

iCARDEA

“An Intelligent Platform for Personalized Remote Monitoring of the Cardiac Patients with Electronic Implant Devices”

SPECIFIC TARGETED RESEARCH PROJECT

PRIORITY Objective ICT-2009.5.1: Personal Health Systems - a) Minimally invasive systems and ICT-enabled artificial organs: a1) Cardiovascular diseases

iCARDEA Deliverable D7.2.1 Data Analysis and Correlation Tool

Due Date: January 31, 2012
Actual Submission Date: January 31, 2012
Project Dates: Project Start Date : February 01, 2010
 Project End Date : January 31, 2013
 Project Duration : 36 months
Leading Organization: *Contractor* OFFIS

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Document History:

Version	Date	Changes	From	Review
V01	January 14, 2012	Draft version of the Deliverable	OFFIS	SRDC / SALK
V02	January 20, 2012	Integrated suggestions of SRDC	OFFIS	Internal
V03	January 27, 2012	Pre-final version for review	OFFIS	All Partners
V04	January 31, 2012	All partner reviews consolidated	OFFIS	All Partners

iCARDEA Consortium Contacts:

SRDC	Asuman Dogac	+90-312-2101393	+90(312)2101837	asuman@srdc.com.tr
OFFIS	Wilfried Thoben	+49-441-9722131	+49-441-9722111	thoben@offis.de
SRFG	Manuela Plößnig	+43-662-2288-402	-	manuela.ploessnig@salzburgresearch.at
FORTH	Catherine Chronaki	+302810391691	+302810391428	chronaki@ics.forth.gr
SALK	Bernhard Strohmer	+43-6624482-3481	+43-6624482-3486	b.strohmer@salk.at
SJM	Karl Eberhardt	+43-16073067	-	keberhardt@sjm.com
Medtronic	Alejandra Guillén	34916250361	+34913346453	alejandra.guillen@medtronic.com
HCPB	Josep Brugada	+34932275703	+34932275459	jbrugada@clinic.ub.es

Table of contents

1	PURPOSE	5
1.1	Scope.....	5
1.2	Definitions and Acronyms	5
2	Introduction	6
2.1	iCARDEA System Architecture	8
2.2	iCARDEA Data Analysis System Architecture.....	11
3	Building the Data Analysis Process	13
3.1	IdentifiCation of User Needs	13
3.2	IdentifiCation of available Data sources.....	13
3.2.1	Research on existing datasets.....	14
3.2.2	SALK	16
3.2.3	Physionet.....	16
3.3	INTEGRATION OF DATA SOURCES and Data Preperation	16
3.3.1	Problem of different and changing codings	17
3.3.2	SALK	18
3.3.3	Physionet.....	19
3.4	Builduing Data Analysis	20
3.4.1	OLAP on SALK.....	20
3.4.2	Data Mining on SALK.....	24
4	Design & Implementation	25
4.1	DACT User Interface.....	25
4.2	Software needed for pattern generation	28
4.2.1	Database for Medical Knowledge.....	29
4.2.2	Database for Pattern.....	29
5	CONCLUSION	30
6	Appendix	31
6.1	SAlk USer Forms.....	31
6.2	Results of OLAP Cubes	35
6.3	Physionet descriptions	36
6.3.1	Research of useful MIMIC II Data	36
	Tables:.....	36
	Definitions:	36
	Content:.....	38
6.3.2	Definition-Tables:	39
6.3.3	Content:.....	41
6.3.4	Keyword for cardiac Patients.....	45
6.4	Literature.....	47

1 PURPOSE

1.1 SCOPE

Integrating and harmonizing of patient data enables modern information technology to make sophisticated use of the data. Especially when patient data is available over a long time period, this enables new forms of treatment processes but also it is supporting the treating healthcare actor and the medical research abilities. In Task 7.2 the focus is on the last two mentioned points.

This document describes the achievements in Task 7.2 “Data Analysis and Correlation”. Within the scope of this task, a tool should be provided to the healthcare actors, that correlates the parameters of a current patient with knowledge obtained from other, previous medical cases.

Therefore as a first part, a data analysis process for the medical cases had to be established, that creates statistically valid patterns. They are used for the purpose of making suggestions to the healthcare actors and are also useful for understanding the patient treatments.

The second part is a tool to present this newly obtained knowledge to the healthcare actors with respect to the current treated patient and show suggestions suitable for the patient. Since the data is collected over a longer time period, the semantic changes that are inherited to the data were a major topic addressed at the research to provide suitable suggestions and make the patterns applicable at the evaluation clinic at Salzburg.

1.2 DEFINITIONS AND ACRONYMS

Table 1 List of Abbreviations and Acronyms

Abbreviation/ Acronym	DEFINITION
CDA	Clinical Document Architecture
CIED	Cardiovascular Implantable Electronic Device
CM	Care Management
CRISP-DM	Cross-Industry Standard Process for Data-Mining
DACT	Data Analysis and Correlation Tool
DWH	Data WareHouse
EHR	Electronic Health Record
HL7	Health Level 7
ICD9	International Statistical Classification of Diseases and Related Health Problems – 9 th Edition
ICD10	International Statistical Classification of Diseases and Related Health Problems – 10 th Edition
ICD10-GM	International Statistical Classification of Diseases and Related Health Problems – 10 th Edition – German Modification
IDCO	Implantable Device Cardiac Observation
IHE	Integrating the Healthcare Enterprise
ISO	International Standards Organization
KDD	Knowledge Discovery in Databases
NIH	National Institutes of Health (USA)

OLAP	OnLine Analytical Processing
PHR	Personal Health Record
PIX	Patient Identifier Cross-Referencing
PPM	Patient Parameter Monitor
SAML	Security Assertion Markup Language
SEMMA	Sample, Explore, Modify, Model, Assess – Process of data mining
XACML	eXtensible Access Control Markup Language
XML	eXtensible Markup Language

2 Introduction

Data analysis is a process with the goal of highlighting information, suggesting conclusions and supporting decision making. It has multiple approaches and encompasses diverse techniques under a variety of names. For example, Data Mining is a data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. Online Analytical Processing (OLAP) covers data analysis that relies heavily on aggregation and exploration of multi-dimensional data.

Data analysis can be used for confirming or falsifying existing hypotheses, discovering new features in the data or generating of statistical or structural models for predictive forecasting or classification¹. The used data analysis technique depends on the aimed analysis objective. Most of the statistical algorithms can handle data from various sources and of different kind. Normally all observed parameters that should be used as input for the data analysis must be discredited and stored in a structured way [Han2006].

The standard process in data analysis project is as follows [Azevedo2008]:

- Business Understanding,
- Data Understanding with the steps data selection, preprocessing, data preparation,
- Modelling (the step where the statistical algorithms are used),
- Evaluation / Interpretation of the models and at last
- Deployment of the results.

¹ Data analysis overview by Wikipedia from May 2010 http://en.wikipedia.org/wiki/Data_analysis

This process is the same in all three standard processes in data analysis [Azevedo2008], named CRISP-DM, SEMMA² and KDD³. In Figure 1 the de facto industrial standard process for data mining in CRISP-DM is shown⁴.

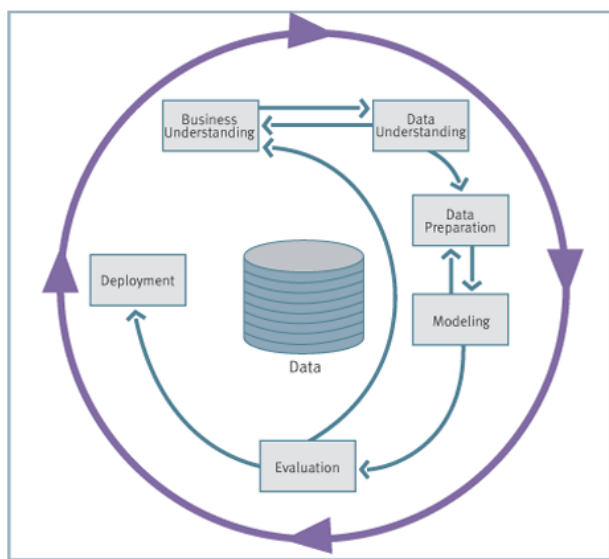


Figure 1 CRISP-DM Process Model⁵

Data stands at the beginning and in the center of the data analysis process. To obtain reliable results, a statistically sufficient volume of suitable data is needed. 80% of the effort in data analysis projects is used for the data preparation [Bauer2008]. Therefore the data is often provided by specialized databases, called data warehouses (DWH). Data warehouses serve as the data centers providing integrated, cleaned and structured data from various sources to all different kinds of data analysis techniques. The different observed real world parameters for analysis purposes are stored in so called dimensions which provide exploration information about the data and are accessible via OLAP-services.

