layout of the Diploma Thesis

This thesis is organized in five main chapters.

The first chapter gives some insight about the aims and motivation.

Chapter 2 characterizes basic information about Vanderbilt University Medical Center (VUMC), its research and describes basic principles about secondary use and privacy described. This section is important to understand the motivation for secondary use of patient data and how VUMC operates in regards to research. It also outlines TMF's data privacy concept and gives a first insight into the statutory situation.

Chapter 3 describes the methods of how information about the miscellaneous topics was gathered, which search terms were used, how the laws were evaluated and how information was compiled. An overview of the tool used to create this work is presented as well.

Chapter 4 covers identified processes, use cases and the system architecture of SD and BioVU followed by a comparison of concepts. This chapter also describes the differences and similarities of data sharing/research from a legal point of view and outlines risks and risk prevention of Vanderbilt.

Chapter 5 summarizes the results, discusses the methods used, arrives at a conclusion of this thesis and gives some thoughts on future prospects.

2 Fundamentals

This chapter provides an overview of secondary use in clinical research and its risks for personal privacy in general. It also highlights the associated laws and regulations in the US and Germany, introduces data privacy concepts and re-identification risks will likewise be introduced.

2.1 Secondary Use & Research Databases

A topic that got more attention over the past few years is secondary use of patient data. *Secondary use* means, that information will be used for a purpose for which it was not primarily acquired. In this context it means that patient information that was collected through routine documentation in the hospital, is used for medical research. Applications of secondary use in the medical field can be "disease specific clinical or epidemiological research projects, health care research, assessment of treatment quality and health economy"[Pommerening and Reng, 2004, page 1]. Pommerening also states that typical aspects of secondary use are that "the data leave the context of the physician where they are protected by professional discretion and that the identity of the patient doesn't matter."[Pommerening and Reng, 2004, page 1]

Some positive aspects are that "Secondary use of health data can enhance health care experience for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about effectiveness and efficiency of health care systems, support public health and security goals, and aid businesses in meeting customers' needs" [Safran et al., 2007, Abstract].

The information in EMR's can't be used directly. Privacy policies forbid the use of identifying attributes called Protected Health Information (PHI)[1], which are mandatory

[1] More detailed information on PHI can be found in Chapter 2.3.

for patient care. One way to use that information nevertheless is the creation of databases for research purposes only. No official definition could be found so in this thesis those databases will be called ***Research Databases***. Research databases are data repositories that store information solely for the purpose of research, in other words they are the primary tools for information gathering and analysis. This work will focus on research databases where the patient information gets added in a automatic way, manually maintained databases will not be considered.

2.2 Vanderbilt Medical University Center

This section describes the research environment used in Vanderbilt University. It also describes the EMR of VUMC that is used for documenting the day to day patient care.

2.2.1 General Information

StarChart/StarPanel

In 2001 Vanderbilt completely replaced its first generation electronic patient record and started using its self-made EMR system *StarChart*. By 2009 StarChart contained data of over 1.8 million patients with more than 65 million documents[2], going back to the 70's and with comprehensive data especially for the past ten years [Presentation Masys, 2010, slide 11].

Its web-based component *StarPanel* integrates seamlessly and is generally used for documentation nowadays.

Advantages of StarPanel over StarChart are that several patient charts can be opened at the same time. A messaging system was implemented as well, so that clinical users, health care professionals and patients can interact with each other, without the need of a personal visit, for example when the patient just has a small question, which doesn't warrant a visit. Whole communications can arise this way, which can be saved and included in the patients record. The documentation that is not done electronically will be scanned, added to the record and the paper-based documents will be destroyed [Giuse, 2003, page 1].

[2]http://informatics.mc.vanderbilt.edu/archives/starchart 04/16/2009; visited 29.07.2011.

StarBRITE (Biomedical Research Integration, Translation and Education)

StarBRITE is Vanderbilt's web-based research portal, designed to "bind data, information, and knowledge to effective action in order to promote and speed the design and conduct of research" [Harris et al., 2011, page 2]. Additionally, StarBRITE allows to collect and analyze metrics routinely to analyze the effects of interventions designed to improve the research process [Harris et al., 2011, page 2].

The system focuses on the researchers needs and supplies educational resources, assists with research projects and their implementation, finding funding, data management and provides access to Vanderbilt's research databases.

As just mentioned, part of StarBRITE are the research databases, which are the *Synthetic Derivative,* a de-identified[3] copy of StarChart, and *BioVU*, Vanderbilt's biobank. Since the focus of this work lies on the secondary use systems both will be described in detail in the next sections.

For more detailed information on StarBRITE [Harris et al., 2011] can be used as a reference.

2.2.2 Research Databases

This subchapter gives an insight on Vanderbilt's Research Databases *Synthetic Derivative* and *BioVU*.

Synthetic Derivative

The *Synthetic Derivative (SD)* is a copy of StarChart, VUMC's EMR, stripped of patient identifying attributes called PHI.

The name Synthetic Derivative is composed of *Derivative* and *Synthetic.* "Derivative" stands for the information content that is derived from StarChart but reduced by removing patient identifiers. "Synthetic" stands for data that is systematically changed. Date attributes for example are modified, by changing their value to address privacy concerns while leaving the chronological context intact. [Presentation Masys, 2010, slide 2].

[3] Data stripped of identifying information; see Chapter 2.3.1.

By 2008 SD contained about 120GB of information (not including images) and more than 300 million observations for 1.4 million patients. By 2011 it included more than 1.95 million patients already [Roden, 2008, Harris et al., 2011].

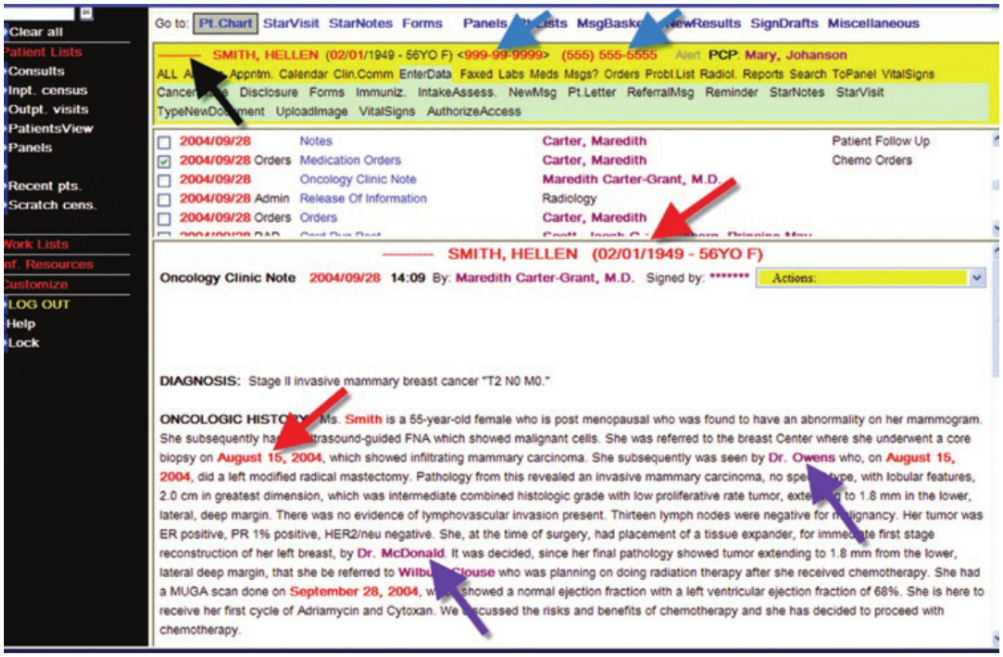


Figure 2.1: Synthetic Derivative [Roden, 2008]

De-Identification As mentioned, PHI get removed before the data enters SD. This is done by a method called *de-identification*, which is explained next.

De-identification for clinical data is done by *DE-ID,* a commercial tool developed by DE-ID Data Corp[4]. DE-ID is a Natural Language Processing (NLP) tool, that "uses a set of rules, pattern matching algorithms and dictionaries" [Meystre et al., 2010, page 9] to find and remove PHI. The software removes the identifiers by replacing them with tags that indicate the type of PHI that was removed[5]. This is done to keep readability of the text intact. Same identifiers, e.g. names, gets replaced with the same tags to keep the context. Vanderbilt additionally feeds "date shift values" into DE-ID so that dates don't get removed, but shifted back in time by 1-356 days.

[4]http://www.de-idata.com/; visited 29.07.2011.
[5]An example text can be found in the Appendix Figure 5.7.

The whole de-identification process gets explained in detail in Chapter 4.2.2.

General problems with NLP are *over- and undermarking,* where words are marked as PHI which are not and actual PHI get missed. Examples for over- and undermarking can be found in the Appendix[6].

Natural Language Processing An important task in the de-identification of patient documents in Vanderbilt is Natural Language Processing (NLP). It allows the scrubbing of documents that are written down as narrative text, so where the location of PHI is not known beforehand.

De-identification using NLP algorithms can be separated into manual and automated de-identification. Manual de-identification however is costly and the quality of the de-identification is very depending on a persons abilities [Presentation Wasserstrom, 2010, slide 8], therefor its not an option to use on a large scale. Automated de-identification on the other hand is less costly and can have a high performance [Presentation Wasserstrom, 2010, slide 9][Aberdeen et al., 2010][Roden, 2008, page 2], depending on the algorithm used. Automatic NLP is based on mostly two different groups of methodologies, *pattern matching* and *machine learning* or a combination of both[Aberdeen et al., 2010, page 3]. DE-ID uses the former, in form of lists and dictionaries.

A more detailed, in-depth comparison of different NLP algorithms can be found at [Meystre et al., 2010].

BioVU

Vanderbilt's second research database is *BioVU.* BioVU is a biobank containing genetic information extracted from leftover blood samples. The collecting commenced in 2007 and the database had about 91.000 samples at the end of 2010 [Harris et al., 2011, page 4], with every week between 500 and 900 new samples being added [Ritchie et al., 2010, pages 2][Roden, 2008, page 3].

The information in BioVU is de-identified as well to fulfill the HIPAA privacy requirements and is linked to data in SD (from the same patients), without any risk of discovering the identity. The exact mechanism of linking both systems is described in Chapter 4.2.6.

[6] Figures 5.8; 5.9.

Since genetic information is a delicate matter patients have the option to "opt out" of BioVU. On the last page of the consent to treat form patients can check, that they don't want their leftover blood to be used for the DNA database. This procedure will subsequently be termed **opt out model**.

Storage and usage processes of BioVU are explained in detail in Chapter 4.2.3.

2.3 Laws and Regulations

This chapter gives an overview about the judicial situation concerning privacy and data sharing. Here the laws and regulations of the US and Germany are summarized and briefly explained.

2.3.1 U.S.A.

In the United States a law ensures patient privacy while allowing data sharing called *Health Insurance Portability and Accountability Act (HIPAA)*. It allows Covered *Entities*[7] to share data in three ways, which are explained below.

Safe Harbor

Safe Harbor is the first method that allows data transfer. When Safe Harbor is applied to a data set, 18 fields with identifying attributes called Protected Health Information (PHI), like names, address, date of birth,... have to get scrubbed before the data may be shared. This is done to ensure that neither "the individual, the individual's relatives, employers, nor household members" NIH [2004, page 10] can be identified.

A list with all identifying attributes can be found in the Appendix (Figure 5.11)

[7]HIPAA defined Covered Entities as
a) a health care provider that conducts certain transactions in electronic form
b) a health care clearinghouse.
c) a health plan.
source: https://www.cms.gov/HIPAAGenInfo/06_AreYouaCoveredEntity.asp; visited 12.07.2011.

Limited Data Sets

Limited Data Sets are similar to Safe Harbor. The difference is, that not all the identifying attributes have to be de-identified. 16 identifying attributes have to be scrubbed, but fields like city, state, ZIP Code, elements of dates, and other numbers, characteristics, or codes not listed may be kept. These attributes are useful and sometimes needed for certain type of studies like epidemiology studies. Privacy concerns get satisfied by signing a binding *Data Use Agreement*. The researcher getting access to the Limited Data Set has to sign, that he will not try to re-identify patients among other things.

A list with all identifying attributes can be found in the Appendix (Figure 5.12)

Statistical Methods

Patient data can also be shared though other means than Safe Harbor or Limited Data Sets. The shared data can have any format as long as an expert attests, by using generally accepted statistical and scientific methods and principles, that the risk of re-identification is "very small". The statistician must document his methods and results and a covered entity must keep the documentation for at least 6 years [NIH, 2004, page 10].

2.3.2 Germany

In Germany the *Bundesdatenschutzgesetz* (eng. Federal Data Protection Act (FDPA)) together with *Landesdatenschutzgesetze* (eng. State Data Protection Act (SDPA)) and *Landeskrankenhausgesetze* (eng. State Hospital Act (SHA)) manage the protection of data privacy. In this thesis the State Data Protection Acts and State Hospital Acts of the states Baden-Württemberg, Bavaria, Hesse, Mecklenburg-West Pomerania and Rhineland-Palatinate were looked at.

[Schütze and Oemig, 2010] compiled a table that shows in which state which law regulates the privacy protection. An excerpt for the relevant states is shown in Table 2.1.

As can be seen in the table, a distinguishment has to be made between *internal* and *external usage*. In this context internal usage means that the data is directly used by the collector (usually the hospital or department itself) and external usage means that data gets transferred to a different site. External usage has to be divided into

	external usage		internal usage
	public sites	non-public sites	research
BaWü	SHA	FDPA	*
Bavaria	SDPA/SHA	FDPA	permitted by SHA
Hesse	SDPA/SHA	FDPA	*
MeckPom	SHA	SHA	*
RP	SHA	SHA	permitted by SHA

Table 2.1: Summarized table [Schütze and Oemig, 2010, Poster]

FDPA = Federal Data Protection Act
SDPA = State Data Protection Act
SHA = State Hospital Act
* subject to the possibility of authorization

öffentliche Stellen (eng. public bodies/sites), like hospitals financed by the state, and *nicht-öffentliche Stellen* (eng. non-public bodies/sites), like private hospitals.

For secondary use of patient data only external usage applies, since data leaves the oversight of the original source, as has been established in Chapter 2.1.

The following part of this section will highlight the articles that are of interest for external usage. A short summary will be given as well.

Federal Data Protection Act (FDPA):

The FDPA[8] manages the data privacy in Germany in regards to personal data. Its legality applies to privacy in IT as well as manual processing of personal data. To account for the differences and uniqueness of every state, the FDPA states that the statutes of the SDPA's overrule those of the FDPA[9]. If a state doesn't specify a statute, the FDPA's statutes automatically take effect. [BRD, 2009a, Sec.1 subsec.2]

- **Section15 subsection 1 paragraph 2** and **Section 16 subsection 1 paragraph 1** define when data may be transmitted to public and private sites. Data may be transmitted if usage as defined by **Section 14** is permitted.
- **Section 14** allows usage among others if consent is given.

[8] [BRD, 2009a]
[9] Section 12 subsection 2 FDPA.

Baden-Württemberg

In Baden-Württemberg non-public sites fall under the jurisdiction of the FDPA while for public sites Baden-Württemberg's SHA[10] applies.

State Hospital Act:

- **Section 46 subsection 1 paragraph 2a** says that patient data may be transmitted outside the hospital, if its necessary for medical research (by the hospital).
- **Section 50 subsection 1** states that patient consent has to be obtained in individual cases and that its not enough if the consent is obtained through general admission requirements.

A hospital/physician will only share its/his data if something can be gained, so the goal would always be research. This means that the requirements of the SHA for data transmission would automatically be fulfilled.

Bavaria:

In Bavaria the SDPA[11] as well as the SHA[12] apply to public sites. Non-public sites are covered by the FDPA.

SHA:

- **Section 27 subsection 5** states that patient information may be transmitted to a third party if the care context is given, permitted by statutory regulation or patient consents are given.

SDPA:

- **Section 15 subsection 1 paragraph 2** says that data processing (among others) is allowed if the patient gave consent.
- **Section 15 subsection 7** processing health related data (and others) is only allowed if the patient explicitly gave consent to the usage of that data.