By that way the data can be reused in different data analysis projects [Bauer2008]. At data analysis projects the appropriate data has to be chosen based on the business and data understanding and integrated into a DWH. Then the data has to be transformed to the required formats of the statistical algorithm, chosen for the aimed objective.

In practice the adaption of this process to iCARDEA was as follow:

- Business Understanding →
 - What do the healthcare professionals need?
 - What are the questions of the healthcare professionals to be answered by data analysis?
 - What Data is needed to answer these questions?
- Data Understanding with the steps data selection, preprocessing, data preparation →
 - What Data is available?

² Official SEMMA site by the SAS Institute last visited 1st May 2010 <http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>

³ [Fayyad1996]

⁴ Poll of main methodology used for data mining from August 2007 http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

⁵ Official CRISP-DM website last visited 1st May 2010 <http://www.crisp-dm.org/Process/index.htm>

- In what kind of format, under which conditions (time, legal regulations) is it available?
- How can the data be stored (codings, classifications)
- Modelling (the step where the statistical algorithms are used) →
 - Based on the available data, data analysis processes are produced for
 - OLAP: The so called “cubes” were defined and the data was integrated for this to answer questions about the kind of treatment
 - Data Mining: A data mining process was defined, to identify correlations at the patient data
- Evaluation / Interpretation of the models
 - The results of the data analysis have to be prepared and explained to medical experts,
- Deployment of the results
 - The results are, if found to be potential useful, are deployed to the iCARDEA architecture, to be presented to the healthcare actors when they are treating a similar patient.

Since it is important for data analysis to have comparable data items with respect to syntax and also semantic, research efforts were spent on developing a theory and tool to provide a visual helper to identify semantic changes in the data over the time. This supports the data analyst in the step of data understanding and for iCARDEA also at the deployment step, when the patterns should be adapted to the iCARDEA platform.

2.1 ICARDEA SYSTEM ARCHITECTURE

The iCARDEA system aims to automate and personalize the follow-up of cardiac arrhythmia patients with implanted CIED devices with computer interpretable clinical guideline models using standard device interfaces and integrating patient EHRs. Figure 2 shows the overall architecture and the environment in which iCARDEA provides interoperation services.

The tools developed in Workpackage 7, called the Patient Parameter Monitor (PPM) in Task 7.1 and the Data Analysis and Correlation Tool (DACT) in Task 7.2 are intended to provide the Healthcare actors an integrated view of all available patient related information, including patterns based on historical patient cases. Since this is one main point to access the patient data, there has to be arrangements to ensure the privacy of the data. For the creation of pattern, historical cases have to be analyzed by specialists, who are normally not allowed to access patient specific data using separate medical knowledge bases (see Figure 11). Therefore privacy and security mechanisms have to be provided at two stages: At the iCARDEA system itself which is installed at the hospital and also at the infrastructure of the data analysis process which may be located at an external data center. These topics are discussed in detail in Deliverable 7.3.1.

The major components of the iCARDEA system, in which PPM and DACT are integrated, are as follows:

- Personalized Adaptive Care Planner for the CIED Recipients: In the iCARDEA project, the personalized follow-up of CIED patients is coordinated through a “care plan” which is an executable definition of computer interpretable clinical guideline

models. The care plans are represented in GLIF, and the Care Plan Engine is capable of semi-automatically executing the care plan by processing its machine processable definition. The control flow of the care plan is dynamically adapted based on the patient's context derived from the data coming from CIEDs and the medical context obtained from the EHRs. Through a graphical monitoring tool, the physicians are allowed to follow the execution of the care plan in detail, and coordinate the flow of actions when consultations to physicians are required. Also the Adaptive Care Planner provides a central Care Management Database, where patient related data is collected from the different components and stored.

- The CIED Data Exposure Module uses "IHE Implantable Device Cardiac Observation Profile (IDCO)" to expose the CIED data from different vendors in a machine processable format to be used in the care plan of the patients and for the presentation of patient data at the PPM. For this, it has a component that allows accessing the CIED Portal of the vendor and triggers the CIED data export from the CIED Data Center. It extracts the CIED data from vendor specific formats and the Data Translation Service sub-system creates a valid IHE IDCO format (HL7 v2.5 ORU Message) and makes the CIED data available to the iCARDEA Adaptive Care Planner through PCD-09 Send Observation message.
- EHR Interoperability Infrastructure: To execute the clinical guidelines and to provide the healthcare actor with complete patient information, it is also necessary to have access to medical history of the patients in the EHR systems. Considering that there are many EHR systems with proprietary interfaces, in iCARDEA, "IHE Care Management (CM) Profile" is used. In our system, the proprietary hospital information systems export "Discharge Summary" and also "Laboratory Report Summary" CDA documents in conformance to IHE CDA Document templates⁶ to an EHR Server which is implemented as an IHE XDS Repository⁷. This EHR Server also acts as a "Clinical Data Source" by implementing the IHE CM Profile. In this way, Adaptive Care Manager can subscribe to receive update notifications for the clinical data that is necessary to execute the care plans. IHE Care Management Profile specifies standard interfaces to extract this data that is needed by the care plans from the EHR systems.
- There is also a Patient Empowerment component that aims to provide active and informed involvement of patients in management of their own health. Through the web based PHR, patients are able to view their medical history, CIED data, and manage their medication summaries, daily nutrition information.
- The Healthcare Professional is also supported by a single point of information access for patient data. This Patient Parameter Monitor (PPM) developed within the Task 7.1, provides all data collected to one patient from the EHR, PHR or CIED Integration via the Adaptive Care Planner Engine to the Healthcare actor. The PPM contains also a link to the Data Analysis and Correlation Tool, where, based on the parameters of a patient, statistically valid patterns are provided to the healthcare actors. For creation of the patterns, historical cases obtained either from existing clinical knowledge bases or from the legacy HIS are stored into a data analysis database. This knowledge base is only available to data analysts and possible healthcare actors.

⁶ IHE Care Coordination Framework, Content Modules, [http://wiki.ihe.net/index.php?title=1.3.6.1.4.1.19376.1.5.3.1.1#Medical Documents Specification 1.3.6.1.4.1.1_9376.1.5.3.1.1.1](http://wiki.ihe.net/index.php?title=1.3.6.1.4.1.19376.1.5.3.1.1#Medical_Documents_Specification_1.3.6.1.4.1.1_9376.1.5.3.1.1.1)

⁷ IHE Cross Enterprise Document Sharing (XDS) Profile, http://www.ihe.net/Technical_Framework/index.cfm#IT

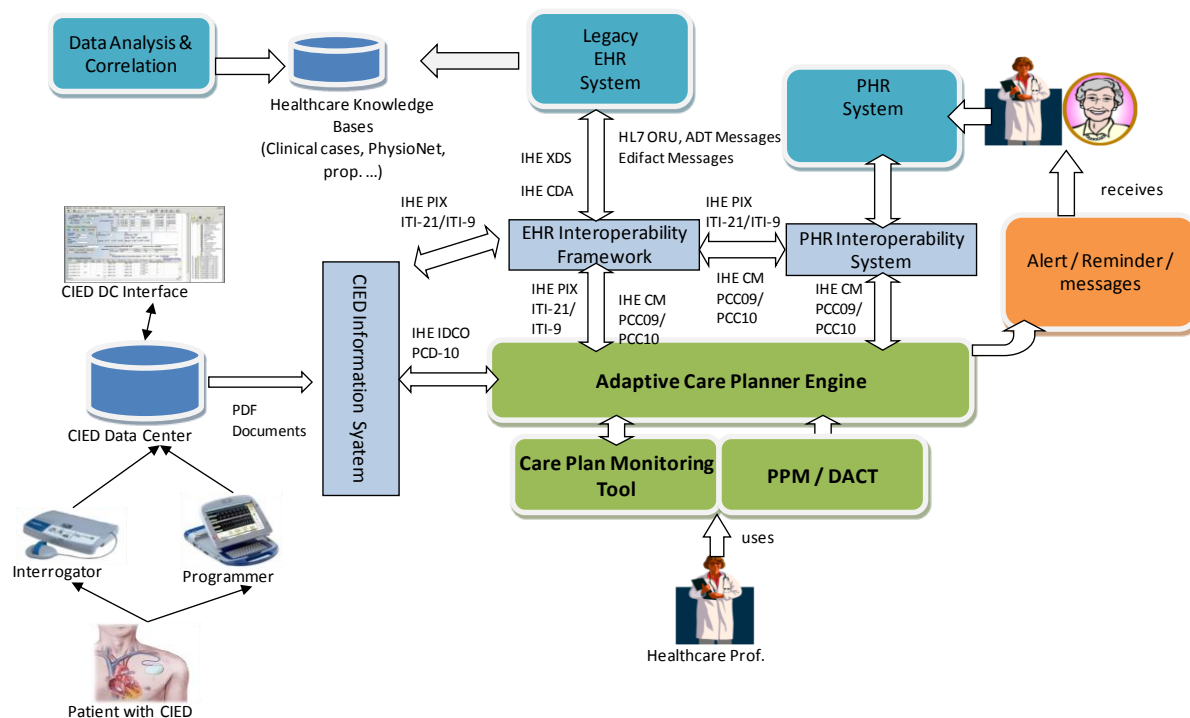


Figure 2 iCARDEA Architecture Overview

Clinical practice guidelines present and formalize medical knowledge required for clinical decision-making and try to standardize the patient care delivery by guiding the healthcare practitioners regarding next actions to be performed. Despite the potential benefits of the clinical guidelines, at the moment they are underutilized in clinical practice due to interoperability problems of healthcare data sources. iCARDEA exposes healthcare resources such as electronic healthcare records (EHR), personal healthcare records (PHR) and CIED data of a patient through standard interfaces. Furthermore, the integration problem does not occur only between EHR and PHR systems. Care providers may provide care plans using different types of codes systems, and these code systems may not be used in patients' EHR and PHR systems.

EHR includes patient functional status in coded form, which is needed in decision steps of clinical guidelines. Current clinical healthcare follow-up systems cannot employ available EHR data in their process flows due to the interoperability problems with legacy EHR systems.

The iCARDEA platform provides EHR interoperability so that information about patients' medical history such as history of non-cardiac conditions; more detailed information about severity of each condition (e.g., record of prior hospitalizations or specifics of therapy for the condition); the medications being taken at the time of spontaneous arrhythmia occurrence or the non-cardiac conditions denoting contraindications to the proposed therapies can be obtained from the patient EHR data and used in the clinical workflow. One of the major challenges to be addressed related with EHR interoperability is the interoperability of the code system used (semantic interoperability). For this purpose iCARDEA Platform provides a Code Mapping API.

While interoperability between all these EHR and PHR systems are provided, security and privacy mechanisms are needed for controlling access to data on these PHR and EHRs. One of the privacy and security challenges in iCARDEA system is the possibility of the access of

patient related information is being accessed from unauthorized users on purpose or improbable way. For example, finding out that a political person having a disease or reaching medical history of a person that does not want her medical history is discovered would be unfavourable in both legal and ethical conditions.

Therefore the iCARDEA Consent Editor, in detail described in deliverable 5.4.1, is used to provide a patient centric access control mechanism for PHR users. iCARDEA Consent Editor is a compatible tool and can be easily integrated to different PHR systems through its PHR interface. Consent Editor produces XACML documents to accommodate the standards and uses those documents while generating access decision, hence PHR's are also able to use Consent Editor's decision making service just by sharing the restriction policies that created in XACML standard, without being completely integrated with Consent Editor.

iCARDEA Consent Manager is also integrated to the Care Management Database System. The Care Management Database (Caremanagement DB) is the central repository of the Care Plan Engine, the Patient Parameter Monitor and Data Analysis and Correlation Tool are using this data source. The Care Management Database holds the unified information from different data sources of iCARDEA system such as PHR, EHR and CIED Data Exposure System. The Care Plan Engine, the Patient Parameter Monitoring Tool and the Data Analysis and Correlation Tool accesses patient related data from this repository. Consent Manager is integrated to the Care Management Database system in order to ensure that the request of accessing patient related data stored in this repository should be authorized according to the consent of the patient in question. Since the part of the medical knowledge bases accessible via iCARDEA contains only anonymized patient pattern, which can not be related to an individual patient, these data access is not controlled directly via the Consent Editor but through protection via the database management system and the DACT Tool using access rights of the medical professional. Access to the medical knowledge base for creating the patterns is especially restricted to the data analysts due to legal requirements and the fact that the DACT Knowledgebase is stored outside the iCARDEA environment.

2.2 ICARDEA DATA ANALYSIS SYSTEM ARCHITECTURE

Since the iCARDEA system architecture does not represent the system of data analysis, which is outside the iCARDEA system as stated in Deliverable D3.4.1, Figure 3 shows the complete architecture together with the data flows representing the process steps. Before discussing them in more detail in the following sections, here a brief summary is presented.

The origin of the data analysis architecture is on the left, with the different potential useful datasets of historical medical information about treatments. These are integrated at OFFIS into a first database without any manipulations done to the data, except transferring it into a relational database format. Now the data can be evaluated for its structure quality and needed metadata, which is also integrated into this database. The metadata for iCARDEA consists of the used classifications, mappings of non-standardized description to suitable representation and also pre-definitions for the OnLine Analytical dimensions. Based on this, the data is harmonized, quality assured, suitable partitioned and then integrated into the Data Warehouse (DWH). The DWH consists of all available data in a cleaned and for the single source comparable format. These steps are, as stated in D3.4.1, most time consuming. Based on this data, special data representations (So called cubes for OLAP) or special relations for data mining (propositionalized relations consisting only of Boolean and Integer values) are created

for every data analysis task. The data mining tasks has to be done by an expert several times to optimize the parameters of the algorithms to obtain potential meaningful results. The results, either from OLAP or Data Mining, are preprocessed for discussion with the healthcare actors. After the discussion, either steps in the data analysis are reconfigured or the found patterns are integrated into the Pattern Database at the iCARDEA environment at SALK.

There the patterns are used for providing suggestions to the healthcare actors with respect to the parameters of the current treated patient. These parameters are taken from the Caremanagement DB, provided and enabled by iCARDEA data exchange via standards.

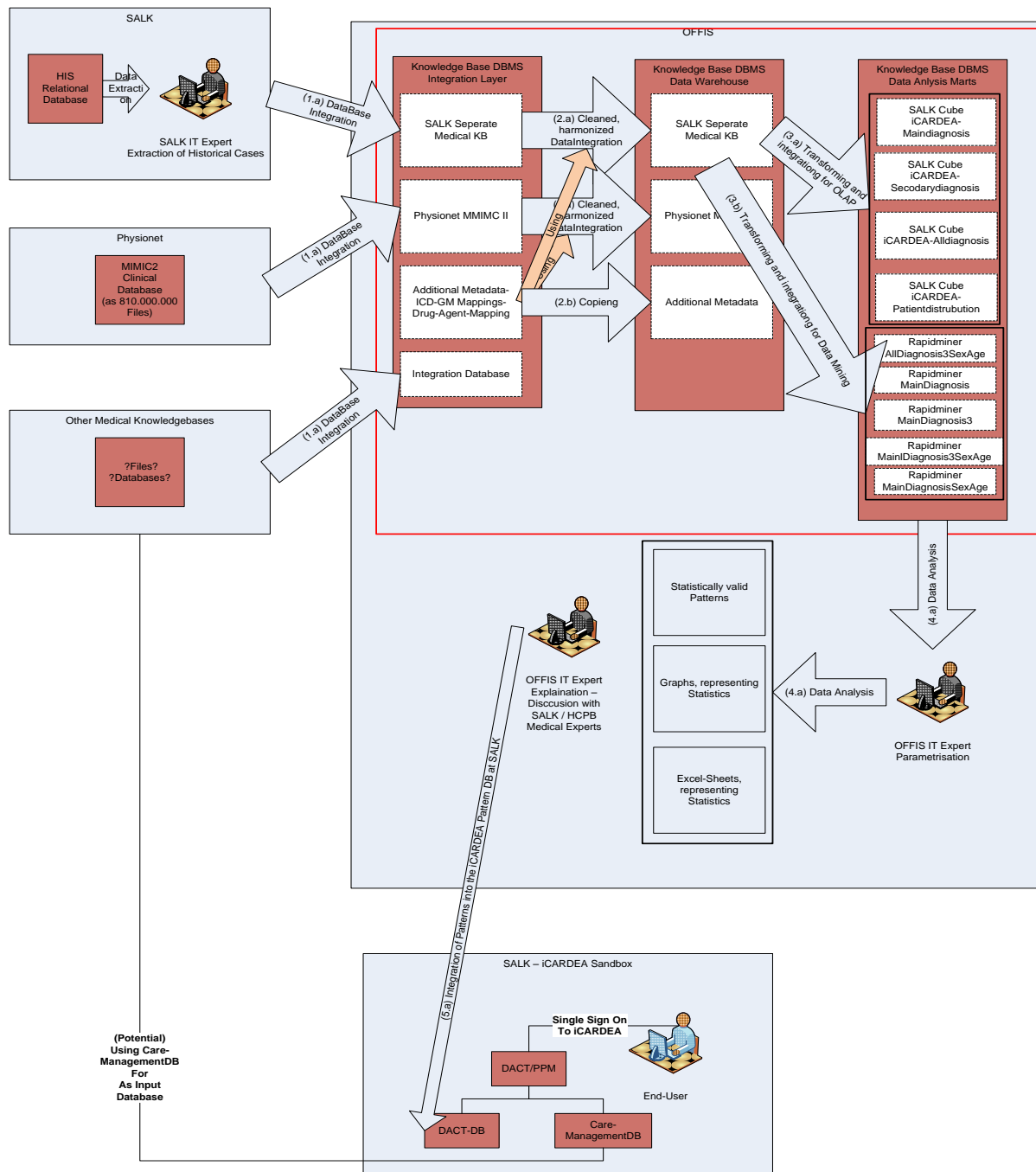


Figure 3 Process of Data Analysis with its architecture

Since the Caremanagement DB collects all patient data from the target scenario, it could serve as a historical medical case database for data analysis in the future. Since it collects data soonest at the evaluation, this was no option for the data analysis task. But the collected data would be very useful for data analysis for answering the specific medical questions, since it should fulfil all the requirements stated at section 3.2.