[10] [Baden-Württemberg, 2011]
[11] [Bayern, 2009]
[12] [Bayern, 2008]

- **Section 18 subsection 1** and **Section 19 subsection 1** allow data transmission to public and non-public sites, if the usage is permitted by **Section 17 subsection 1 paragraph 2** and **subsections 2 - 4**, of which the following is of interest:
  - **Section 17 subsection 2 paragraph 2** permits the usage of data if consent was given.
- **Section 23 subsection 3** says that patient information has to be anonymized once the research purpose permits it. Until then identifying attributes have to be saved separately.

The bottom line of both laws is, that patient data may be transmitted if the patient gave consent. However the last section cited has a technical implication, since it requires a separation of data, so identifying attributes and medical data have to be separated in research databases as well.

Hesse:

In Hesse, similar to Bavaria the states SDPA[13] as well as the SHA[14] apply to public sites while the FDPA applies to non-public sites.

SHA:

- **Section 12 subsection 1** explicitly states, that for hospitals the SDPA applies.

SDPA:

- **Section 7 subsection 1 paragraph 3** says that data processing is only allowed if patient consent is given (among others).
- **Section 7 subsection 2** states that the consent has to directly relate to the data specified.
- **Section 7 subsection 4** processing health related data is only allowed as specified in **Section 33-35 + 39**, of which the following is of interest:
  - **Section 33** deals with usage for scientific research. Of importance is **subsection 2,** which states that identifying attributes have to be saved separately as soon as its possible and deleted when acceptable.

[13] [Hessen, 1999]
[14] [Hessen, 2011]

- **Section 13 subsection 2** allows the usage of data for secondary use for the same reasons that are stated in **Section 12 subsection 2+3**, where the following is of interest:
  - **Section 12 subsection 2 paragraph 1** allows data acquisition[15] if the person concerned gave consent (among other thing).

The legal situation in Hesse is similar to Bavaria. The most important thing is that consent has to be obtained and that medical and identifying attributes have to be separated as well.

Mecklenburg-West Pomerania:

In Mecklenburg-West Pomerania only the SHA[16] applies.

- **Section 17 subsection 1 paragraph 6** permits data transmission for research purposes through **Section 20**, which outlines the data usage for research.
  - **Section 20 subsection 1** allows patient data to be shared, if it got acquired for **Section 15 subsection 1**, which states:
    - **Section 15 subsection 1 paragraph 1:** patient data may only be acquired and saved, if its for the contract governing medical treatment, including the obligatory documentation requirements.
- **Section 20 subsection 4** says that identifying patient attributes have to be saved separately and that they should be deleted once the research purpose permits it.

The SHA allows data transmission for research purposes, if it got acquired for the patients treatment. A separation of identifying attributes and medical data applies in Mecklenburg-West Pomerania as well.

Rhineland-Palatinate:

In Rhineland-Palatinate similar to Mecklenburg-West Pomerania the states SHA[17] covers the patient privacy.

[15] In the case of Section 13 subsection 2: data transmission.
[16] [Mecklenburg-Pommern, 2008]
[17] [Rheinland Pfalz, 2011]

- **Section 37 subsection 3** permits transmitting to and usage of patient data by third parties for scientific research if the patient gave consent.
- **Section 37 subsection 4** says that patient information has to be anonymized once the research purpose permits it. Until then identifying attributes have to be saved separately.

Similar to the other states, consent has to be given before the data may be transmitted. A separation of data is noted as well.

German Genetic Diagnostics Act (GGDA)

The German Genetic Diagnostics Act (GGDA)[18] regulates the genetic diagnostics and the usage of genetic samples and data.

- **Section 13 subsection 2** says that genetic samples may be used for a different purpose than they were taken for, if permitted by law or if the patient gave consent in writing after a briefing.

As Section 12 subsec. 2 shows, genetic samples can also be used for research, if the patients gave consent.

2.4 TMF Data Privacy Concept

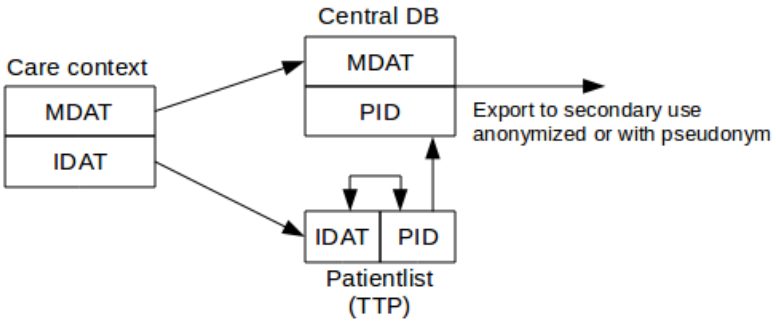
The *Technologie- und Methodenplattform für die vernetzte medizinische Forschung* (TMF) is a Non-Profit-Organisation in Germany with the goals to

- improve the conditions for medical research in regards to quality, organization and collaboration,
- answering of questions in networked medical research, legal & ethnic fundamentals and quality assurance & quality management[19].

The generic data privacy concepts A and B by the TMF will be described next as well as their concept for biobanks.

[18] [BRD, 2009b]
[19] http://www.tmf-ev.de/Ueber_uns/Ziele.aspx; vistited: 29.07.2011.

Model A



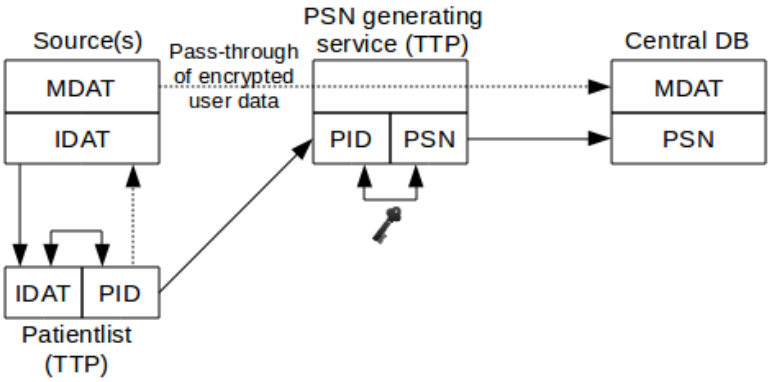
MDAT = Medical Data
IDAT = Identity Data
PID = Patient Identifier
TTP = Trusted Third Party

Figure 2.2: Model A [Pommerening et al., 2009, page 6]

A characteristic of Model A is that the research database is in close relation to patient care, which means that a physician has direct access to the database. Medical Data (MDAT) and Identity Data (IDAT) are only known to the physician at the same time though, in the system they are separated. The IDAT gets linked to a Patient Identifier (PID) and the PID in turn to the MDAT. When the physician wants to enter data the "patient reference" is dynamically generated via an identity management component called *Patientlist,* where the IDAT-PID relation is managed. The Patientlist itself is located at a *Trusted Third Party (TTP).* After positive authorization a temporary ticket will be handed out to the physician and the *Central Database (DB)* for the duration of the session. The Central DB can then be the starting point for secondary use. [Pommerening et al., 2009, Pommerening, 2009]

Model B

In comparison to Model A in TMF's Model B direct care context is not given. The IDAT and PID are managed in a Patientlist, at a TTP as well. Additionally to Model A the Patientlist is part of the quality management and assures that data from several sources get assigned correctly. Before the data can enter the Central DB a second security step has to be executed. At a second, different TTP a *Pseudonym (PSN)* is



MDAT = Medical Data
IDAT = Identity Data
PID = Patient Identifier
PSN = Pseudonym
TTP = Trusted Third Party

Figure 2.3: Model B [Pommerening et al., 2009, page 7]

generated out of the PID. Unlike the Patientlist, no link is stored and only the key for the generation is known. At the second TTP the MDAT can't be accessed, it is encrypted in a asynchronous way and can only be decrypted at the Central DB.

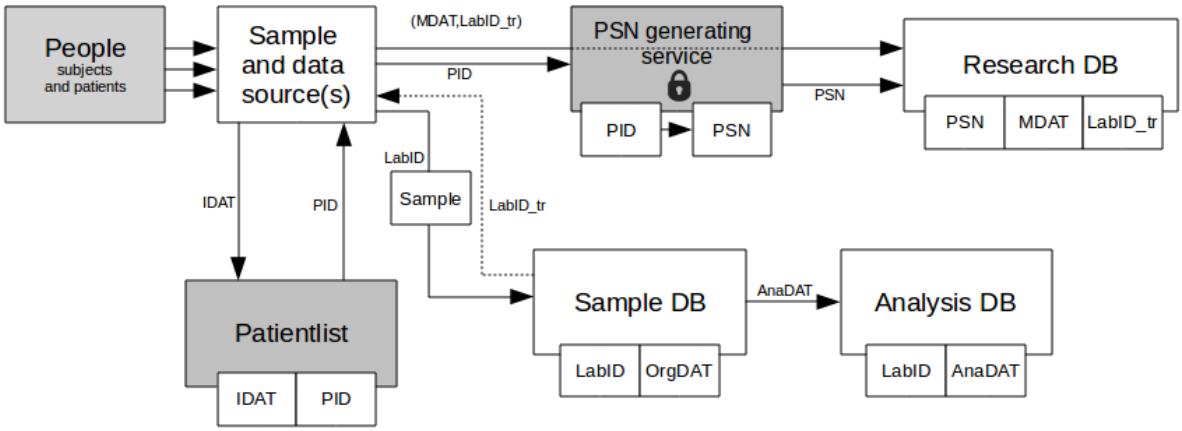
The generation of PSN's out of PID's makes the system error-tolerant, as a typing error of the patient name for example gets ignored. The information can still be linked together in the database because of the correct assigning of identity and PID.[Pommerening et al., 2009, Pommerening, 2009]

Model BMB

The goal of *Model BMB*[20] is to allow for the collection of samples in biobanks for research. In this respect the model complies with Model B which gets extended for genetical research and the storage of samples. The separation of IDAT and MDAT and the generation of pseudonyms out of PIDs are consequently the same as in Model B.

Samples get stored in the *Sample DB* and are referenced as *LabID* for unique identification. The ID gets created in a similar way to the PID, at the location where the sample is

[20]BMB = dt. Biomaterialbanken = eng. Biobanks.



MDAT = Medical Data
IDAT = Identity Data
PID = Patient Identifier
LabID = Laboratory ID
OrgDAT = Organizational Data
AnaDAT = Analysis Data

Figure 2.4: Model BMB [Pommerening, 2009, page 21]

taken. The MDAT can reference a transformed LabID_tr instead of the regular LabID, to add an extra level of security. LabIDs are also linked to Organizational Data (OrgDAT) for data management purposes and are needed for the identification of *Analysis Data (AnaDAT)* in the *Analysis DB*.

In conclusion, the *Research DB* links the PSN, MDAT and LabID/LabID_tr together. [Pommerening, 2009, 2007]

2.5 Re-Identification Threat/Prevention

Patient privacy is a big concern when sharing data, and even though data sets might seem un-identifiable, chances are that by combining data from several sources people's identities might be found nonetheless. In the mid 90's Sweeney was able to identify the medical record of the Governor of Massachusetts by combining publicly available discharge summaries and voter registration lists. The re-identification was achieved by

using the three non explicit[21] identifiers *ZIP code*, *Date Of Birth (DOB)* and *Gender [Sweeney, 2002, page 3].* As a result her work was cited in the original publication of HIPAA's Privacy Rule and influenced the Safe Harbor policy [Malin, 2010a, page 4]. Research has also be conducted that shows that re-identification risk and costs[22] vary for the different states in the US [Benitez and Malin, 2010]. This results from different states providing different information and not charging the same price for the voter lists.

El Emam and Dankar [El Emam and Dankar, 2008] describe two different scenarios of re-identification attacks, one, termed *prosecutor re-identification scenario,* where a "intruder" knows that a certain person is represented in the data set and wants to find those records. The second scenario is called *journalist re-identification scenario* and describes that a "intruder" doesn't care about a specific person, but just wants to re-identify someone. Sweeneys "attack" would fall into the latter domain.

People have their information in a multitude of places, for example when visiting healthcare providers and leave a trail of information behind in this way. Using these trails to re-identify individuals resulted in what's called *trail re-identification*, which will be described next after a quick excursion to *Risk mitigation* methods.

Risk Mitigation Information can be chanced for a variety of reasons, de-identification being just one of them. There are four general methods, how this can be done*,* suppression*,* generalization*,* randomization and synthesization. De-identification allows for all methods to be applied and Vanderbilt uses three out of the four, however not all at the same time and for the same purposes. The privacy model discussed later on in this chapter[23] also resorts to these methods.

Suppression is the simple removal of attributes that are too risky to be shared, the name of a person for example. *Generalization* is often done with ZIP codes, when not all the digits are displayed but only the first few which refer to a bigger region. With *randomization*, as the name applies, an attribute gets different random value. Vanderbilt is planning to randomize the names in SD for example to add an extra level of security. A different method of risk mitigation is *synthesization.* Here the data doesn't get shared,

[21] Explicit identifiers allow for direct identification. An example for explicit identifiers would be a set of data containing name and address [Sweeney, 2000, page 6].
[22] Voter registration lists may be publicly available, but have to be purchased.
[23] See k-Anonymity 2.5.

but compiled to aggregate statistics. From these "fake" statistics new, synthesized records get generated.[Malin et al., 2011, page 3]

Trail Re-identification

Sweeney showed that by using seemingly anonymous data from the *1990 U.S. Census summary* and *hospital discharge information,* 87% of the US population could be identified using only three attributes, 5-digit ZIP code, gender and date of birth[24] as can be seen in Figure 2.5. Even when using a two year age range instead of the date of birth, 0.01% of the population is still identifiable, which would be about 3.1 million people today[25].

County	18.1	0.04	0.00004	0.00000*
Place	58.4	3.6	0.04	0.01
ZIP	87.1	3.7	0.04	0.01
	DOB	**Mon/Year**	**BirthYear**	**2yr Age**

DOB = Date of birth

Figure 2.5: Percentage of re-identification in the US based on gender (not displayed), geography and age vary; [Sweeney, 2000, page 30]

There are different models which determine the trail re-identification risks by assessing how many records relate to unique individuals or how many people a record can be derived from [Malin et al., 2011, page 2]. Those models will not be highlighted in detail, rather the concept of trail re-identification will be explained.

The idea behind trail re-identification is described by means of a example taken from [Malin, 2010b] in Figure 2.6.

Imagine Ali, Bob, Charlie and Dan visiting three out four different hospitals in their region and leaving their genetic information there for various purposes. In this example it is assumed, that the hospitals disclose the identity data while the sensitive genomic data is not revealed.

First the names and DNA sequences will be separated into different tables *A (Identities)* and *B (DNA).* In the original table can be seen, that Bob and Dan did not leave their DNA at every hospital they visited. Next the tables A and B are converted into matrices,

[24] Values from the discharge summaries and voter registration lists: Figure *5.3.*
[25] Population of the US: 312,131,801
http://www.census.gov/population/www/popclockus.html (09/03/11 at 18:07 UTC).