3 Building the Data Analysis Process

The following section shows in a chronological order the steps performed in the iCARDEA project for data analysis. They are inline with the model presented in section 2.

3.1 IDENTIFICATION OF USER NEEDS

The first part of useful data analysis is the part of business understanding. This especially includes the needs of the users. The users in iCARDEA were identified as mainly the clinicians at the evaluation clinic at SALK. Therefore in September 2010 a small workshop was hold at SALK to discuss the intended analysis results with the medical staff. The kind of expected analysis was related with better understanding for the patients. Therefore questions were stated like: “How many patients had (none) arrhythmias under (no) medications”. This was the general question for questions like “how many patients had atrial fibrillation while taking anticoagulation (blood thinner)”. This is a question which has to be answered by OLAP, since statistics are wanted to address these pre existing questions. This means a hypothesis from the healthcare actors should be verified.

The second identified “user” was the demonstration potential of iCARDEA. Since the data analysis should provide suggestions for the potential “development” of a patient in the future, more predictive questions had to be answered. These were not stated by the clinicians as requirement, but assumed form the intended iCARDEA functionality from the task description. This question could be described as, what combinations of parameters are most likely for the patients. The answers can then be used as “recommender system”. This means hypothesis are created from the existing data sets.

3.2 IDENTIFICATION OF AVAILABLE DATA SOURCES

To answer the questions by the users and to offer more potentially interesting questions, available data sources had to be identified.

The requirements for data to be fully use- and meaningful for data analysis are as follows:

- Reliable data:
 - The data must be from a trustful environment and be based on real cases, not artificial ones.
- Approved data:
 - The conclusions and actions in the knowledge base must be approved by experts in CIED. Random data or data without an indicator for the success of a treatment will lead to inappropriate conclusions.
- Available data:
 - The data must be available to the project use. It is important that no legal regulations prohibits the access to the data or the use of the extracted patterns
 - Useful data:

- The data must have the potential to hold useful patterns. A data set without conclusions or proper actions in the area of ICD / CIED handling will not be suitable for the purpose of the project.
- Information technology
 - Structured data: The data must be in the form of attribute–value pair representation.
 - Facts: The data has to consist of facts. Ambiguous colloquial descriptions are not interpretable for computers.
 - Every attribute has to be from a specific type of data with a defined value range.
 - The attributes have to be available also at the target environment to use the pattern.

An additional requirement at iCARDEA is that the data has to be comparable to the iCARDEA environment. A diagnosis from Spain can not be compared automatically to a diagnosis from Austria, since there could be different instructions / underlying conditions at the healthcare systems to code them. This leads to biases or complete different interpretations. Since the evaluation of iCARDEA should be done at an Austrian Clinic, there would be coded diagnoses in DIMDI ICD10-GM (German Modification) Version, which is also enhanced by some Austrian specific Codes. This makes data from other countries than Austria, Germany and Swiss incompatible. Even more, as in Germany and in Switzerland are different Healthcare Systems compared to Austria, not even these are without a bias. To use the patterns for meaningful suggestions at SALK, the data should be comparable.

3.2.1 Research on existing datasets

Since the data integration is a big time consuming block, especially from sources which are not prepared for data analysis, a research was done on online available medical cases but also directly through discussions at information and medical fairs and conferences.

A quick research was done on the following online available sets at the beginning of the iCARDEA project. The goal was to prevent to do a time consuming data extraction of an individual source:

- <http://mlr.cs.umass.edu/ml/datasets.html> The UCI Machine Learning Repository with 189 different data sets. Following data sets were identified as potentially useful for iCARDEA :
 - Arrhythmia Data Set: Distinguish between the presence and absence of cardiac arrhythmia and classify it in one of the 16 groups.
 - Echocardiogram Data Set: Data for classifying if patients will survive for at least one year after a heart attack. *There is no information about surgeries or implants.*
 - Heart Disease Data Set: Dataset with 4 databases and 76 attributes to predict the presence of a heart disease.
 - Post-Operative Patient Data Set: The classification task of this database is to determine where patients in a postoperative recovery area should be sent to next. Because hypothermia is a significant concern after surgery (Woolery, L. et. al. 1991), the attributes correspond roughly to body temperature measurements.
 - SPECT Heart Data Set: The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the

patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images.

- SPECTF Heart Data Set: The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images.
- Statlog (Heart) Data Set: This dataset is a heart disease database similar to a database already present in the repository (Heart Disease databases)
- <http://www.cardiosignal.org>
 - Two datasets are available:
 - Liver Specific Gene Promoters: Liver specific genes the provider of the data sets referred to are either expressed exclusively in liver or in a small number of tissues including liver. The liver promoters of genes in this dataset are curated manually. A comprehensive literature review is performed in order to determine the pattern of expression of the corresponding genes. This strategy allowed identification of 47 non-orthologous genes preferentially expressed in liver.
 - Muscle-specific regulatory regions: Here, the provider of the data sets further mapped the each region in the corresponding UCSC genome assemble of the current version. These are the 200 bp sequences which were used as the positive portion of the training set for the logistic regression analysis.
- <http://www.ncbi.nlm.nih.gov/pubmed/>
 - PubMed, mentioned in the Description of Work, has no electronically datasets available. It is a knowledge base for medical papers and not for clinical trials / use cases.
 - <http://www.ncbi.nlm.nih.gov/sites/GDSbrowser/> has datasets about gene expression.
 - <http://www.physionet.org/> : PhysioNET mentioned in the Description of Work has a lot of datasets available including ECG Data with annotations. IT cons
- <http://disco.neuinfo.org/webportal/discoResourceListTable.do>
 - A list of databases, but no links to CIED / ICD / Cardio datasets except PhysioNet available.
- <http://www.meddb.info/>
 - Finder for medical and molecular biological databases. No further interesting database (cardiosignal and Physionet is listed)
- http://nmtracking.unm.edu/resources/web_links.html
 - Building a Healthy New Mexico: Datasets about health statistics. But no one for cardiology or CIED/ICD

None of these datasets matches the requirements stated at the beginning of this section and didn't seem promising for generating useful patterns for iCARDEA or could possibly answer the questions stated from the clinical professionals at SALK. The problem for all of the stated data bases was that none of them was trusted by the medical partners. They seemed not reliable to them to produce patterns, they would trust. From their point of view the data was non-reliable and non-approved. Also from the technical point of intended data analysis, most of the dataset were either artificial, working on genome level or about identifying heart

diseases by interpreting ECGs or other waveforms. This means that the data provided by the data sets would not be available at the environment at SALK to create suggestions.

3.2.2 SALK

Since the research on preexisting databases wasn't successful, it was identified, what data is available at the evaluation clinic itself. These data by nature fulfills the stated substantial requirements. The investigation had to be done on the technical requirements. Therefore the IT-department was interviewed and the computer forms, used for clinical documentation, were examined. Based on this, following items were identified as potential available:

- All personal data about the patient and his family, like birthdate, civil status, gender
- Data about the implemented devices and the dates of the implantation
- All diagnosis of a patient at the different stays at the clinic together with the treating department and major or secondary diagnosis
- Prescriptions at discharge
- Lab results
- Accounting codes for the healthcare system

This was based on analysis of the forms you can see in anonymized way in the appendix 6.1. After these were identified, the legal regulations had to be clarified for a data exchange. This took several months and is described in detail in D7.3.1.

After it was clarified that and under what circumstances the data could be used with respect to legal issues, the IT department at SALK extracted the requested data.

The requested data couldn't be provided in total due to technical and legal limitations. But data about the medication, the devices, the patients, diagnosis together with dates were provided.

3.2.3 Physionet

As the initial analysis showed, no readily available database was found suiting the technical requirements of iCARDEA and the healthcare actor's needs at Austria. For this reason, we had extracted the historical cases of previous CIED patients from SALK and performed initial data analysis on this data set. As a result of the first iCARDEA Review meeting, the reviewers requested to extend the datasets used by the data analysis task, by making use of existing knowledge bases. For this reason, a survey with lowered requirements was done, and MIMIC II database provided by Physionet located at the United States of America was identified as the most suitable one. Although this dataset is not specialized to ICD or CIED patients, this database seemed to be potentially useful to extract patterns for cardiac patients.

This assumption was based on the descriptions shown in appendix 6.3.

The databases consist of 28.000 patients of intensive care units at the United States, treated between the years 1990 to 2000. The description of the datasets showed, that there are a lot of patient information (also illegal in Europe for data analysis like religion and race) together with medications, diagnosis, chartevents, lab results and a lot more. Since this seemed promising to obtain meaningful results, it was decided to use MIMIC II as second database for generating patterns.

3.3 INTEGRATION OF DATA SOURCES AND DATA PREPERATION

After identifying the available potential data sources for data analysis, the data had to be integrated into two databases as shown in Figure 3. This is first the integration layer and the

second the data warehouse. The first step consists of integrating the different formats like textfiles, excel-sheets, databases into a relational database. Then the data has to be cleaned, that means correction of syntactic or semantic errors, dealing with missing values, or getting it into one comparable format. It also means that the data should be coded with comparable or even better the same coding schemes.

3.3.1 Problem of different and changing codings

The problem of different and changing codings arises at data analysis task, since data analysis is intended to make statements over the meaning of data. As already mentioned in section 3.2, the meaning of data elements should be comparable. Therefore in data analysis metadata is used to describe the meaning of data elements. In the medical domain of Germany and Austria for diagnosis the *International Statistical Classification of Diseases and Related Health Problems (ICD)* is commonly used in a German Modification (ICD-GM) of the WHO ICD version. The current WHO Version is 10 (ICD10). This metadata provides a taxonomy with predefined semantics for diagnosis and is intended to be used for documentation and also statistic purposes.

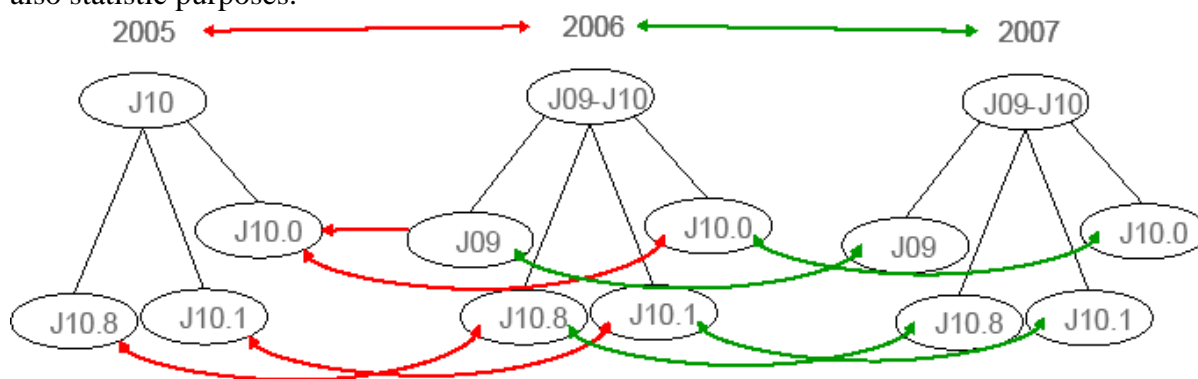


Figure 4 Subgraph of ICD-GM metadata for influenza viruses and its development for 2005 to 2007

But since the data used for data analysis normally span over a longer time period, the metadata evolves. The ICD-GM, is updated every year together with transition rules between the old and new classifications [ICD2011]. Also, if a new disease is identified, a new code is introduced. For example, the bird flu was identified as a disease and introduced in 2006 with a new code *J09*. Figure 4 shows the subgraph for the years 2005 to 2007 of the coding for influenza viruses. The graph represents the official taxonomy of the ICD-GM codes and the directed edges represent the officially provided transformation rules for conversion from one years version to the following [ICD2005,ICD2006,ICD2007]. The vertical lines show the connection to the higher level parent node.

Data to be analyzed will be stored in a data warehouse on the finest level, i.e. the data is stored according to the values at the leaves. If an analysis about the values *J10.8* or *J10.1* is required there is no change at the data because of the existence of bijective edges to the same nodes in every year.

The problem of semantic shift occurs if an analysis for the years 2005 to 2007 of *J10.0* should be done, because it is ambiguous which data to use for the following reason: If *J10.0* from the year 2005 is intended, the code *J09* has also to be considered for 2006 and 2007. If the concept *J10.0* of 2006 or 2007 is intended, the human analyst must be aware that there was a change in the meaning of the data, even when the data is syntactically identical and a transition in both directions exist.

The domain specific background knowledge needed for analysis is that *J10.0* is a collecting node for not yet known types of influenza. As mentioned before, in 2006 the bird flu was identified and introduced as new code. Thus, the meaning of *J10.0* as all unknown viruses is unchanged but compared to 2005 it is *without bird flu*. This would lead to inaccurate results if not considered in statistical analysis on such fine-grained data. For the analysis on the higher level parent node *J10* in 2005 or *J09-J10* for 2006 and 2007, the results would be accurate because all existing transformation edges are only referring to child nodes. A broader description with respect to the embedding for data analysis can be found in [Luepkes2011]

To deal with this problem, in iCARDEA a first prototype was developed to show the experts the evolutions, that can occur and also, if the code is available. In Figure 5 two complete different meanings of the ICD-GM code C83.3 (type of cancer) are shown. For the years of 2004 to 2010 C83.3 was stable, but in 2011 it changed to C85.5. If the meaning of C83.3 in 2011 was interesting, for the years 2004 to 2010 the code C83.4 has to be taken as reference value.

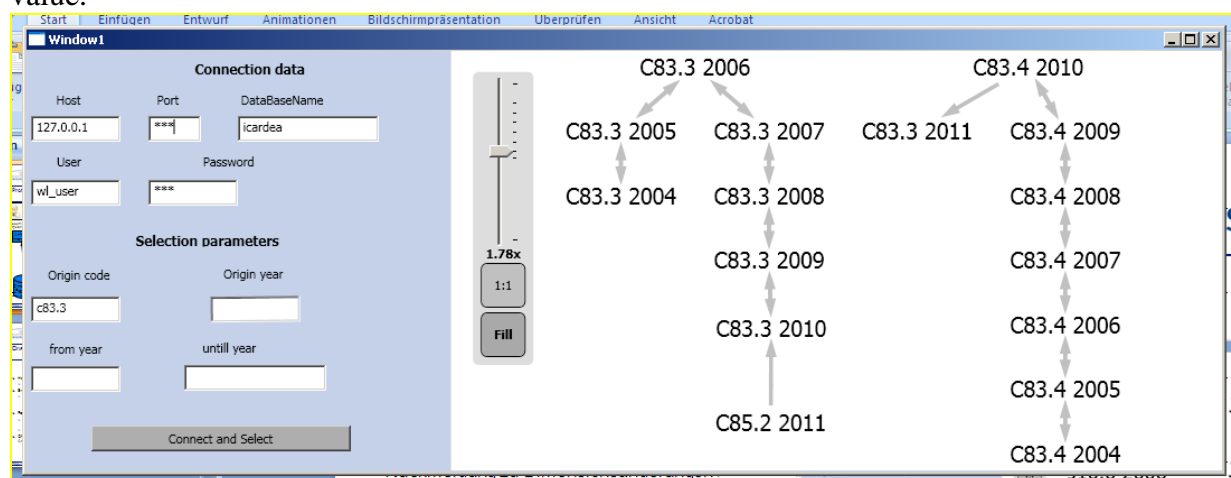


Figure 5 Prototype showing different meanings of ICD-GM code C83.3

3.3.2 SALK

The SALK datasets were extracted and provided as encrypted excel files like stated in Deliverable 7.3.1. For these excel files first relations had to be created and the excel files were integrated into them using standard tools. The data provided consisted of 228 ICD / CIED patients with their gender and birth data.

Also diagnosis information were provided. These consisted of ICD10-GM Codes together with the year and month of the treatment, the department where the diagnosis was done and two different attributes describing if the diagnosis are major or secondary nature and if the diagnosis was done at clinical / department admission, discharge, follow-up or treatment. In total for 199 patients sets of diagnosis were available. In Figure 14 you can see the form with all the information about diagnosis of a patient.

Also medication information were provided. These consisted of the intake time (Morning, Noon, evening, Night) with amount and the name of the brand together with the units. Since the brand names are not comparable a mapping was done: For 58 different drug names are mapped to their agents and doses. Only for eight of ICD/CIED patients the medications were provided. Investigation at SALK showed that no substance administration information were electronically available for more ICD / CIED patients. Also no timestamp was provided for

the substance administrations of these patients. These drawbacks resulted in the decision not to use the medication information.