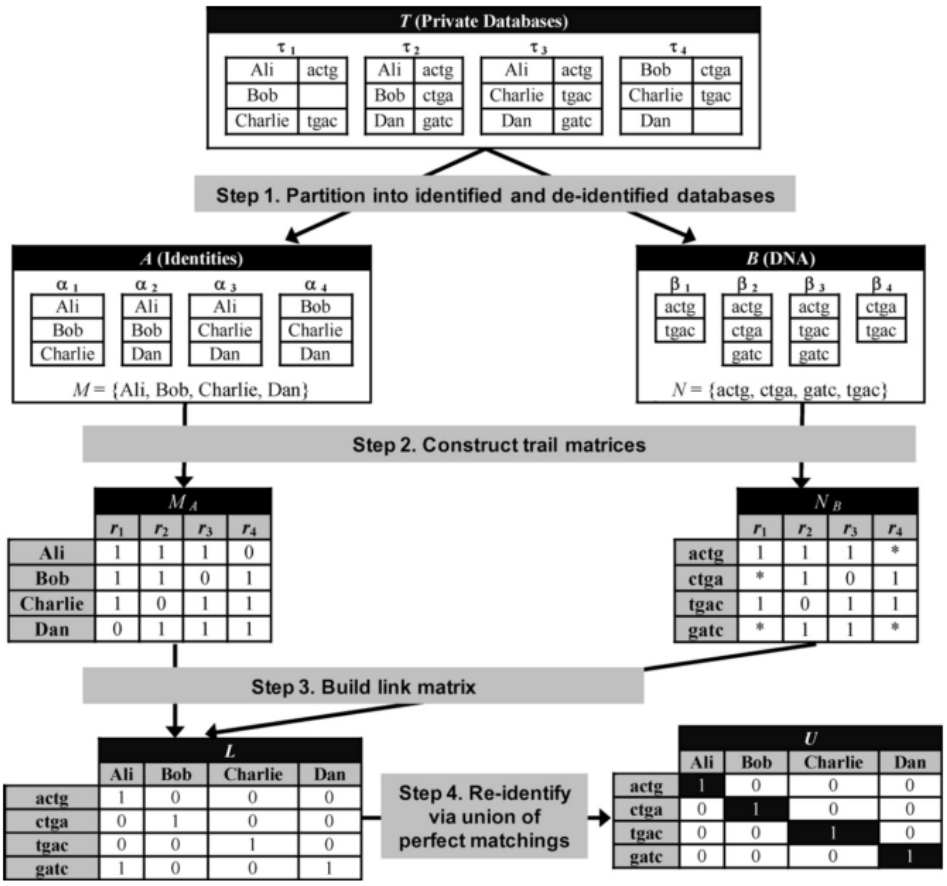


Figure 2.6: Trail re-identification example; [Malin, 2010b, page 3]

where the * acts as a wildcard. This means that it represents both possible values, true (1) and false (0) alike, since no DNA in table B doesn't automatically mean that the patient didn't visit the hospital (as can been seen in Bob's and Dan's case). Now both matrices are matched together and form the *link matrix L.* This is done by comparing each row from one table with each row from the other one. If the visits match with the DNA-patterns, the corresponding cell in matrix L gets a positive entry (1). As can be seen, L indicates that Ali might have the DNA sequence *actg* or *gatc*, since the * in the table could be a match as well. At the end the matrix will be reduced to *perfect matches*, which link the DNA sequences to the identities and results in the re-identification (black boxes). In Ali's case actg was linked to him, because the sequence and his visits matched perfectly while gatc had a wildcard at hospital 1.

One model trying to prevent re-identification is briefly described next. Such privacy models usually use suppression and generalization approaches described earlier to achieve their privacy goals. The concept and idea of the model will be described only, for more thorough information the stated sources should be visited.

k-Anonymity

Age	Gender	Zip
30	Male	15213
33	Male	15217
33	Female	15213
30	Male	15213

(a) Quasi-identifiers in a private table.

Age	Gender	Zip
30	Male	15213
33	*	1521*
33	*	1521*
30	Male	15213

(b) Quasi-identifiers in a 2-anonymous table.

Age	Gender	Zip
30	Male	15213
33	Male	1521*
33	*	15213
30	Male	15213

(c) Quasi-identifiers in a 2-ambiguous table.

Figure 2.7: k-Anonymity example; [Malin, 2008, page 5]

K-Anonymity is a privacy model, where data has to be changed in a way, that the combination of quasi-identifying attributes appears in k pieces of the data set. The model is based on the concept of *indistinguishability* in reference to the data set, as opposed to its original source [Malin, 2008, page 4]. A data set with a range of attributes is said to be k-anonymous, if for every row at least k-1 other rows exist that are undistinguisable based on quasi-identifiers [Ciriani et al., 2007, page 4].

In a prosecutor re-identification scenario the worst case scenario would be that the attacker matches his "target" against the smallest possible subset, which would be k for k-anonymous data. So the greater k, the safer the identities are, but at the same time the data set will be less differentiated.

Figure 2.7 shows an example of a 2-anonymous table, where attributes got removed by suppression. As can be seen, between rows 1 & 4 and 2 & 3 can not be distinguished so the table complies with 2-anonymity.

Other examples for privacy models are *k-Ambiguity* or *k-Unlinkabilty*. K-Ambiguity is similar to k-Anonymity, but it defines indistinguishability as *indiscernability* in respect to the original table, instead of equivalence [Malin, 2008, page 4]. A data set is considered

k-unlinkable on the other hand, if every piece of data can be related to at least k identities [Malin, 2006, page 107]. An example for a 2-ambiguous table can be found in Figure 2.7 and an example for for 2-Unlinkability can be found in the Appendix Figure 5.10.

3 Methods and Materials

This chapter describes the methods and materials used for this thesis.

3.1 Literature Research

To get an overview of the publicly available information a systematic literature research was conducted. To find relevant publications the search-engines PubMed[1] and Google Scholar[2] were used and the period of searching ranged from October 2010 until April 2011. The literature research was done in two major stages, one prior to an on-site visit at VUMC and one while at Vanderbilt.

The Following search terms were used prior to the visit:

database nih, database privacy rule, distributed databases, distributed databases privacy, research database nih, research database, research database usa, research database nashville, research repository, research repository hipaa, safe harbor, safe harbour, secondary use, secondary use privacy, secondary use safe harbor, secondary use hipaa, secondary use database, secondary use research database, hipaa, privacy rule, privacy rule hipaa, data sharing privacy.

Where the search term was to generic no abstracts were reviewed and it was tried to specify the term to a greater extent.

When starting with this thesis initial literature was given as well. This included information about Greifswald Approach to Individualized Medicine (GANI_MED) and the underlying TMF data privacy concept, as well as a list of Assist. Prof B.Malin's

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://scholar.google.com/

publications, Dept. of Biomedical Informatics Vanderbilt University, who has published about data privacy topics, re-identification methods and prevention[3].

The second research phase was not only about finding publications but also to gather information about more general topics. To find publications PubMed and Google Scholar were used and for general information Google's regular search engine. Search terms included: dbGaP, HIPAA, NLP, UMLS[4], ICD[5] 9, ICD 10, StarPanel, WizOrder/Horizon Expert Orders, Synthetic Derivative, BioVU, De-ID, MIST[6], HMS Scrubber.

Upon arrival in Vanderbilt a list with related publications and topics of interest was provided as well.

3.2 Legal Texts

Finding legal texts for the *Bundesdatenschutzgesetz*, the *Landesdatenschutz- & Landeskrankenhausgesetze* of Baden-Württemberg, Bavaria, Hesse, Mecklenburg-West Pomerania and Rhineland-Palatinate Google was used. Additional information about German privacy policies was provided by Department of Information Technology and Medical Engineering (ZIM) as well. Information about HIPAA was found via Google.

The articles in Chapter 2 are structured as follows and correspond to their German counterparts:

- (eng.) Section (Sec.) = (ger.) Paragraph
- (eng.) Sub-section (Subsec.) = (ger.) Absatz
- (eng.) Paragraph (Para.) = (ger.) Satz

3.3 Interviews

During the stay at Vanderbilt several interviews were held. Given the different expertise of the interview partners, no standardized form was prepared, but the questions were directed

[3] http://hiplab.mc.vanderbilt.edu/people/malin/CV.php#publications; visited: 09.03.2011.
[4] Unified Medical Language System (UMLS).
[5] International Statistical Classification of Diseases and Related Health Problems (ICD).
[6] MITRE Identification Scrubber Toolkit (MIST).

at the respective field of interest. Interview partners included various IT/Privacy-Experts and a Medical Doctor/Scientist.

3.4 Diagrams

The process model diagrams in Chapter 4 where created using the Business Process Model and Notation (BPMN) 2 standard.

The most important elements used in this thesis are briefly described.

- *Events* are depicted as circles and indicate that something happened, like the start or the end of a process.
- *Activities* are represented as rounded-corner rectangles and describe that something has to be done. The most used Activity type used in this work are *Tasks* which are single units of work.
- *Gateways* are represented as diamond shaped objects. At a Gateway, the process flow can split or merge again, depending on the conditions of the process modeled. At a Parallel Gateway two or more outgoing branches get executed simultaneously for example.
- There are several *Data* objects, which show what kind of data is used, how its used and where its saved. *Data Output* for example illustrates the data which is a result of an Activity or Task.

An overview of BPMN's core elements can be found in the Appendix[7] and a thorough description can be found in the standards specification[8].

The use case diagram was created by the specification of Unified Modeling Language (UML)[9]. The elements of use cases are:

- An *Use Case* symbolizes a certain functionality and sequence of actions. Use cases are depicted as ellipses.

[7] Core elements: Figure 5.1.
[8] http://www.omg.org/spec/BPMN/2.0/ visited 12.07.2011.
[9] http://www.omg.org/spec/UML/2.3/; visited 12.07.2011.

- An *Actor* represents a "user", which can be a Person, Organization or external system which interacts with the system being modelled. A stickman is the usual graphical representation.
- A *System Boundary Box* is a rectangle drawn around use cases to indicate the boundary of a system. Anything inside the rectangle is part of the system, everything outside is not.
- There are four use case relationships*: Include, Extend, Generalization* and *Association.* In this thesis only associations and includes were used, which are symbolized by dotted arrows and simple lines.

The system architecture diagram was done in a non standardized but simple and self-explanatory way.
For databases a cylinder form was used and the different sub-systems are symbolized as rectangles. Arrows show the relation between components and the standard or format that is used for communication. Their color indicates the context of the data (identifiable or de-identified data).

3.5 Tools used

For modeling the use case diagram Eclipse Modeling Tools[10] in combination with the Papyrus plugin[11] were used. The open source tool OpenOffice[12] Draw was used for the system architecture. BPMN diagrams were created using the open source tool Yaoqiang BPMN[13]. For literature research the reference manager Mendeley[14] was used and typesetting was done with LyX 1.6.7[15]. A literature database, using BibTex[16], was set up as well which was kept synchronized with Mendeley.

[10] Version: Helios Service Release 2.
[11] The plugin enables Eclipse to model UML 2.2 diagrams including Activity, Class, Communication, Composite Structure, Package, StateMachine, Sequence and aforementioned Use Case diagrams. Version 0.7.2.x; 2011/02/02.
[12] http://www.openoffice.org/; Version 3.2.
[13] http://bpmn.sf.net; Version 2.0.69.
[14] http://www.mendeley.com/
[15] http://www.lyx.org/
[16] http://www.bibtex.org/

4 Results

This chapter contains all the information that was collected for this thesis. It starts with Vanderbilt's system architecture, use cases and processes followed by a comparison of privacy laws of Germany and USA and ends with an analysis of the applicability of Vanderbilt's research databases using German laws and TMF's data privacy concept.

4.1 Vanderbilt

4.1.1 System architecture

This section describes the system architecture of Vanderbilt relating to its secondary use system and research databases.

The system can be separated in a hospital-internal and a hospital-external part with the data superiority and the responsibility to ensure patient privacy being on the internal side. As such, VUMC can't transmit data that leaves its oversight, unless it's already de-identified. This is why the de-identification process[1] has to take place under the supervision of Vanderbilt. DNA is sensitive material in itself and will not leave the oversight of the hospital. In this work the sphere of influence of VUMC will include the *Hospital Information System (HIS)*, *DNA storage* and an Extract Transform Load (ETL) Layer.

The hospital-external part of the system is named *Research resource*[2] in this thesis. It includes databases, procedures and processes that enable the researcher to use the system for medical research purposes.

[1] Described in Sections 4.2.2 and 4.2.3.

[2] Def: "The funding, research facilities, and materials available for research" http://de.dict.md/definition/Resource; visited: 13.07.2011.

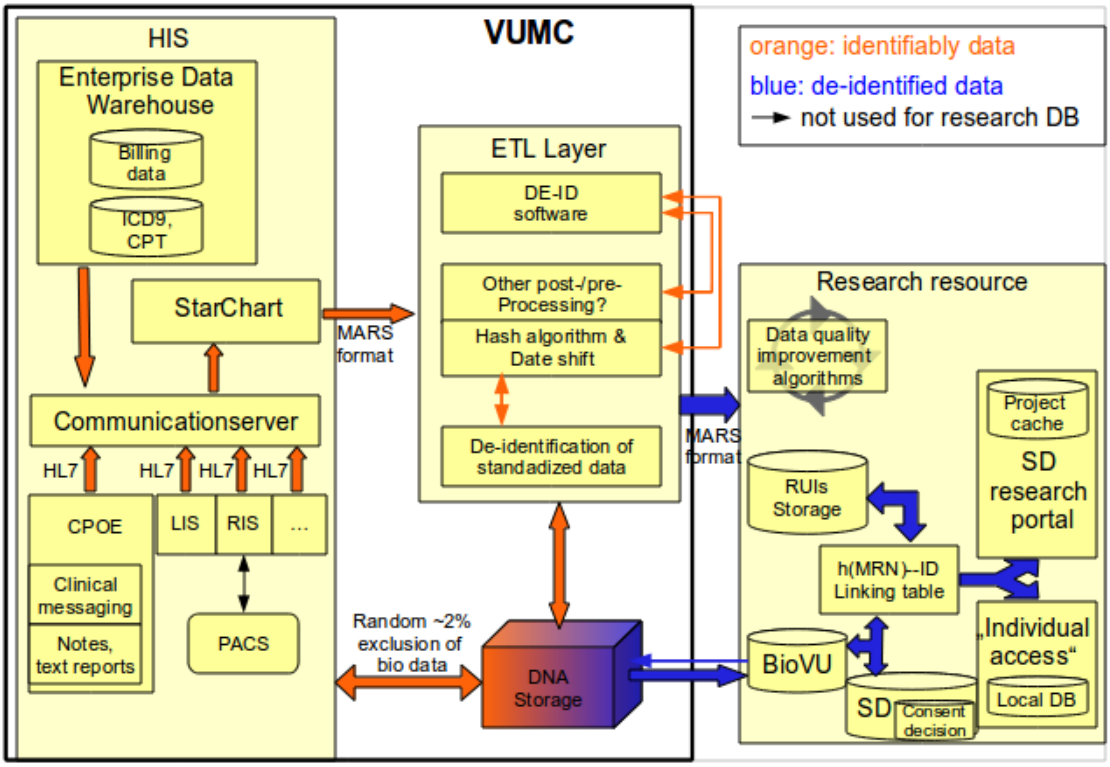


Figure 4.1: System architecture

All the source information is hosted in the *HIS*. It is located in one central system called *StarChart*[3]. StarChart is the central storage and all the other sub-systems like the *Enterprise Data Warehouse, Laboratory Information System (LIS), Radiology Information System (RIS),...* communicate with it through a *Generic interface engine,* which acts as a Communication server. This star-like topology reduces the amount of interfaces to n from n(n-1) for point-to-point models for example.

The Enterprise Data Warehouse has two databases with relevant information, one with ICD 9 and Current Procedural Terminology (CPT)[4] codes and a second one, which contains billing information of every patient. More comprehensive information can be found in the clinical systems. They include databases that contain notes and text reports,

[3]See Section 2.2.1.

[4]Used to code procedures (tasks, services,..) delivered to a patient; http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page; visited: 13.07.2011.

clinical messages, structured data like laboratory information from the LIS, medication and Care Provider Order Entry (CPOE) data.

StarChart is connected to the ETL Layer and sends documents in the Medical Archival and Retrieval System (MARS) format[5], where they get processed before they enter the Research Resource.

The ETL Layer is responsible for the de-identification of every document and several processes de-identify the different documents types (narrative text documents, semi-structured and structured data). Those processes are explained in detail in the subsequent chapters. Additionally the ETL Layer takes care of the de-identification of the blood samples as well. The samples themselves are stored in the *DNA storage* and are only taken out when needed for sequencing. A random exclusion of about 2% of the blood samples is done before the DNA gets extracted and stored. This is done to mask the patient group with genetic information in BioVU and in this way strengthen the patients privacy.