Another provided input was information about the implants. The electronic form presented in Figure 12 shows structured information about the manufacture of the aggregate and the potential three leads divided in right, left and atrium location together with the type (active and passive). Unexpectedly the electronic data provided by SALK was not in the structured format of the forms. There was information about the device, but not in electronically interpretable way. Most consisted of free text of the special device names without even containing the vendor. Also the date of the device implantation could not be extracted from the SALK-Systems.

To improve the data amount and quality a second data provision round was initiated. Complications of the ICD and CIED patient were provided and also the information about the implants was extended. But this data was like the previous not in an unstructured format and without useful information for data analysis.

All the mentioned data was imported into the integration layer and quality assured. This means for the diagnosis information provided as ICD-GM Codes starting 2007 versions, all 7000 provided items were checked for correctness and then integrated using the tool presented in section 3.3.1 to the version of 2010 to have a comparable data source and especially semantic. The diagnosis was separated into different types of diagnosis; namely all, major and secondary diagnosis.

The medication of the patients was harmonized using the declared medication mapping. This was done before it was clear, that no further medications of patients will be available and therefore not used.

Especially for every data item the age of the patient was calculated and assigned to the data item itself to indicate, when something happened.

3.3.3 Physionet

Since the chosen MIMIC II Clinical Database is provided for research, the data integration process was completely different from the task of SALK data integration. Physionet provided already descriptions of the available relations and the intended content of those. You can find them with the remarks from the investigation in the Appendix 6.3. Since the descriptions seemed promising for iCARDEA data analysis, it was thought that it can be used to create patterns based on patients from the USA in the 1990s. The plan was to adapt these patterns using the mapping tool for patients at Austria. Yet we were aware of the risk that the end users at SALK Clinic may not trust this data and patterns.

The integration of MIMIC II was a time consuming task. Since the datasets were provided for each patient and each datasets, it in total consisted of about 80.000.000 files with 80 Gigabyte of data volume. This led to system instability of the windows operating system running the PostgreSQL database, in which it should be integrated. Since for iCARDEA the waveform data were not so interesting at pre-analysis, these were not in totally integrated to reduce the amount of data. All other data was integrated to be used. You can see the count of the itemsets per relation at Figure 18 and Figure 19 in the Appendix.

After integration was complete, a research was done on the used codings, classifications, acronyms and kind of input of the data, since the North American Health Documentation is not similar to the German or Austrian one.

3.3.3.1 Decision to drop Physionet

The pre-analysis of the integrated Physionet MIMIC II Clinical database was very promising. With 27903 different patients together with 270.000 diagnosis codes, background including abuses and chart events seemed like a good repository candidate to produce patterns on a broader patient basement. This was also one of the requirements highlighted by our reviewers in first annual review of iCARDEA Project. The drawback came at the finer analysis of the integrated patient data. After the integration, it was possible to search the patient's clinical history. For iCARDEA usage, the patients for data analysis were intended to have cardiovascular problems and / or an implanted device. To identify these patients, it was searched for keywords indicating such patients. At first such an analysis was done by the data analysts, without deeper medical knowledge, finding about 2000 potential patients. Therefore a keyword list was produced using strings like 'ventricular', 'VT' and other strings

These keyword lists were updated by the clinical partners, which knew the domain. You'll find the keywords used for identifying relevant patients from Physionet at section 6.3.4, Figure 20 at the Appendix. It turned out that a search on the patients with SVT / VT tags returned only one patient. This amount was the same when searching on the patients with ICD tags. All other patients had with high likelihood no chance to be similar to cardiac patients of iCARDEA.

Since the amount of potentially interesting patients was dramatically lower then the already existing amount of SALK patients, it was decided to spend no further efforts in preparing the data for analysis and to obtain or provide the required metadata.

3.4 BUILDING DATA ANALYSIS

After the data was cleaned and the metadata for understanding and using was integrated, the core data analysis could be prepared. Since the healthcare actors wanted statistics about their patients, OLAP cubes were provided. For generating more predictive patterns data mining was also performed on the prepared data.

3.4.1 OLAP on SALK

For creating OLAP cubes on the SALK data first so called dimensions had to be defined. These dimensions are taxonomies describing the semantic of the data and providing paths for analysis.

For the SALK data, an age dimension was produced for structuring the age of the patients. This was binning the patient ages as year in groups of five years, a group of unknown age, and one special group holding people older then 84. All of these groups are connected to an "All" element representing the totality of patients.

The next dimension was created for sex. This consists only of male, female and other gender and the "All" element, representing the totality of patients.

The hugest dimension was created for the ICD10-GM 2010. This consisted of all 16000 ICD-codes, organized and structured as provided by the DIMDI (translation "German Institute of Medical Documentation and Information").

Using these dimensions following analysis cubes were build:

- iCARDEA-Patientdistribution consisting of
 - Groups of Age
 - Sex
- iCARDEA-AllDiagnosis consisting of
 - Groups of Age
 - Sex
 - ICD10Codes → All provided diagnosis to the patients
- iCARDEA-MajorDiagnosis consisting of
 - Groups of Age
 - Sex
 - ICD10Codes → Only provided major diagnosis to the patients
- iCARDEA-Patientdistribution consisting of
 - Groups of Age
 - Sex
 - ICD10Codes → Only provided secondary diagnosis to the patients

After creating the cubes, software was written to transfer the pre-cleaned data into a special OLAP Server provided by Palo⁸. This software was developed in C#.

3.4.1.1 Results

The results were provided to the medical experts for interpretation. To have some information about the patient dataset here are the PatientDistributions as table and as chart.

The age represents the first appearance of a patient at the cardiac clinic. Notable is that it seems that the amount of women is constant over their age. But men have really a peak at the age between 60 and 75.

		This tables shows the distribution with respect of sex and age of first stay at the clinic of the patients at the SALK Database.											
iCARDEA - Patientdistribution													
Absolut Amount	All(Age 85 Plus)	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-x
All (Sex)	198	2	7	6	12	18	17	34	45	24	23	8	2
Female	35	0	3	2	2	4	0	6	6	6	4	2	0

⁸ <http://www.jedox.com/de/enterprise-spreadsheet-server/excel-olap-server/palo-server.html>

Male	163	2	4	4	10	14	17	28	39	18	19	6	2
Percentage per Age	All(Age 85 Plus)	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-x
Female	18%	0%	43%	33%	17%	22%	0%	18%	13%	25%	17%	25%	0%
Male	82%	100%	57%	67%	83%	78%	100%	82%	87%	75%	83%	75%	100%