The de-identified data then get stored in the *Research resource*, more precisely EMR documents in SD's database and de-identified DNA samples get stored in the hospital-internal DNA storage. When a research project needs genetic information the DNA gets sequenced and the data gets stored in BioVB's database, which is located in the Research resource. All the documents in the SD and the information in BioVU get linked via hashed Medical Record Number (MRN) codes that are in turn connected to an identifier (ID). This ID is later linked to one or more Research Unique Identifier (RUI)'s. Every research project and every researcher has its individual and unique RUI. This is done so that the same patient will appear with a different number for different researchers. However, if one researcher is part of two or more research projects, the same patient gets the same number for all the projects the researcher partakes in.

Depending on the "project type" (standard SD GUI or "individual access") research project information gets stored in a *Project cache* or it gets extracted out of the regular SD DB into a *Local DB*.

A separate part of the Research resource are *Data quality improvement algorithms* which are executed on an irregular basis. Their goal is to improve the data quality by finding

[5]The MARS format was developed at the University of Pittsburgh to make its university hospitals EMR more accessible. An example of a MARS message can be found in the Appendix Figure 5.13.

missed PHI and removing or scrubbing the found information.

Vanderbilt does not include MMO's like X-ray and ultrasound images or sound files in its research databases, only Electrocardiogram (ECG) images are available. MMO's are not included because of their generally large sizes, which would result in slow response times.

Consent storage

The patients consents get captured either with *consent to treat form* documents, which get scanned and saved into the EMR or electronically via "check in kiosks", depending on the clinic the patients check into.

The patients decision regarding the consent of using leftover blood samples for BioVU is stored as a *bit* in SD's database.

The scanned documents themselves get not stored in SD, since their de-identification can't be guaranteed yet.

4.1.2 Use cases

Figure 4.2 shows the diagram of identified use cases and the involved actors.

System: VUMC

Use Case Name: Opt out

Description: This use case allows the *Patient* to opt out of the usage of his leftover blood for research purposes.

System: VUMC

Use Case Name: Discard blood sample

Description: In Vanderbilt leftover blood usually gets discarded every three days. However the leftover blood can be used for genomic research if the patients didn't opt out. This use case includes the use case *De-identify blood sample* and is extended by the use case *Extract DNA*.

System: VUMC

Use Case Name: De-identify blood sample

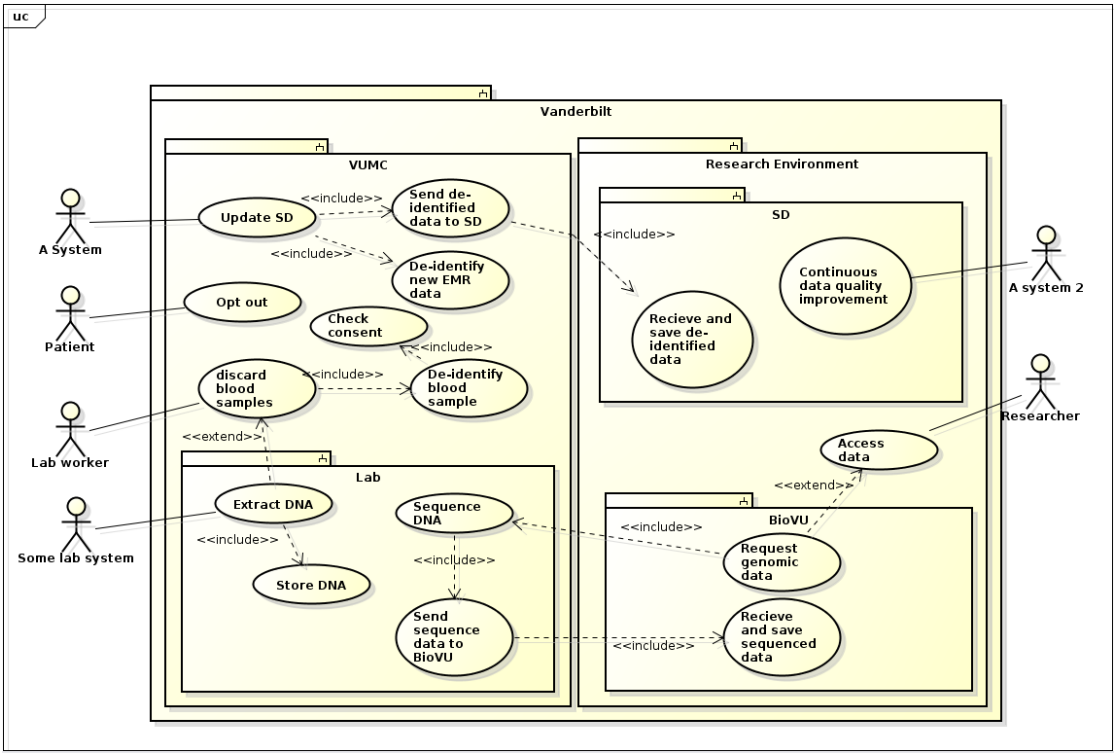


Figure 4.2: Use Case diagram of SD and BioVU

Description: Every patients data must be de-identified before it can be shared and used for research. De-identification for blood samples is done by hashing the MRN and relabeling the tube with the hashed number. The exact de-identification process is described in Section 4.2.3. This use case includes the use case *Check consent.*

System: VUMC

Use Case Name: Check consent

Description: When a patient gets admitted, he or she has to sign a *Consent to Treatment Form.* On the last page, right above the signature line is an option to *opt out* of BioVU. If the patient does not check the box, his or her leftover blood can be used for genomic research later on. This consent is checked before a blood sample is about to be discarded.

System: Laboratory

Use Case Name: Extract DNA

Description: DNA gets extracted out of the blood samples so that it can be stored more easily. Storing the tubes themselves would be to impractical. This use case includes the use case *Store DNA*.

System: Laboratory

Use Case Name: Store DNA

Description: This use case is included by *Extract DNA* and handles the physical storage of the extracted DNA samples.

System: VUMC

Use Case Name: Update SD

Description: Data of new patients and new documents of existing patients have to be send to SD. This gets triggered by *A System.* This use case includes the use cases *De-identify new EMR data* and *Send de-identified EMR data to SD.*

System: VUMC

Use Case Name: De-identify new EMR data

Description: Every patients data must be de-identified before it can be shared and used for research. De-identification is done for patients documents to ensure a patients privacy rights. The exact de-identification process for EMR data is described in section 4.2.2. This use case is included by *Update SD.*

System: VUMC

Use Case Name: Send de-identified EMR data to SD

Description: De-identified patient data gets send to the SD where it is used for medical research. This use case is included by *Update SD.*

System: SD

Use Case Name: Receive and save de-identified data

Description: After the EMR data gets de-identified it gets send to the SD. Here it is received and saved for later use. The data is saved in a database and in an adequate form. This use case is included by *Send de-identified data to SD.*

System: Synthetic Derivative

Use Case Name: Continuous data quality improvement

Description: Natural language processing will always miss PHI, even if its only a small percentage. To meet privacy concerns, algorithms get executed on an irregular basis to increase the data quality of SD by searching and removing missed PHI. *A System 2* executes this use case.

System: Research Environment

Use Case Name: Access data

Description: To allow the researcher to use the information in the database, the data needs to made queryable and converted into an analyzable form. The *Researcher* accesses SD to perform his research. This use case is extended by *Request genomic data.*

System: BioVU

Use Case Name: Request genomic data

Description: When a patient is eligible for a project but his genomic information got not sequenced yet, a request for the sequencing of the genotype gets initiated. This use case includes *Sequence DNA.*

System: Lab

Use Case Name: Sequence DNA

Description: Before genetic information can be used for research it needs to be sequenced. Depending on the research project specific genotypes get sequenced from the stored DNA. This use case includes *Send sequenced data to BioVU.*

System: Lab

Use Case Name: Send sequenced data to BioVU

Description: After the sequencing of the genotypes, the genetic information gets send to BioVU for permanent saving and later use for research. This use case includes *Receive and safe sequenced data.*

System: BioVU

Use Case Name:Receive and save sequenced data

Description: After the DNA gets sequenced in the laboratory the genotype information gets send to BioVU. Here it is received and stored for research. The genomic information is saved in a database in an adequate form.

4.2 Workflow and Processes

The following section describes the processes that are related to secondary use in Vanderbilt. The first subchapter covers the authorization processes for SD and BioVU, as well a "feasibility determination process". The second subchapter highlights the de-identification process of textual data, while the third subchapter focuses on the de-identification and storage of genetic information and its usage. The fourth subchapter explains the data extraction process, more precisely the tasks a researcher has to perform in order to get his research cohorts and subsequently the data. The fifth subchapter shows ways, how patient data from the SD is shared to other sites and the last subchapter highlights the mechanism used to link the separate types of data.

4.2.1 Record Counter and Authorization Process

This section describes the steps that are needed to get approval for access to information and data from SD and BioVU. Additionally this chapter covers and starts with the process of the *Record Counter* which can be used to determine the feasibility of conducting research using the SD.

There are two different ways for researchers to get access to data at Vanderbilt, depending if only clinical data is needed or clinical data in combination with genomic data. It might be beneficial, prior to starting either *Authorization Process,* to determine if the preliminary work for the authorization is actually worth the time and effort. For this purpose Vanderbilt developed the *Record Counter*, which allows users, who have a Vanderbilt University ID to query the SD for aggregated information.

Record Counter: As can be seen in Figure 4.3, the Researcher dispatches a query, for example "male patients" and the ICD code for headache. The Record Counter receives and processes the query before it sends a response with the aggregated information back. The response is a number that has to be greater than five and contains limited demographic information like gender. The Researcher receives the result and determines if the SD has enough patients for his contemplated research project[6].

[6]Screenshots from the Record Counter can be found in the Appendix A, Figures **5.4, 5.5** and **5.6.**

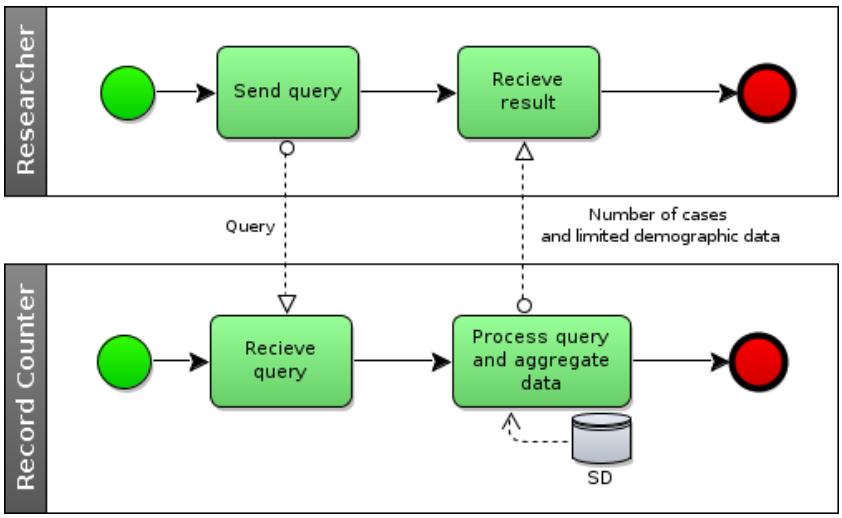


Figure 4.3: Record Counter

SD authorization: Figure 4.4 shows the business process for authorization to SD. First a researcher has to prepare a *research proposal* and submit it to the Institutional Review Board (IRB), which is indicated as a message flow in the diagram. After receiving and reviewing the proposal, the IRB committee has to decide whether to accept or decline it. If the proposal is declined a negative message get send back to the researcher. He then has the alternative to improve and resubmit it or end the process. The process ends *at End Event B,* if the researcher decides that he doesn't want to re-file the proposal.

Alternatively, if the IRB accepts the proposal a positive message gets send to the researcher whereupon he has to sign a Data Use Agreement (DUA), which forbids attempts of re-identification (among other things). A message that the DUA has been signed gets send to VUMC, which then in turn approves the authorization and gives the researcher access to SD. The process ends with *End Event A* in the case of a positive application.

BioVU authorization: The process for getting access to BioVU (alongside SD) is similar to getting access to SD only and is depicted in Figure 4.5. In addition to the SD authorization process, for access to BioVU, the proposal has to be send to the *BioVU Overview Board* (BioVU OB) as well. If either the IRB or the BioVU OB decline the proposal a message gets send to the researcher, that either one (or both boards) declined it and the researcher can decide if he wants to improve the proposal again. The process

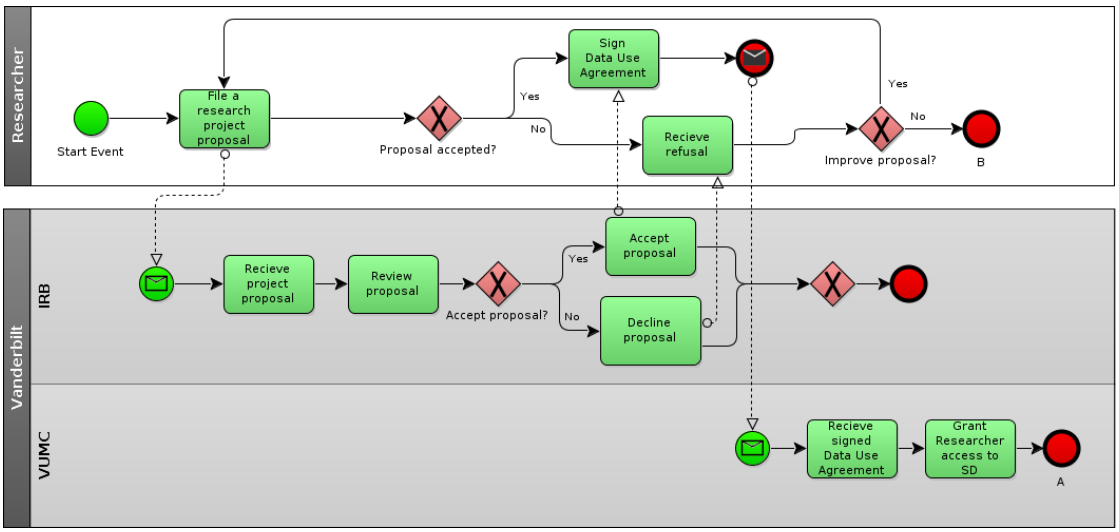


Figure 4.4: Authorization SD

ends at *End Event B*, if the researcher doesn't want to mend it.

However if both boards accept the proposal the researcher gets notified and has to sign a DUA as well and a message gets send to VUMC. After receiving the DUA VUMC will grant the researcher authorization to use both research databases for his project.

The two processes for data authorization and the Record Counter process are not represented in the use case diagram. These processes represent "policy processes" and the use case diagram focuses on the technical aspects.

4.2.2 De-identification Process

The following subsection will describe how patient information from the EMR is de-identified before it ends up in the SD. The de-identification process (Figure 4.6) covers the de-identification of different types of documents.

Three different kinds of patient data were found: *Structured data, Semi-structured data* and *Narrative text* documents, which could also be described as unstructured data.

The process starts with the hashing of the MRN. The hashing algorithm computes *h(MRN)*, which will always be the same for the same patient and doesn't allow a

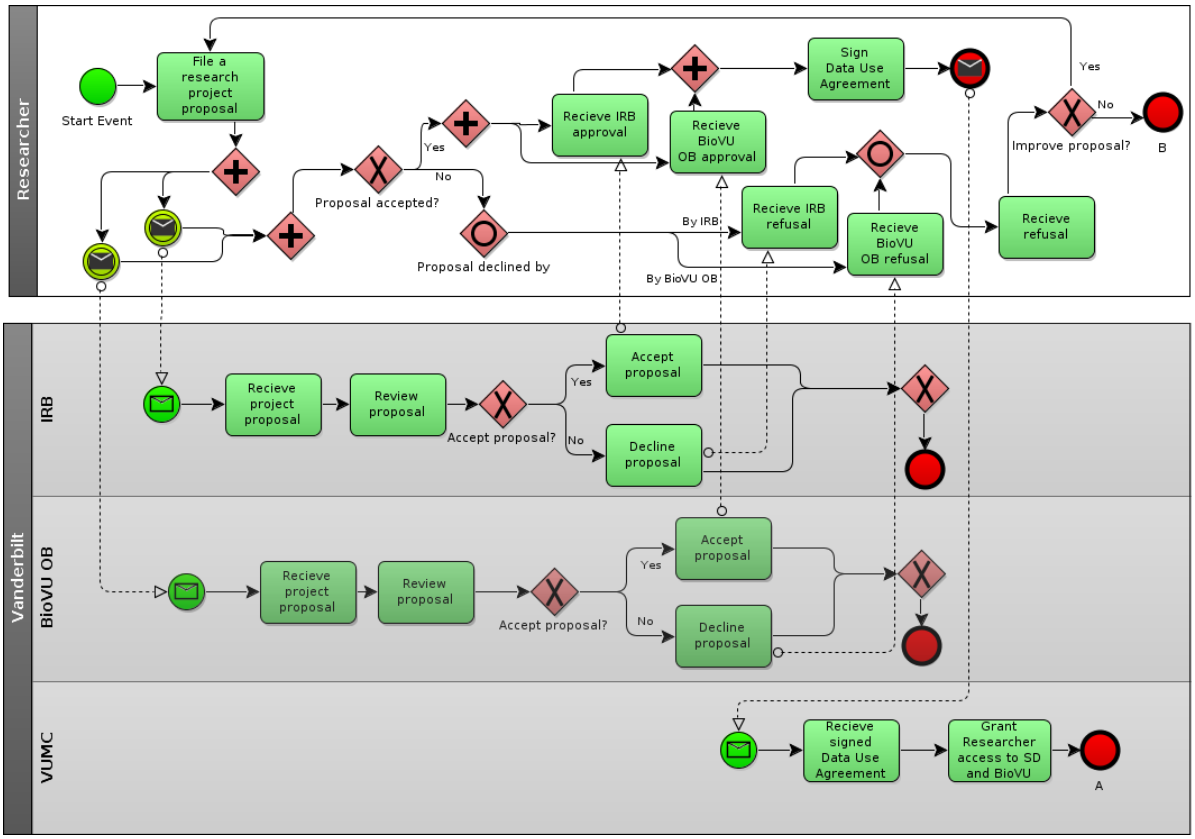


Figure 4.5: Authorization BioVU

conclusion to the original number. After the hashing, the h(MRN)-output gets used as input for the next task, which creates a *Date shift value*. The task uses a certain, undisclosed part of h(MRN) to calculate the value, which in turn gets used to de-identify date attributes later on. After the calculation of the date shift different routes are taken, depending if structured or non-structured documents have to be de-identified.

Non-structured documents first enter a *Pre-processing task* where the data gets converted into a specific input format for the following steps. Certain regions in the documents, that should not be changed, get tagged and PHI gets removed, where it is known that the next step, the actual de-identification task, misses them. The *De-identification* task uses a commercial tool for the natural language processing called *DE-ID* from DE-ID Data Corp.[7]. DE-ID replaces all PHI found with tags, for example

[7]See Chapter 2.2.2.

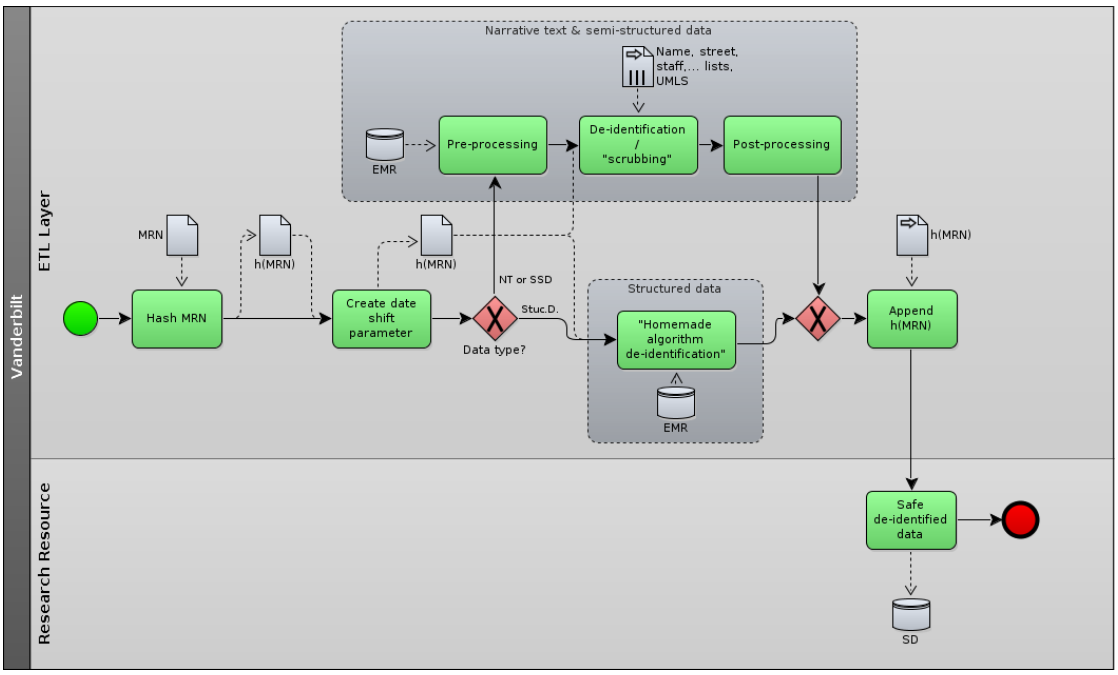


Figure 4.6: De-identification of medical data

NAME[XXX,YYY] and shifts all dates back in time. The "Date shift value" from earlier is used as input for the date shifting and miscellaneous lists, such as staff lists, street names, etc. as are used as input as well, to help with PHI identification. Libraries like UMLS get included to help distinguishing between names and diseases. An example is the name *Chron*, which could be a patients name but in combination with the prefixed word *Morbus* its apparent to the system that a disease it meant.

Structured data gets de-identified on a different way. Since the PHI locations are known, de-identification is done by the task "Homemade algorithm de-identification". Sophisticated and computationally intensive NLP algorithms are not needed.

After the de-identification of structured or non-structured documents, the h(MRN) gets appended to the data again, to ensure that all documents of a patient can be linked together. Before the process ends, the de-identified documents get stored in the SD for later usage by researchers.

The process is represented by the use cases *Update SD* and subsequently by *De-identify*

new EMR data, Send de-identified data to SD and *Receive and safe de-identified data.*

4.2.3 DNA Processes

The next two processes characterizes how genetic data ends up in BioVU and how it gets extracted again for research. Figure 4.7 shows how the genetic information gets stored while Figure 4.8 shows the usage process of the data.

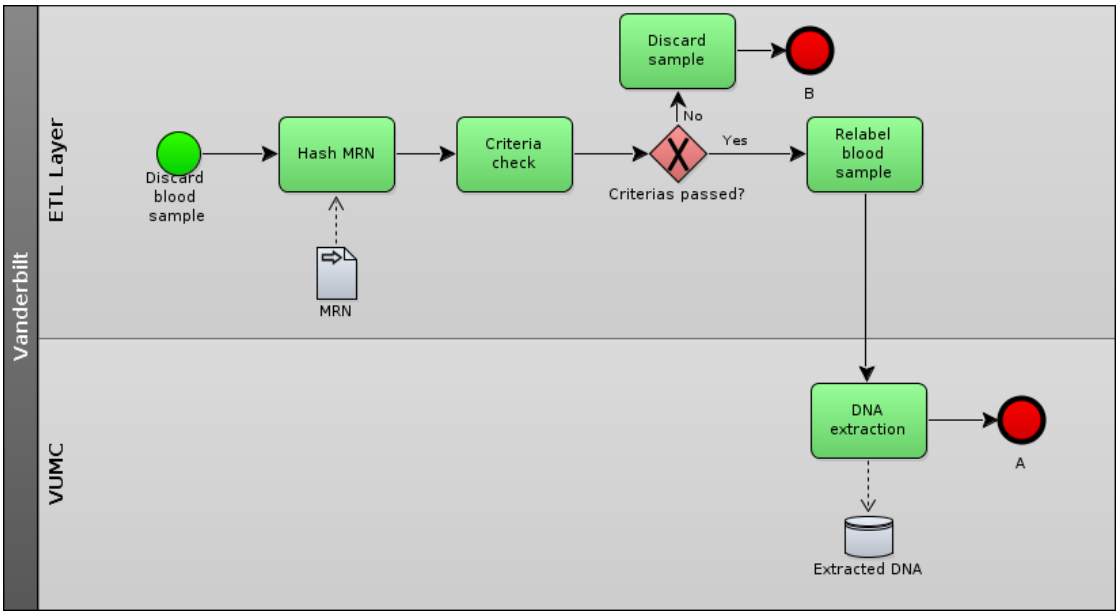


Figure 4.7: DNA storage

Blood samples that are about to be discarded still contain the MRN in form of a barcode on their tubes. These codes get scanned and hashed in task *Hash MRN*. Afterwards the samples get checked for eligibility[8] and if they pass the tubes get re-labeled with the hashed MRN-code h(MRN). The blood sample gets discarded as they normally would if they are not eligible and the process ends with *End Event B*. After the relabeling of the eligible tubes, the samples move on to the next task where the DNA gets extracted and stored in before the process ends. This storage is not BioVU yet, which only contains the sequenced data, but a store for physical DNA samples.

[8] A list of criteria for eligibility can be found in the Appendix Figure 5.2.

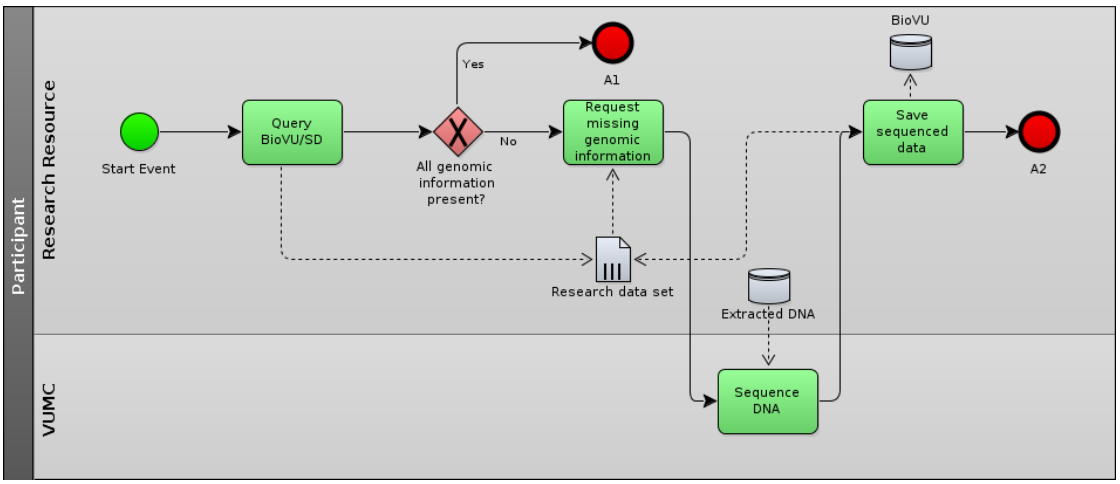


Figure 4.8: DNA usage

The DNA usage process starts with a query to SD/BioVU. If all the patients in the result set have the wanted genomic information in BioVU already, the process ends with *End Event A1.* Otherwise the task *Request missing genomic information* gets executed. The message flow from the collection *Research data set* to the task symbolizes the input of IDs, so that only DNA from not yet available patients gets requested. After the request, the DNA of the patients with the missing genomic information gets sequenced. The sequenced data gets added to the *Research result set* as well as to BioVU. This is done so that later research projects can reuse the existing data as well. The process ends with *End Event A2.*

The use cases *Extract DNA* with *Store DNA* and *Discard blood samples* with subsequently *De-identify blood sample* and *Check consent* represent the DNA storage process. *Request genomic data, Sequence DNA, Send sequence data to BioVU* and *Receive and save sequenced data* represent the DNA usage process.

4.2.4 Data Extraction Process

The Data extraction process describes the steps researchers have to take, in oder to get their research data out of the SD and BioVU. Preliminary work for authorization is dealt with in Section 4.2.1.

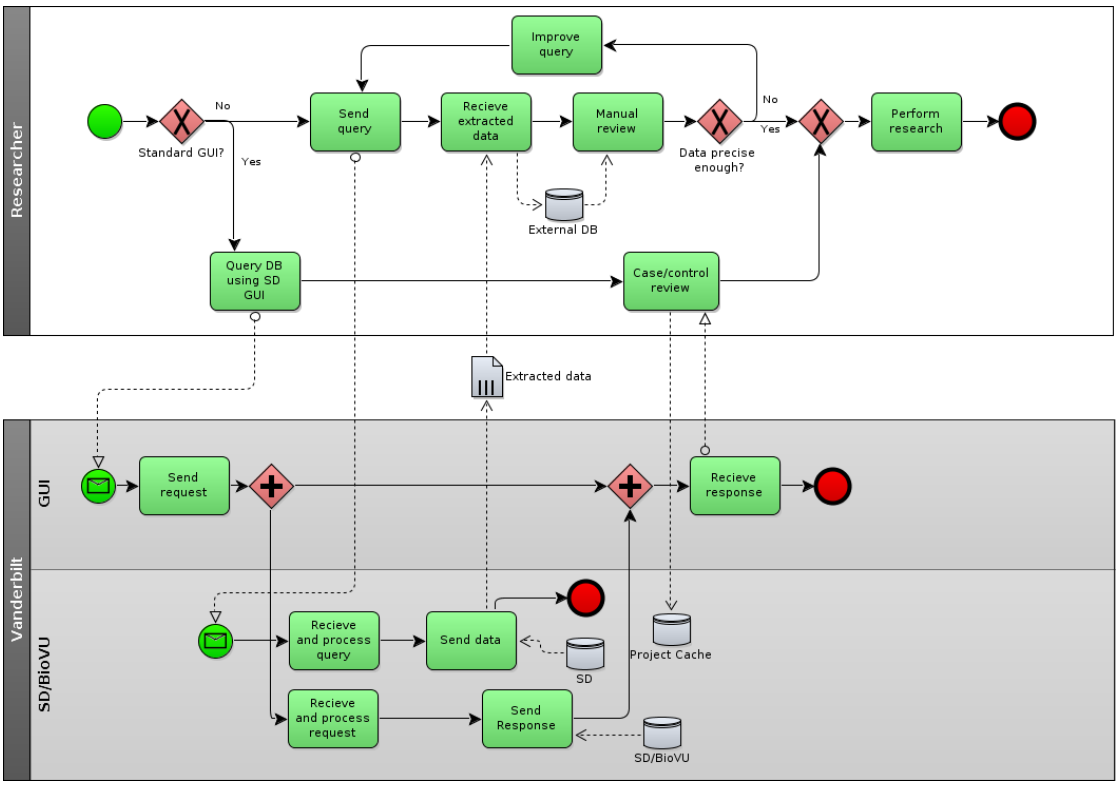


Figure 4.9: Data extraction process

There are two different ways a researcher can acquire his data from SD, either through the *standardized way* using the build in Graphical User Interface (GUI) or a *custom way* that includes extracting data out of SD. The GUI or tools for the custom way to view the extracted files have to be developed or bought separately. For the usage of genomic data only the standard GUI can be used.

The custom way starts with a researcher sending a query to the SD. The SD receives the query and processes it. Afterwards a result set gets send back to the researcher, who receives the *Extracted data* and stores it in an external database. The researcher then has to check the patient documents manually and decide if his query result is precise enough[9] or if he has to improve his query/query algorithm again. If the PPV for the

[9] A result set is not precise enough, if too many patients do not have the illness or condition, the researcher was looking for. False positive hits might occur for example when symptoms are similar for two different conditions or for wrong documentation. The Positive Predictive Value (PPV) thats high enough has to be set by the researcher themselves.

result set is high enough, the researcher can continue and start his actual research.

The standard way uses a GUI that is accessible to all researchers with access to the research databases. The researcher sends a request to the databases using the standard GUI. The DBs receive, processes and sends a response back. The researcher then gets a list of patients and has the option to view the patients documents and include them in his research record set or not. If the researcher uses BioVU and need DNA from patients sequenced, the process from the previous section gets executed. For simplicity of the process this was not modelled.