Table 2: Results of OLAP analysis – Cube iCARDEA-PatientDistribution

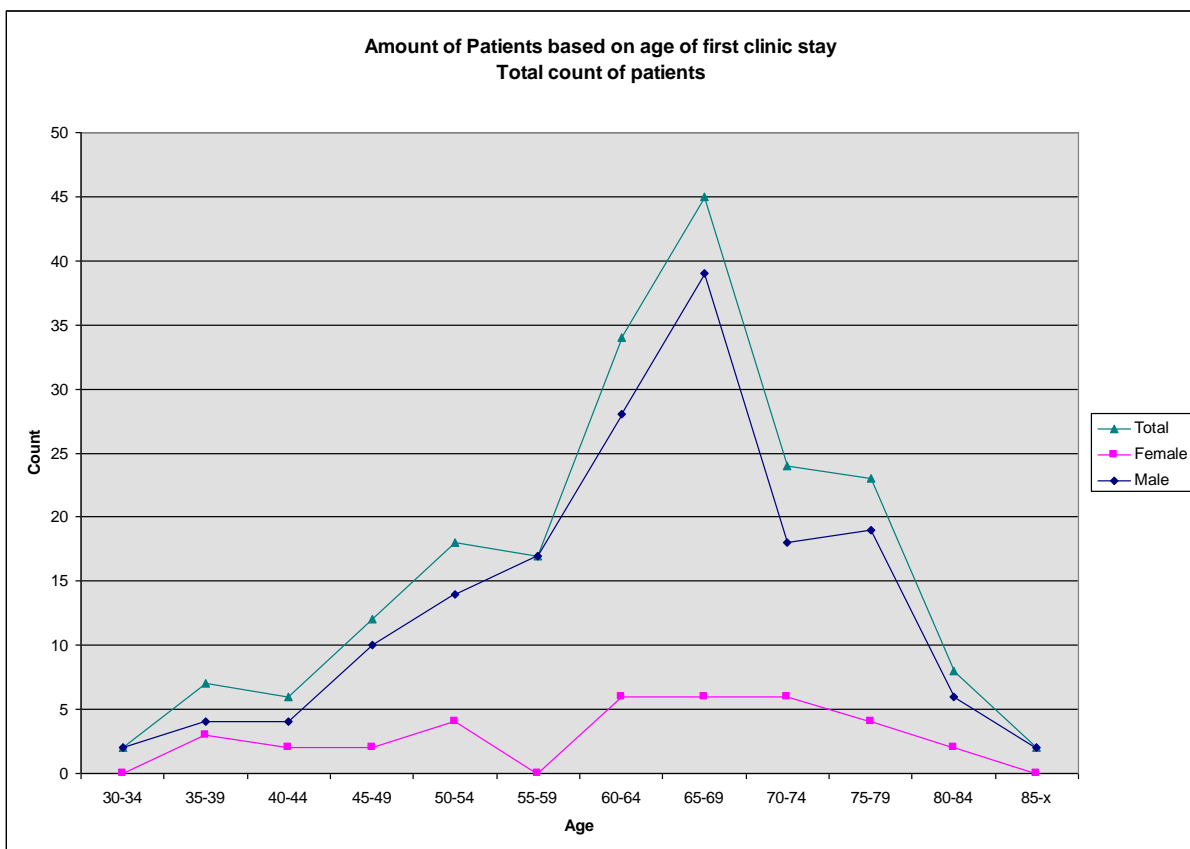


Figure 6 Graphical representation of Table 2: Results of OLAP analysis – Cube iCARDEA-PatientDistribution

Here are also some results from the iCARDEA-MajorDiagnosis part.

There are some underlined percentages. These are fields where for a higher amount of patients a derivation from the suspected distribution of diagnosis between male and female is spotted. This should illustrate the interpretation of OLAP results.

	All(Age85Plus)				
	All (Sex)	Female	Male	% Female	% Male

All (Icd)	768	162	606	21%	79%
I	5	0	5	0%	100%
II	38	25	13	66%	34%
III	6	2	4	33%	67%
IV	17	3	14	18%	82%
V	40	6	34	15%	85%
VI	14	2	12	14%	86%
VII	15	7	8	47%	53%
IX	394	68	326	17%	83%
I05-I09	1	0	1	0%	100%
I10-I15	4	1	3	25%	75%
I20-I25	67	9	58	13%	87%
I26-I28	3	1	2	33%	67%
I30-I52	280	50	230	18%	82%
I60-I69	8	1	7	13%	88%
I70-I79	25	4	21	16%	84%
I80-I89	5	2	3	40%	60%
I95-I99	1	0	1	0%	100%
X	14	0	14	0%	100%
XI	28	3	25	11%	89%
XII	6	0	6	0%	100%
XIII	31	5	26	16%	84%
XIV	26	11	15	42%	58%
XVII	2	1	1	50%	50%
XVIII	23	7	16	30%	70%
XIX	100	19	81	19%	81%
S00-S09	3	0	3	0%	100%
S20-S29	1	0	1	0%	100%
S30-S39	1	1	0	100%	0%
S40-S49	2	0	2	0%	100%
S70-S79	1	1	0	100%	0%
S90-S99	2	0	2	0%	100%
T00-T07	1	1	0	100%	0%
T08-T14	2	0	2	0%	100%
T66-T78	3	0	3	0%	100%
T79-T79	2	0	2	0%	100%
T80-T88	78	14	64	18%	82%
T90-T98	4	2	2	50%	50%
XXI	9	3	6	33%	67%

This table shows in the first column the ICD10 GM codes as its groups or even chapters for the patients presented previously. The next column presents the absolute amount of diagnosis, with no respect to the gender. The next two columns present the total amount of diagnosis separated by gender. The last columns show the distributions of the diagnosis as comparable percentages.

The first row shows that the overall distribution is that 21% of the diagnosis belongs to the female patients and 79% to the male patients. This would be expected to be the normal distribution for this dataset. As you can see in the row starting with “II” there are the female three times often than the male patients. The ICD-Chapter II stands for “Neoplasms”. But this is nearly the only amazing trend. This table shows that the diagnosis is nearly independent from the gender, if there are enough diagnosis/patients involved. The second statically interesting information is XIV “Diseases of the genitourinary system”.

3.4.2 Data Mining on SALK

For identifying more sophisticated patterns, also several attempts with different data mining tools were made. For iCARDEA purposes rapidminer⁹ was chosen.

In rapidminer an association approach was followed. This means identifying sets of patient data that occur often together. To improve the results, the association analysis was enhanced by a binning algorithm to create the best age ranges for associations. As input for this processes the data had to be proposionalised and also the ICD10GM-attributes had to be transferred to a Boolean matrix. The trickiest task in data mining was to tune the parameters of the age binning algorithms and especially the apriori algorithm to produce association rules.

Associations rules are of the form “*antecedent* → *consequent*” or more natural “prerequisite → conclusion”. They consist of items (in our case ICD10 Codes, age-bins, gender). The idea is, to identify first itemsets, that often occur together and then to identify, which item is most likely a derivation of the other items. To show how relevant a rule is, a confidence is given, representing the percentage, how often the rule is correct for patients with the prerequisite. As an example, the ICD10 Code for birth should lead to the gender ‘female’ with a confidence of 100%. But the gender ‘female’ as prerequisite will lead to ‘birth’ with less confidence, since not every woman at the clinic is treated for birth. But for every birth, the patient is female.

For the iCARDEA data following parameters for creating the rules produced the most suitable results:

- The ages should be divided into 5 bins
- Confidence: The found rules had to have at least 30 % of confidence. This means that the rule is true for at least 30% of the datasets fulfilling the prerequisite.
- Delta of 0.05 to lower minimum support to find frequent itemsets. Minimum support means, how many datasets must be available with the used itemsets.
- ICD10 codes should not be to specialized, only 3 digits to produce rules with more then 50% confidence

3.4.2.1 Results

These are the results, only the ones with more than 50% confidence are shown, since the following ones are only permutations of these frequent item sets.

The binning produced following borders for the attribute age:

Range-Name	Start and End age	Count of Patients
range1	[-∞ - 53.500]	40
range2	[53.500 - 62.500]	45

⁹ <http://rapid-i.com/content/view/181/196/>

range3	[62.500 - 67.500]	38
range4	[67.500 - 73.500]	36
range5	[73.500 - ∞]	40

These are the association rules. The prerequisite is mostly, the conclusion is always an ICD10 Code.

Prerequisite	Count of Support		Conclusion	Count of Support	Confidence
E78 and I10	31	→	I25	26	84
I10 and I47	34	→	I25	28	82
E11	25	→	I10	20	80
E78	44	→	I10	26	76
I10 and I50	30	→	I25	23	77
E78 and I25	34	→	I10	26	76
E78	44	→	I10	31	70
I10	64	→	I25	45	70
I25 and I50	35	→	I10	23	66
I10 and I25	45	→	I47	28	62
I50	57	→	I25	35	61
I25 and I47	47	→	I10	28	60
E78	44	→	I25	26	59
I10 and I25	45	→	E78	26	58
I47	85	→	I25	47	55
Age = range3	38	→	I25	21	55

From a technical point, it is remarkable, that the attribute sex is not included in any of the created top confidence 50 rules. Also the nearly non presence of age was not to be expected.

Although remarkable is the low diversity of used ICD10 Codes. From the total available 253 used diagnosis on the group level, only the six groups E11, E78, I10, I25 I47, and I50 appears together. All other 247 ICD groups are not significantly used with other codes. Since I10 till I50 are cardiac diseases this would be expected. E11 and E78 - diabetes and hyperlipidaemia – seems to be indicating the most two common spotted risk diagnosis.

4 Design & Implementation

The design and implementation of software related with data analysis and correlation can be divided into two parts, the end user tool for presenting and correlating patterns to the medical professionals and the software components needed for the successful generation of patterns.

4.1 DACT USER INTERFACE

The DACT user interface is provided to the end users at the hospital as subsystem of PPM. This decision was made, since PPM is intended to be the central point of iCARDEA, where the healthcare actor can access all data available for the treated iCARDEA patient. Due to

intended user friendliness, this decision was made. While the healthcare actor is treating the patient and is looking at the patient data, he can also have a look at statistically valid patterns that could be relevant for the patient.