The difference between both approaches is that in the custom way data gets pulled out of the SD and stored in an separate DB, while the data stays in the SD for the standard way. The extraction requires the SD to search through all the data contained in the DB just once, while the standard way requires the SD to operate on the DB every single time a patient is accessed. This results in slower response times for the standard way. This advantage of the custom way gets compensated by the need for an own GUI and tools to process the extracted data. Another thing to consider is that the extracted data is object to privacy concerns, so the external DB has to ensure data privacy as well.

The process for data extraction is represented by the use case *Access data* in the Use case diagram.

4.2.5 Data Sharing

The following section describes how data that was once in Vanderbilt's EMR can be shared with other institutions, repositories or the researchers in general. The dotted arrow in the picture depicts a possible way of data sharing, while the continuous arrows describe actual scenarios.

As explained in Chapter 2.3.1 HIPAA defines three distinct methods of data sharing where a patients consent is not needed, *Safe Harbor*, *Limited data sets* and *Statistical methods.* Natural language processing can never be perfect and there is always a possibility that PHI are missed. In order to address this issue Vanderbilt considers the data in the SD a sort of Limited data set and constrains the researcher to sign a *Data use agreement*. The exact course of action for this method is described in Chapter 4.2.1.

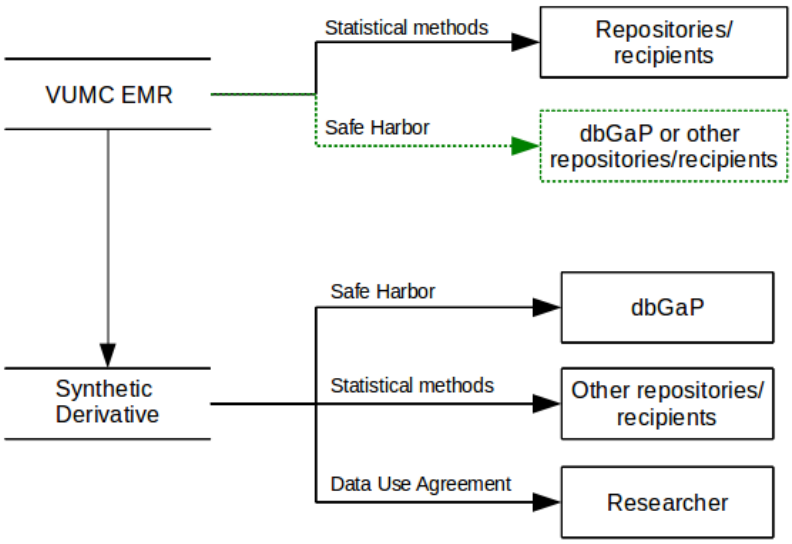


Figure 4.10: Data sharing diagram of Vanderbilt

A different way to data sharing is the aforementioned *Safe Harbor* method. Here all 18 PHI as defined by HIPAA have to be removed from the data sets, before the *covered entity*[10] is allowed to share the data. Public sharing is permitted with this method as well as sharing to public repositories. Vanderbilt uses this practice to share its data to dbGaP, the database of Genotypes and Phenotypes[11].

The more sophisticated way to share data are *Statistical methods*. Data sets that are de-identified with his method are shared to other repositories and recipients on a case to case basis by Vanderbilt. Source data can come from the identifiable EMR or the de-identified SD, as is depicted in the diagram.

A theoretical method for data sharing, but not used by Vanderbilt, would be using the Safe Harbor method on non de-identified EMR data. This possibility is pictured by a dotted arrow in the graph.

[10] See Chapter 2.3.1.

[11] Public repository for individual-level phenotype, exposure, genotype, and sequence data, and the associations between them created by National Center for Biotechnology Information (NCBI)[Kelley, 2008] which is part of NIH; http://www.ncbi.nlm.nih.gov/gap; visited: 13.07.2011.

4.2.6 Data Linking Method and Hashing

In SD and BioVU several different data types, like structured laboratory data, unstructured CPOE documentation or genomic information get linked together. This section highlights the linking mechanism that is used (Figure 4.11).

The main component in the linking of multiple documents is an undisclosed hashing function. The algorithm computes always the same output for the same input value, in this case a MRN, and the resulting h(MRN) can not be calculated back to its original number (= one-way-hashing function). The one-way-hash satisfies the privacy concerns, since the identity can't be backtracked and the h(MRN) ensures that all documents of one patient can be easily linked together.

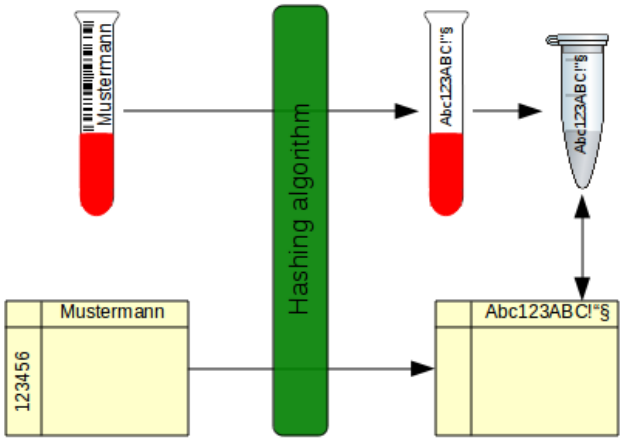


Figure 4.11: Linking method using hashing algorithm

4.3 Vanderbilt's Data Security Concept

As can be seen in the chapters above, the security concept of Vanderbilt consists of technological and policy means. This section also highlights the concepts in regards to (trail) re-identification.

Policy Means

On a policy level, the privacy gets protected by two major methods. First of all not everyone has access to SD and BioVU. Access is only granted to Vanderbilt employees and only after IRB and BioVU Overview Board/Operations Advisory Board approved of the research project.

Secondly, even though the documents in SD are de-identified users are required to sign a DUA. NLP algorithms will never find all PHI, so this method ensure that the privacy of the patients is better protected. The DUA requires missed PHI to be reported and forbids re-identification attempts and its violation will result in penal consequences. In this aspect SD's policy is similar to HIPAA's Limited Data Set, but not the same.

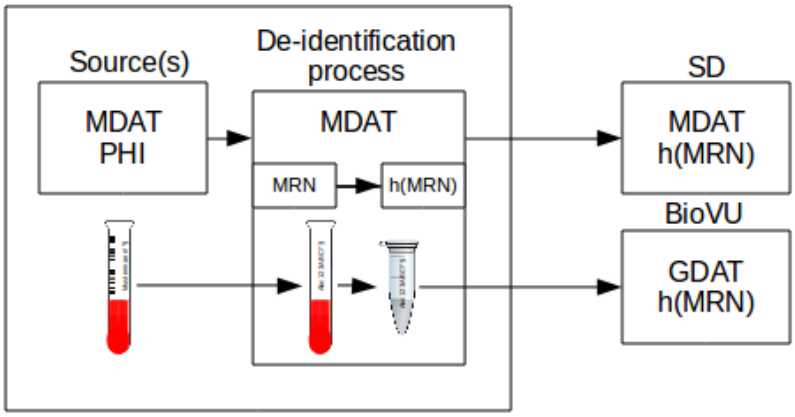
Technological Means

The first and most important technological method of securing the patients privacy is the de-identification[12]. Vanderbilt uses the methods of suppression, generalization and randomization described in Chapter 2.5. In SD suppression is done on identifiers like names and numbers like the Social Security Number for example. Those attributes are highly risky and don't serve any medical research purpose so they can get removed. HIPAA's privacy rule requires dates and ages over 89 years to be aggregated into a single category of age 90 or older, so DE-ID automatically uses generalization here. Vanderbilt uses a sort of randomization for its date representation in SD. Dates get shifted back in time by 1-365 days. This date-shift-value however is not completely random, but gets generated using the hashed MRN. The algorithm to creating the date-shift-value is not disclosed and the value can't be predict for a given MRN. So to the outsider the date shift appears to be random.

Another means of technologically securing the privacy is the project specific login and password that get created after the overview boards approve of the projects. Additionally all access is logged and the log-files can be checked if any inconsistencies appear.

The technological data privacy means can be roughly depicted as seen in Figure 4.12.

[12] See Chapter 4.2.2.



GDAT = Genomic Data
MDAT = Medical Data
PHI = Protected Health Information
MRN = Medical Record Number
h(MRN) = hashed MRN

Figure 4.12: Data privacy concept - technological means

Re-identification Risks / Trail Re-identification

The data saved in SD is de-identified, which means that names, IDs, geographical attributes, contact information,... got removed, dates shifted and ages changed respectively[13]. Chapter 2.5 highlighted that anonymous looking data doesn't necessarily protect a persons privacy. This chapter approaches the risk of re-identification for SD.

Since most of the identifiers specified in HIPAA get removed via suppression they are irrelevant for risk assessment. These specific attributes have no medical relevance, so it can be removed without any concern, with the exception of geographical data which could be used for epidemiological studies. Dates and gender are more interesting, since medical conditions can be gender constrained or more/less likely at certain ages. Furthermore time spans can be of interest, if studies are conducted where its evaluated how long it took for a certain therapy to be effective for example. A journalist re-identification scenario[14] is unlikely, since no attribute of the PHI list is left unchanged. For example, dates can't be match against those of other (public) databases and names are completely removed.

[13]Complete list see Appendix Figure 5.11.
[14]See Chapter 2.5.

Figure 2.5 shows that the likelihood of re-identification drastically sinks if the geographical region increases and even more if the age or DOB get less specific. Sweeney used three attributes for those numbers however, of which SD doesn't have the geography identifiers, the DOB is randomized and only the gender is available. With the DOB shifted back by up to a year, the plausibility for this attribute would lie between Birth year and 2yr Age, so even if a additional third attribute could be found to replace geography, the percentage of re-identification would most likely still be small.

Re-identification based on prosecutor re-identification scenario: Even if an "attacker" gets access to the data from SD it doesn't mean that he's automatically able to identify a person of his interest. No attribute of the PHI list is left unchanged, but since SD is not k-anonymous or k-unlinkable, no guarantee can be given that the attacker wouldn't find his target. However, if the attacker knows the visitation dates he could match those against visits from patients in SD and might be able to re-identify his target based on the intervals.

SD was not designed to ensure patient privacy using methods like k-anonymity or k-unlinkability, to avoid the loss of data quality which is inherent with those methods. Instead Vanderbilt applies policy and technological means, which allows for ample security while allowing maximum research possibilities. Furthermore is SD not meant to be publicly shared but supposed to be used as a means for medical research at Vanderbilt.

4.4 Law and Concept Comparisons

4.4.1 Comparison German Laws

A small analysis of the privacy laws of Germany for five of the 16 states (Baden-Württemberg, Bavaria, Hesse, Mecklenburg-West Pomerania and Rhineland-Palatinate) and the Federal Data Protection Act concluded that the usage of patient data for secondary use or research is permitted, with the general constraint that patients have to give consent first.

All state laws except the *privacy law* of Baden-Württemberg allow the usage and transmission of data after the patient gave consent. Baden-Württemberg permits transmission generally, if its for scientific research for the hospital.

Some paragraphs state that identifying patient attributes have to be saved separately from the medical data and be deleted (Mecklenburg-West Pomerania: Hospital Act Section 20 subsection 4) or anonymized (Bavaria: SDPA Section 23 subsection 3) once they are not needed for research anymore.

Baden-Württemberg's Section 50 subsection 1, among others, states that consent has to be obtained in individual cases and consent approval in general admission requirements is not permitted.

4.4.2 Comparison US - Germany

There are a few differences between US and German laws regarding patient privacy and data transmission.

Depending on the state in Germany, different laws apply (see table 2.1) while there is one major law in the US (HIPAA). On the one hand German laws leave the identifying attributes open, whereas in contrast HIPAA exactly specifies which criteria have to be removed and additionally that as an alternative statistic methods can be applied as well. On the other hand in the US re-identification (through means of pseudonyms) are not mandatory, while in Germany patients have to be informed if a study concluded that they have medical condition. German laws also state, that identifying attributes have to be kept separate from non-identifying attributes, while PHI can be scrubbed in the US, to satisfy legislature. As a last point some German state laws explicitly state, that patients have to give consent to the usage of data for secondary use on a research or project-related basis and that general permission for data usage is not allowed.

4.4.3 SD/BioVU with German Laws

The methods and concepts of SD and BioVU can not be copied to Germany one-to-one. Organizational processes like the Authorization process can be adopted, whereas technical processes such as the de-identification are not compatible with German laws.

After de-identification of textual data and blood samples, the usage of the data is considered "non-human subjects" research in the US, which means that it can be shared. In Germany however, data processing is generally only allowed, if permitted by law or

after a person gave consent[15]. The automatic transmission of anonymized data or data with pseudonyms from the EMR to a research database as done with the SD would not be possible. As mentioned in the chapter above, some German states forbid the consent form to be on the *general permission* for data usage. So at least in some German states, BioVU's opt out model would not be possible.

Vanderbilt's authorization process could be adapted to Germany as well, no laws were found that prohibit a procedure in this way.

Germany does not have a list of PHI, but identifying attributes have to be saved separately from medical data in some states, so the de-identification of EMR data can't be copied as a whole. However, if the processes that handle the de-identification tasks would be replaced by processes that separate MDAT and IDAT and anonymize or create pseudonyms (from the IDAT), the general concept could be carried over.

The GGDA states, that if a patient gave his consent, his genetic sample may be used for other purposes than it was acquired for, for example for a research database. Vanderbilt's processes for *DNA storage* and *usage* are applicable with German laws as well. Only the opt out approach would not be feasible and therefore can't be used.

Data extraction and *Data sharing* are applicable with German law as well, if the patient got informed before signing a consent that his non-identifiable data get used for research or send to other research sites.

4.4.4 SD/BioVU and TMF's Data Privacy Concept

This chapter compares SD's and BioVU's privacy concept against TMF's Models A, B and BMB (Chapter 2.4). When talking about Vanderbilt's concept it is referred to as it is depicted in Figure 4.12.

Model A

The biggest difference between Model A and Vanderbilt's data privacy concept is that the direct care context is not given in Vanderbilt.

[15] FDPA Section 4 Subsection 1.

The method for removing the link between a patients identity and his medical data also differs between the two models. In Vanderbilt the data gets de-identified, which means that PHI get removed or replaced with "synthetic information"[16], while Model A propagates a Patientlist at a TTP. There IDAT gets replaced by a PID and the link between IDAT and PID gets stored for later re-identification. Re-identification is not possible in Vanderbilt's concept, since no link between PHI and an identifier gets created and saved. The hashed MRN in Vanderbilt's concept can only be used to link different documents of the same patient[17].

Model B

Model B is more similar to Vanderbilt's concept than Model A, on one hand because the context is the same for both concepts. The data that will be saved in to the central DB (SD in Vanderbilt's case) is already entered into the hospitals EMR and a direct care context, like in Model A, is not given anymore. On the other hand both concepts differ in their underlying ideas and therefore in their methods used. The big difference is, that Model B allows re-identification and Vanderbilt's concept doesn't. Model B masks the link between identity and MDAT on two instances, first at a TTP where the IDAT get replaced by a PID and on a second TTP where a pseudonym is generated out of the PID, before the data enters a central DB. Since Vanderbilt doesn't have a re-identification option, the data gets de-identified and can be stored in SD. In Model B the MDAT get additionally sent to the central DB in an encrypted way. This extra step doesn't have to be done in Vanderbilt, since a TTP doesn't have to be passed through.

Model BMB

As explained in Chapter 2.2 TMF's Model BMB complies with Model B, so the similarities for medical data wouldn't be addressed again. Instead the focus will be on the biobank part of both concepts.

The *Sample DB* and *Analysis DB* from TMF's concept are similarly implemented in Vanderbilt. Samples are saved in the storage for extracted DNA samples (called *DNA*

[16]See Chapter 2.2.2 about de-identification for more detailed information.
[17]See Chapter 4.2.6 for more information.

storage in Figure 4.1) and the Analysis DB equals BioVU, where the sequenced information is saved.

The difference lies in the way the genetic data is accessed. After sequencing the DNA in Vanderbilt, the sequenced data and the hashed MRN get merged together again. In Model BMB however the sequenced data get saved with the ID AnaDAT. So when combining medical information and genetic data, the linkage is done by using the h(MRN) in Vanderbilt and not via a "chain of IDs" as its done in Modell BMB.

5 Discussion and Outlook

This chapter discusses the methods used in this thesis and the outcome of the results. Furthermore the issues that were initially addressed will be highlighted and at the end an outlook will be given.

5.1 Methods

At the beginning the objectives and exact goals of this thesis were not completely specified, so the initial literature research was very unspecific. However the first papers and the initial literature that was supplied gave an introduction and sense for the subject.

In regards to the legal texts the German states were not randomly picked. The state of Baden-Württemberg was chosen, because the University of Heidelberg and the ZIM are located there. Bavaria, Hesse and Rhineland-Palatinate were selected because they border Baden-Württemberg. Mecklenburg-West Pomerania was chosen because the ZIM has a cooperation with the University of Greifswald that's located there. Only five out of the 16 German states were picked, because the focus wasn't to be on law analysis, but to give insight on the judicial framework of Germany. The legal part was not the main focus of this thesis. Goal of that chapter wasn't a detailed and cast-iron analysis, but to give insight for the legal situation.

To get an overview of the judicial situation in the US not the HIPAA itself was read, but rather educational material provided by the National Institute of Health (NIH).

A five week stay in Vanderbilt was done to get a first hand experience and to research the secondary use systems. There interviews were held, a second literature research done and essential knowledge for this thesis gained. The Interviews were not done by any standard or standard form. The goal was information gathering and not a statistical evaluation, so the method seemed appropriate. To capture different perspectives, people from different

working-backgrounds were interviewed. This was also needed because different areas of expertise has to be covered like technical backgrounds and usage experience to name a few.

For the process diagrams the BPMN 2 standard was used for two main reasons. Introductory knowledge about the standard and its usage was already known and the information could be put down on paper in a more understandable way. Standards like UML 2 or Dataflow Diagrams[1] were taken into consideration and prototyped as well, but considered less suited and harder to understand upon review. The possibility that the usage for those standards was less sound can't be excluded, but using BPMN and not a Dataflow or Sequence diagram for example allows for a more strategical and less technical view. This in turn gives a broader range of people the chance to understand the diagrams.

The system architecture diagram was created in a non standardized way. No suitable standard could be found so it was created similar to architecture diagrams found in initial literature. To ensure correctness of depiction several reviews were done.

A use case diagram by the UML standard was created as well. It was made to allow a graphical representation of the functionality of the systems and to show the interaction between different actors.

5.2 Results and Conclusion

An analysis of the secondary use systems SD and BioVU from Vanderbilt was created and the comparison against German privacy concepts and legislature demonstrated that the processes would be compatible with Germany if some adaptions would be made. The biggest discrepancy derives from the differences in German and American laws, that in Germany patients have to be re-identifiable and that IDAT and MDAT have to be separately stored, which is not the case in the US.

The comparison of the processes also identified, that while technical aspects have to be different the overlaying procedures, like the Authorization Process, can be easily adopted. This would mainly impact the anonymization, where de-identification methods can't be

[1] Dataflow Diagrams by Yourden's Notation
http://yourdon.com/strucanalysis/wiki/index.php?title=Chapter_9; visited: 12.07.2011.

applied, but a generation of pseudonyms like suggested by the TMF would have to be implemented.

Storage and usage of genetic information would be compatible in general, but an opt out model like it is implemented in Vanderbilt would not be in accordance with German laws. In Germany the consent has to be acquired on a case-by-case basis and may not be included in the general admission form. However, when a sample is obtained, the same processes could be applied. The same is true for SD, with the exception of the de-identification processes, the general procedure of usage could adopted for, from the technical standpoint as well as the usage of the research environment.

The opt out approach for BioVU and the usage of EMR data from every patient leaves Vanderbilt with the advantage of accumulating huge amounts of clinical and genomic information, which other databases will not be able to collect. Due to this circumstance the databases make attractive targets, since an intrusion would be profitable because of the size, although its questionable if any patients privacy could be compromised because of the security measures undertaken.

Not all types of security risks for the data were covered in this work. Technical security aspects like regular software updates, secure connections and so forth are within the scope of the thesis, but the risk for SD from re-identification was discussed. A trail re-identification attack is less likely in Germany however. Public databases, like hospital discharge summaries or voter registries are not as common in Germany as they are in the US, so the access to information is not as easy to obtain.

This work showed that Vanderbilt's secondary use systems are at large compatible with German laws. Single methods have to be modified, but it was shown that the overall processes can be used in accordance with the TMF's data privacy concepts Model B and Model BMB.

Although, for the regulatory framework of the US, the implementation of Vanderbilt was investigated only, it was shown that procedures between both countries can be exchanged and collaboration for those types of projects is within the realm of possibility. Other projects or theses who investigate the topics of secondary use, research databases and its feasibility for collaboration between German and American institutions or sites can use this thesis as a starting point for their work.

5.3 Outlook

It has been shown that privacy concepts for research databases are similar in the US and Germany, and the hope is that this work helps find a common ground for institutions in both countries to work closer together.

With more and more projects and systems covering the subject of secondary use and the ever increasing costs in the health care sector in general, more standardization, collaboration and exchange of experiences is desirable. The TMF in Germany and the eMERGE network[2] in the US seems to be on a good way towards this goal on a national level.

It has yet to been seen, if the high hopes in secondary use systems get fulfilled, and the resources invested pay off. In a couple of years studies can be conducted that give conclusions about the cost-benefit-estimations and ascertain if secondary use data bases resulted in new opportunities for medical research.

[2] https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page; visited: 13.07.2011.

Bibliography

John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman. The MITRE Identification Scrubber Toolkit: design, training, and assessment. *International journal of medical informatics*, 79(12):849–59, December 2010. ISSN 1872-8243. doi: 10.1016/j.ijmedinf.2010.09.007. URL http://www.ncbi.nlm.nih.gov/pubmed/20951082.

Baden-Württemberg. Landeskrankenhausgesetz Baden-Württemberg, 2011. URL http://www.landesrecht-bw.de/jportal/?quelle=jlink&query=KHG+BW&psml=bsbawueprod.psml&max=true&aiz=true.

Bayern. Bayerisches Krankenhausgesetz, 2008. URL http://by.juris.de/by/gesamt/KHG_BY_2007.htm.

Bayern. Bayerisches Datenschutzgesetz, 2009. URL http://byds.juris.de/byds/009_1.1_DSG_BY_1993.html#009_1.1_DSG_BY_1993_rahmen.

Kathleen Benitez and Bradley Malin. Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association : JAMIA*, 17(2):169–77, March 2010. ISSN 1527-974X. doi: 10.1136/jamia.2009.000026. URL http://www.ncbi.nlm.nih.gov/pubmed/20190059.

BRD. Bundesdatenschutzgesetz, 2009a. URL http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf.

BRD. Gesetz über genetische Untersuchungen bei Menschen, 2009b. URL http://www.gesetze-im-internet.de/bundesrecht/gendg/gesamt.pdf.

V. Ciriani, S. De Capitani di Vimercati, S. Foresti, Samarati, and P. k-Anonymity. In *Secure Data Management in Decentralized Systems*, volume 2, pages 323—-353. January 2007. URL http://dx.doi.org/10.1007/978-0-387-27696-0_10.

K. El Emam and F.K. Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627, 2008. ISSN 1527-974X. doi: 10.1197/jamia.M2716.Introduction. URL http://jamia.bmj.com/content/15/5/627.full.

Dario a Giuse. Supporting communication in an integrated patient record system. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, page 1065, January 2003. ISSN 1942-597X. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1480157&tool=pmcentrez&rendertype=abstract.

Paul a Harris, Jonathan a Swafford, Terri L Edwards, Minhua Zhang, Shraddha S Nigavekar, Tonya R Yarbrough, Lynda D Lane, Tara Helmer, Laurie a Lebo, Gail Mayo, Daniel R. Masys, Gordon R Bernard, and Jill M Pulley. StarBRITE: The Vanderbilt University Biomedical Research Integration, Translation and Education portal. *Journal of biomedical informatics*, February 2011. ISSN 1532-0480. doi: 10.1016/j.jbi.2011.01.014. URL http://www.ncbi.nlm.nih.gov/pubmed/21310264.

Hessen. Hessisches Datenschutzgesetz. 1999. URL http://www.datenschutz.hessen.de/download.php?download_ID=14.

Hessen. Hessisches Krankenhausgesetz, 2011. URL http://www.rv.hessenrecht.hessen.de/jportal/portal/t/5pjx/page/bshesprod.psml?doc.hl=1&doc.id=jlr-KHGHE2011rahmen:juris-lr00&documentnumber=1&numberofresults=52&showdoccase=1&doc.part=R&paramfromHL=true#focuspoint.

K Kelley. The NCBI dbGaP database of genotypes and phenotypes. *Brain, Behavior, and Immunity*, 22(5):629–629, July 2008. doi: 10.1016/j.bbi.2008.05.010.

Bradley Malin. *Trail re-identification and unlinkability in distributed databases*. PhD thesis, 2006. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.2588&rep=rep1&type=pdf.

Bradley Malin. k-Unlinkability: A privacy protection model for distributed data. *Data & Knowledge Engineering*, 64(1):294–311, January 2008. ISSN 0169023X. doi: 10.1016/j.datak.2007.06.016. URL http://linkinghub.elsevier.com/retrieve/pii/S0169023X07001462.

Bradley Malin. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of Investigative*

Medicine, 58(1):11, 2010a. ISSN 1081-5589. doi: 10.231/JIM.0b013e3181c9b2ea. URL http://journals.lww.com/jinvestigativemed/Abstract/2010/01000/Technical_and_Policy_Approaches_to_Balancing.4.aspx.

Bradley Malin. Secure construction of k-unlinkable patient records from distributed providers. *Artificial intelligence in medicine*, 48(1):29–41, January 2010b. ISSN 1873-2860. doi: 10.1016/j.artmed.2009.09.002. URL http://www.ncbi.nlm.nih.gov/pubmed/19875273.

Bradley Malin, Kathleen Benitez, and Daniel R. Masys. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association : JAMIA*, 18(1):3–10, January 2011. ISSN 1527-974X. doi: 10.1136/jamia.2010.004622. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3005867&tool=pmcentrez&rendertype=abstract.

Daniel R. Masys. BioVU : clinically derived Biobank design, 2010. URL http://de-idata.com/sites/de-idata.com/files/UserFiles/PDF/Dan_Masys_NCRR_Pres.pdf.

Mecklenburg-Pommern. Landeskrankenhausgesetz für das Land Mecklenburg-Vorpommern, 2008. URL http://mv.juris.de/mv/gesamt/LKHG_MV.htm#LKHG_MV_rahmen.

S Meystre, F. Friedlin, S. Brett, S. Shen, and M Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10:70, January 2010. ISSN 1471-2288. doi: 10.1186/1471-2288-10-70. URL http://www.ncbi.nlm.nih.gov/pubmed/20678228.

National Institute of Health. Final NIH Statement on Sharing Research Data, 2003. URL http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.

NIH. HIPAA_Booklet, 2004. URL http://privacyruleandresearch.nih.gov/pdf/HIPAA_Booklet_4-14-2003.pdf.

Klaus Pommerening. Das Datenschutzkonzept der TMF für Biomaterialbanken (The TMF Data Protection Scheme for Biobanks). *it - Information Technology*, 49(6):352–359, November 2007. ISSN 1611-2776. doi: 10.1524/itit.2007.49.6.352. URL http://www.oldenbourg-link.com/doi/abs/10.1524/itit.2007.49.6.352.

Klaus Pommerening. Das TMF-Datenschutzkonzept Medizinische Forschung, 2009. URL http://www.datenschutz.fu-berlin.de/dahlem/ressourcen/datenschutzfachtagung-2009/Das_TMF-Datenschutzkonzept_f__r_vernetzte_medizinische_Forschung_und_Biobanken.pdf.

Klaus Pommerening and Michael Reng. Secondary use of the EHR via pseudonymisation. *Studies in health technology and informatics*, 103:441–6, January 2004. ISSN 0926-9630. URL http://www.ncbi.nlm.nih.gov/pubmed/17476066.

Klaus Pommerening, Johannes Drepper, Thomas Ganslandt, Krister Helbing, T. Müller, Ulrich Sax, Sebastian Semler, and Ronald Speer. Das TMF-Datenschutzkonzept für medizinische Daten-sammlungen und Biobanken. *subs.emis.de*, 2009. URL http://subs.emis.de/LNI/Proceedings/Proceedings154/gi-proc-154-137.pdf.

Rheinland Pfalz. Landeskrankenhausgesetz, 2011. URL http://rlp.juris.de/rlp/gesamt/KHG_RP.htm.

Marylyn D. Ritchie, Joshua C. Denny, Dana C. Crawford, Andrea H. Ramirez, Justin B. Weiner, Jill M. Pulley, Melissa a. Basford, Kristin Brown-Gentry, Jeffrey R. Balser, and Daniel R. Masys. Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. *The American Journal of Human Genetics*, 86(4):560–572, April 2010. ISSN 00029297. doi: 10.1016/j.ajhg.2010.03.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0002929710001461.

D M Roden. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical pharmacology and therapeutics*, 84(3):362–9, September 2008. ISSN 1532-6535. doi: 10.1038/clpt.2008.89. URL http://www.ncbi.nlm.nih.gov/pubmed/18500243.

C. Safran, M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, and D.E. Detmer. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *Journal of the American Medical Informatics Association*, 14(1):1, 2007. ISSN 1527-974X. doi: 10.1197/jamia.M2273.Introduction. URL http://jamia.bmj.com/content/14/1/1.short.

B Schütze and F Oemig. Nutzung von Patientendaten zur Forschung und Qualitätssicherung : Datenschutzrechtliche Fragestellungen, 2010.

Latanya Sweeney. Simple Demographics Often Identify People Uniquely. *Health*

(San Francisco), pages 1–34, 2000. URL http://dataprivacylab.org/projects/identifiability/paper1.pdf.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002. URL http://epic.org/privacy/reidentification/Sweeney_Article.pdf.

Dan Wasserstrom. A Commercial Approach to De-Identification, 2010. URL http://www.de-idata.com/sites/de-idata.com/files/UserFiles/PDF/Wasserstrom_OCRPresentation.pdf.

Appendix

A Further Pictures

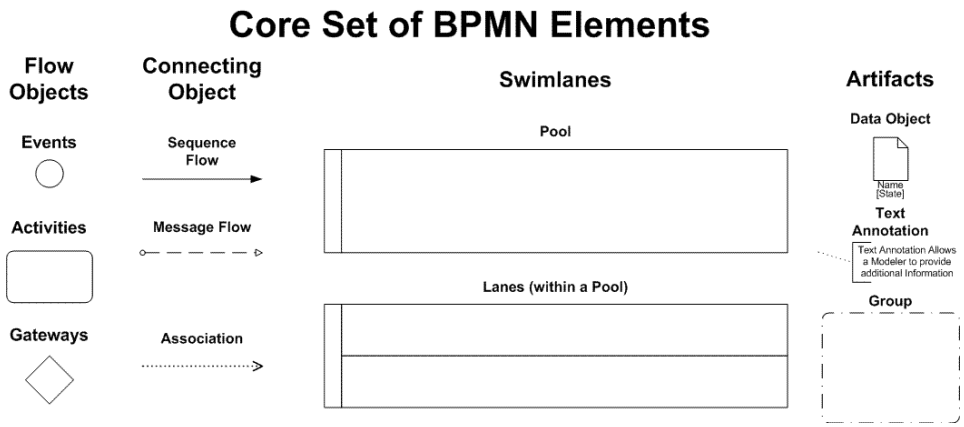


Figure 5.1: BPMN core elements
http://www.bpmn.org/Samples/Elements/Core_BPMN_Elements.htm; visited 28.08.2011

Table 2 Inclusion and exclusion criteria

Manual exclusions
Poor quality
Insufficient blood volume
Automated exclusions
Minors
No signed consent to treatment form (includes the emergency department)
Samples from individuals who opted out
Non-Vanderbilt samples
Duplicate samples
A percentage randomly selected

Figure 5.2: Inclusion and exclusion criteria for blood samples [Roden, 2008]

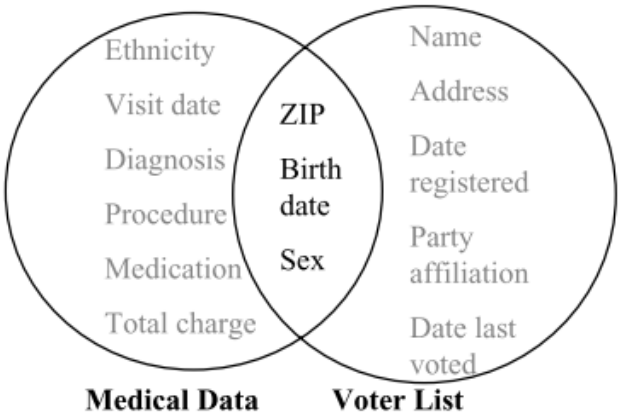


Figure 5.3: Linking to re-identify data [Sweeney, 2000, page 3]

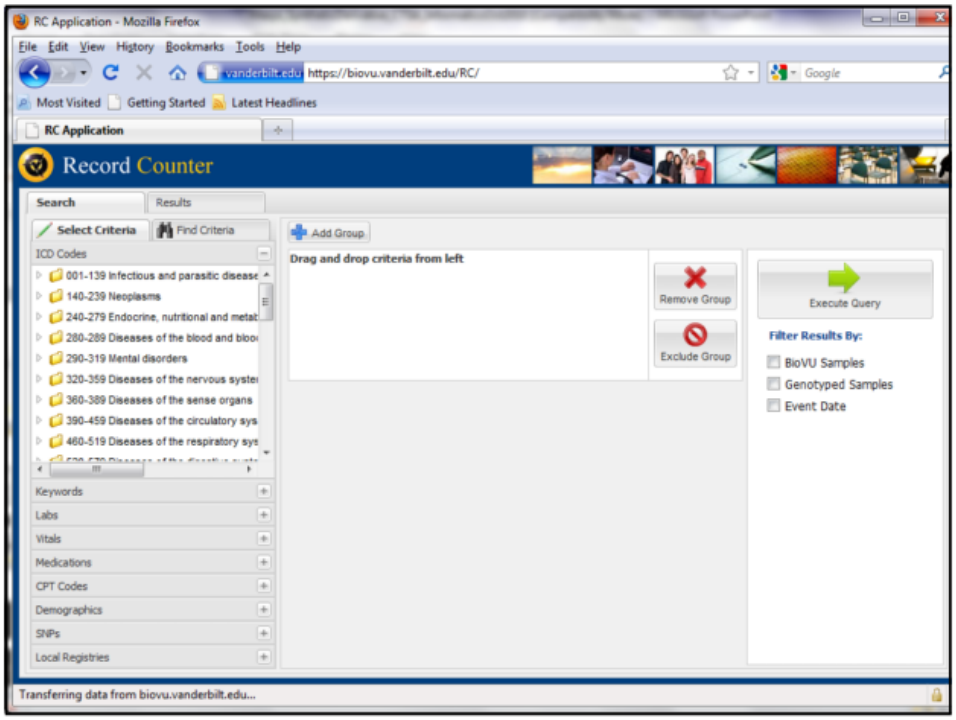


Figure 5.4: Record Counter query screen[Masys, 2010, slide 19]

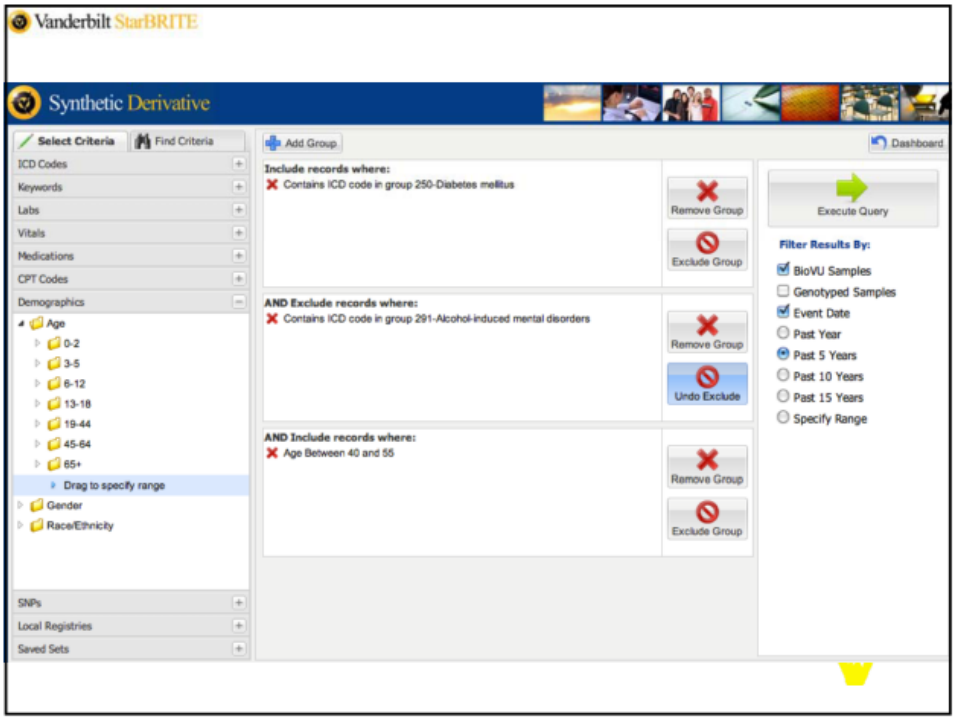


Figure 5.5: Record Counter - multiple query attributes [Masys, 2010, slide 20]

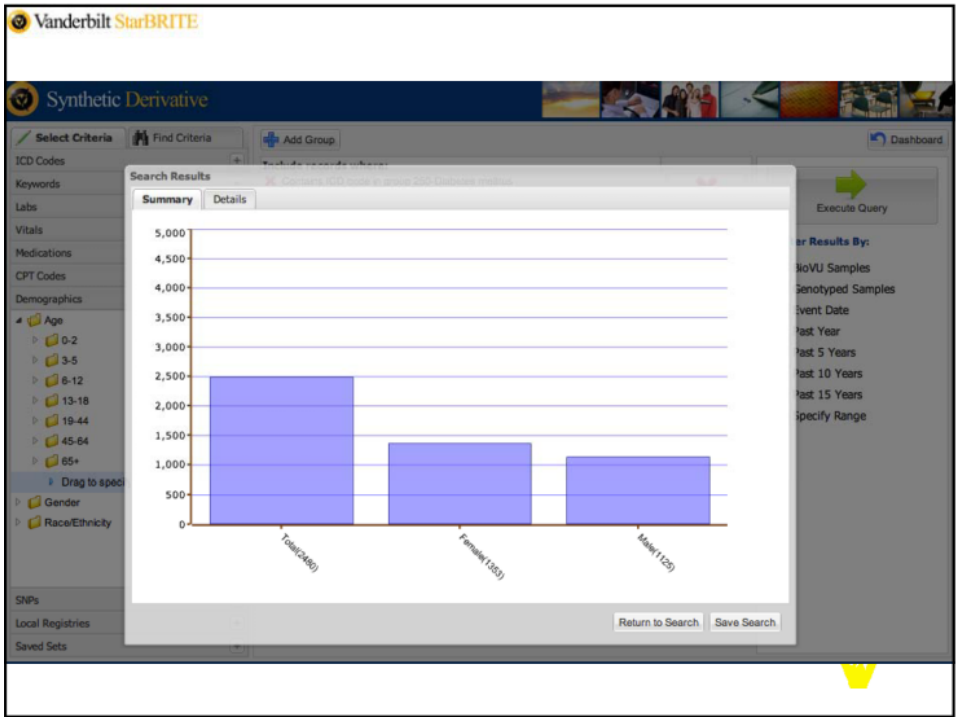


Figure 5.6: Record Counter - Result screen [Masys, 2010, slide 21]

Progress Note	DE-ID Progress Note
December 15, 2005. Rosalind Franklin is a 46 year old woman with a history of hypertension who was started on Norvasc on December 14 by Dr. Schwartz. She presented to the Hope Hospital ED today with a fine red pruritic maculopapular rash with no respiratory symptoms. Dr. Schwartz was called but unavailable, Dr. Lipton, covering advised D/C Norvasc and start Lisinopril 5mg	**DATE[Feb 14 2008]. **NAME[AAA BBB] is a **AGE[in 40s] year old woman with a history of hypertension who was started on Norvasc on **DATE[Feb 13] by Dr. **NAME[ZZZ]. She presented to the [**PLACE] today with a fine red pruritic maculopapular rash with no respiratory symptoms. Dr. **NAME[ZZZ] was called but unavailable, Dr. **NAME[YYY] covering advised D/C Norvasc and start Lisinopril 5mg

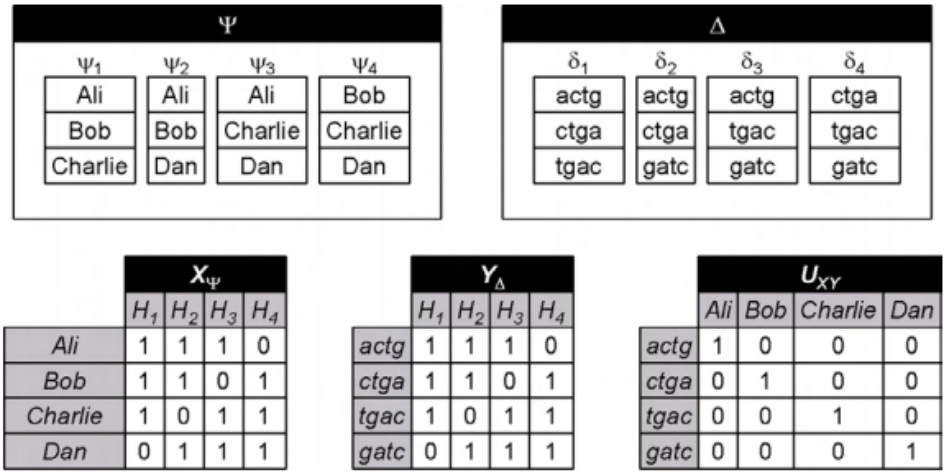
Figure 5.7: DE-ID example [Presentation Wasserstrom, 2010, slide 12]

Pre-scrub	After scrub
GI: soft, ND, normal bowel sounds, non tender, no hepatomegaly, no splenomegaly	GI: **PLACE, ND, normal bowel sounds, non tender, no hepatomegaly, no splenomegaly
with iron, 40 gm protein daily, and 1500-2000 calories daily.	with iron, 40 gm protein daily, and **ID-NUM calories daily.
Standarized Balance Tests: BERG Total score: 34 Pt required frequent rest	Standarized Balance Tests: **NAME[XXX: WWW] score: 34 Pt required frequent rest
An attending Cardiologist was present throughout the diagnostic study.	An attending **NAME[SSS] was present throughout the diagnostic study.
filled through the Easter Seals. The patient is also requesting additional	filled through the The patient is also requesting additional

Figure 5.8: Overmarking examples [Presentation Masys, 2010, slide 16]

Pre-scrub	After scrub	Error Type
Rx for Lortab 10, #60 w/ one refill 12/8/4	Rx for Lortab 10, #60 w/ one refill 12/8/4	Date (Complete but malformed)
SOCIAL HISTORY: He currently lives at 77 Spruce Loop; Crossville, Tennessee	SOCIAL HISTORY: He currently lives at 77 Spruce Loop; **PLACE, Tennessee	Street Address
DATE OF BIRTH: 02/22/1912	DATE OF BIRTH: **DATE[Jun 22 1912].	Age Over 90
number of the ventilator is 98141 Patient being monitored with oximetry The	number of the ventilator is 98141 Patient being monitored with oximetry The	Device ID
Severe Left Thigh Hematoma (Traumatic) 6/00	Severe Left Thigh Hematoma (Traumatic) 6/00	Partial date

Figure 5.9: Undermarking examples [Presentation Masys, 2010, slide 15]



Ψ	Ψ_1	Ψ_2	Ψ_3	Ψ_4
	Ali	Ali	Ali	Bob
	Bob	Bob	Charlie	Charlie
	Charlie	Dan	Dan	Dan

Δ	δ_1	δ_2	δ_3	δ_4
	actg	actg	actg	ctga
	ctga	ctga	tgac	tgac
	tgac	gatc	gatc	gatc

	X_Ψ			
	H_1	H_2	H_3	H_4
Ali	1	1	1	0
Bob	1	1	0	1
Charlie	1	0	1	1
Dan	0	1	1	1

	Y_Δ			
	H_1	H_2	H_3	H_4
actg	1	1	1	0
ctga	1	1	0	1
tgac	1	0	1	1
gatc	0	1	1	1

	U_{XY}			
	Ali	*Bob*	*Charlie*	*Dan*
actg	1	0	0	0
ctga	0	1	0	0
tgac	0	0	1	0
gatc	0	0	0	1

(a) Original tables

Ψ	Ψ_1	Ψ_2	Ψ_3	Ψ_4
	Ali	Ali	Ali	Bob
	Bob	Bob	Charlie	Charlie
	Charlie	Dan	Dan	Dan

Δ′	δ_1'	δ_2'	δ_3'	δ_4'
	actg	actg	tgac	ctga
	ctga	gatc	gatc	tgac

	X_Ψ			
	H_1	H_2	H_3	H_4
Ali	1	1	1	0
Bob	1	1	0	1
Charlie	1	0	1	1
Dan	0	1	1	1

	Y_Δ			
	H_1	H_2	H_3	H_4
actg	1	1	*	*
ctga	1	*	*	1
tgac	*	*	1	1
gatc	*	1	1	*

	U_{XY}			
	Ali	*Bob*	*Charlie*	*Dan*
actg	1	1	0	0
ctga	0	1	1	0
tgac	0	0	1	1
gatc	1	0	0	1

(b) 2-unlinkable tables

Figure 5.10: 2-Unlinkability example; [Malin, 2006, page 119]

B Listings

1. Names.
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census:
a. The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people.
b. The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Telephone numbers.
5. Facsimile numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan beneficiary numbers.
10. Account numbers.
11. Certificate/license numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification.

Figure 5.11: PHI as defined by HIPAA; [NIH, 2004, page 14]

1. Names.
2. Postal address information, other than town or city, state, and ZIP Code.
3. Telephone numbers.
4. Fax numbers.
5. Electronic mail addresses.
6. Social security numbers.
7. Medical record numbers.
8. Health plan beneficiary numbers.
9. Account numbers.
10. Certificate/license numbers.
11. Vehicle identifiers and serial numbers, including license plate numbers.
12. Device identifiers and serial numbers.
13. Web universal resource locators (URLs).
14. Internet protocol (IP) address numbers.
15. Biometric identifiers, including fingerprints and voiceprints.
16. Full-face photographic images and any comparable images

Figure 5.12: Limited Data Sets; [NIH, 2004, page 20]

:ADM:01H
:ID:123456789
:NA:Smith, John
:DOC:Denny, Josh
:DSI:denndxi 2011/07/22 13:21
:TYP:HP
:STYP:Procedure Note
:DAT:2011/07/18
:HP: PROCEDURE NOTE
This is a very-pleasant 55-year-old Vanderbilt professor of law who is here for a thoracentesis for a recurrent plural effusion.
[rest of note]
:UNIQ:b246f8i90
:DB:vumc
:REST:
:STATUS:Unverified transcription
:E_O_R:

Figure 5.13: MARS message example

Universität Heidelberg
Hochschule Heilbronn
Medizinische Informatik

Studiengang Medizinische Informatik
Diplomstudiengang Informationsmanagement in der Medizin

Doods, Justin (Name, Vorname) 164724 (Matrikelnummer)

Thema der Diplom-/Masterarbeit: Evaluation of the Secondary Use approach from Vanderbilt and its usability for Germany

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.
Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistung von folgenden Personen erhalten:

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und ist auch noch nicht veröffentlicht.

(Ort, Datum) (Unterschrift)