Therefore the values of the current treated patient are compared to the prerequisites of patterns stored at the special pattern database. If the patient matches them, they will be shown.

iCARDEA_PPM_DACT					
Prerequisite	Conclusion	Confidence	Support	Approved	
ICD10= E78 and ICD10= I10	ICD10 = I25	84	26 of 31	Approved	
ICD10= I10 and ICD10= I47	ICD10= I25	82	28 of 34	Approved	
ICD10= E11	ICD10= I10	80	20 of 25	Approved	
ICD10= E78	ICD10= I25	77	34 of 44	Approved	
ICD10= I10 and ICD10= I50	ICD10= II25	77	23 of 30	Approved	
ICD10= E78 and ICD10= I25	ICD10= I10	76	26 of 34	Approved	
ICD10= E78	ICD10= I10	70	31 of 44	Approved	
ICD10= I10	ICD10= I25	70	45 of 64	Approved	
ICD10= I25 and ICD10= I50	ICD10= I10	66	23 of 35	Approved	
ICD10= I10 and ICD10= I25	ICD10= I47	62	28 of 45	Approved	
AGE= 62.500 - 67.500	ICD10= I10	55	21 of 38	Open	

Figure 7 Overview of patterns

In addition a tab is presented, where the healthcare professional can access and view all patterns stored at the parameter database.

To ensure that only trusted persons can access patient data, PPM and DACT uses single sign on provided by iCARDEA Platform and Consent editor to ensure, that only people with sufficient rights view this data. The idea of consent management was briefly provided at section 2 and can be found in detail in deliverable D5.4.1 – Patient Consent Management and Security. A more detailed description is given in D7.3.1.

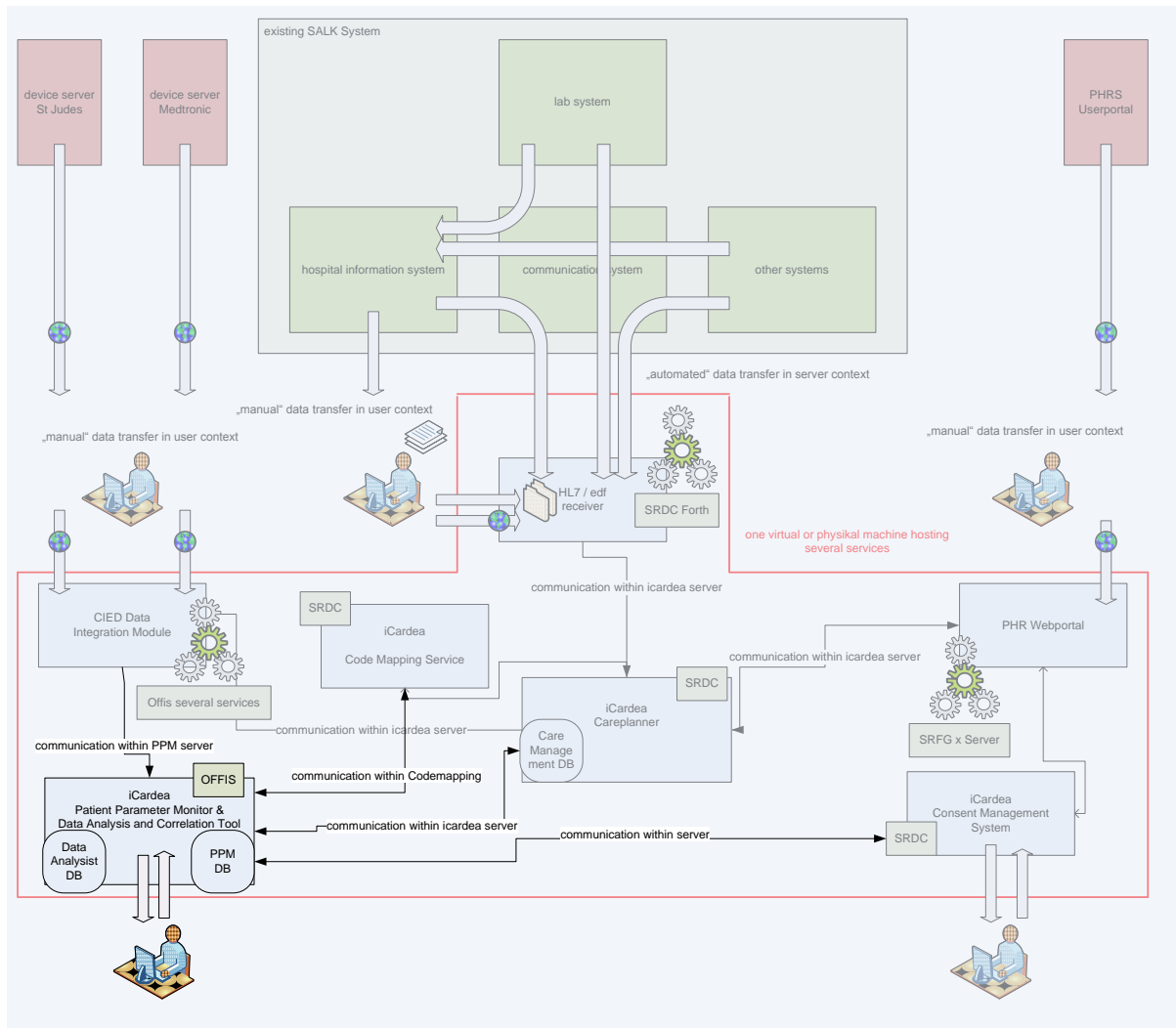


Figure 8 iCARDEA System Architecture and the Role of the end user DACT in this Architecture

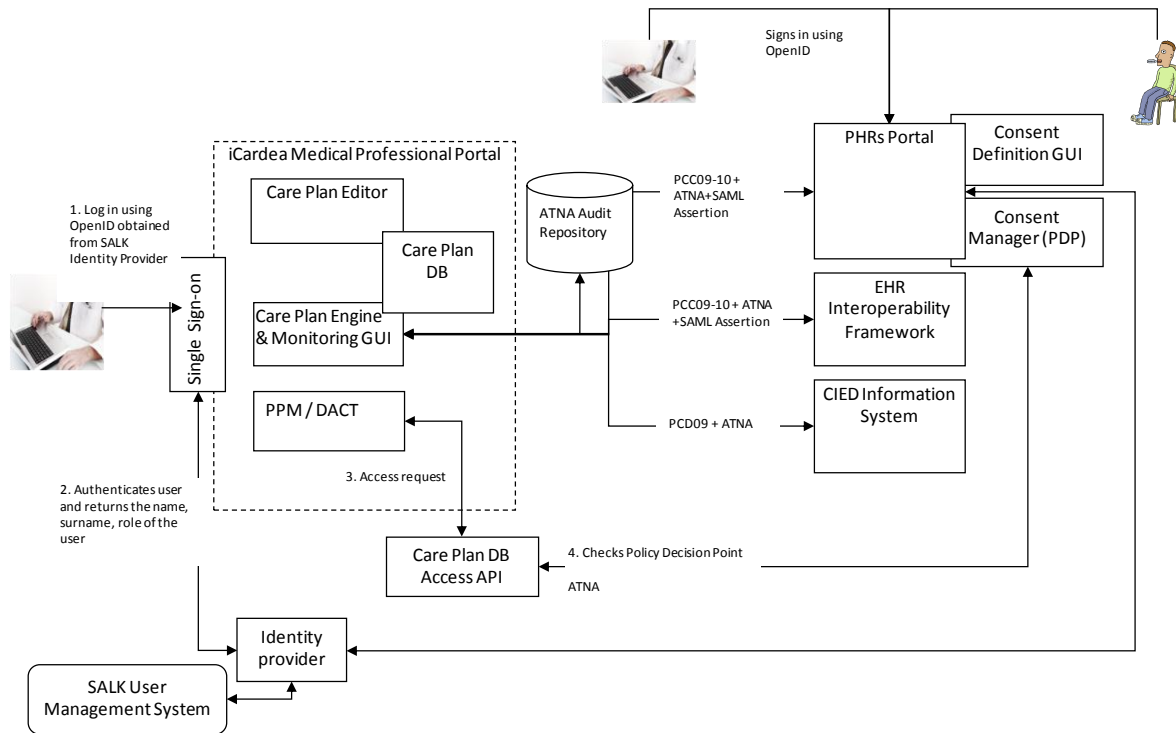


Figure 9 Single Sign On infrastructure

4.2 SOFTWARE NEEDED FOR PATTERN GENERATION

For the processes of pattern generation different software tools were developed. As already presented in section 3.3 there are different needs for software. For integrating the source data from Excel-Files and especially for the integration into the data warehouse and the cubes, special software was written, preprocessing the cleaned data to the needed data mining algorithm or to produce the needed cubes for OLAP out of the quality assured data warehouse data. Since these processes and software components have no user interface and are developed for these specialized cases these modules are not presented in this document.

The second set of software tools developed in this part consists of two tools for handling the classifications of data. A screenshot of the graph view for changing and evolving classifications nodes was given in Figure 5. Based on the same database representation and development, a second tool was built to quickly identify the meaning of the special codes. This is shown in Figure 10. These are both pre steps for the temporal enhancement of changing OLAP dimensions.

The needed underlying data is stored in the database for medical knowledge.

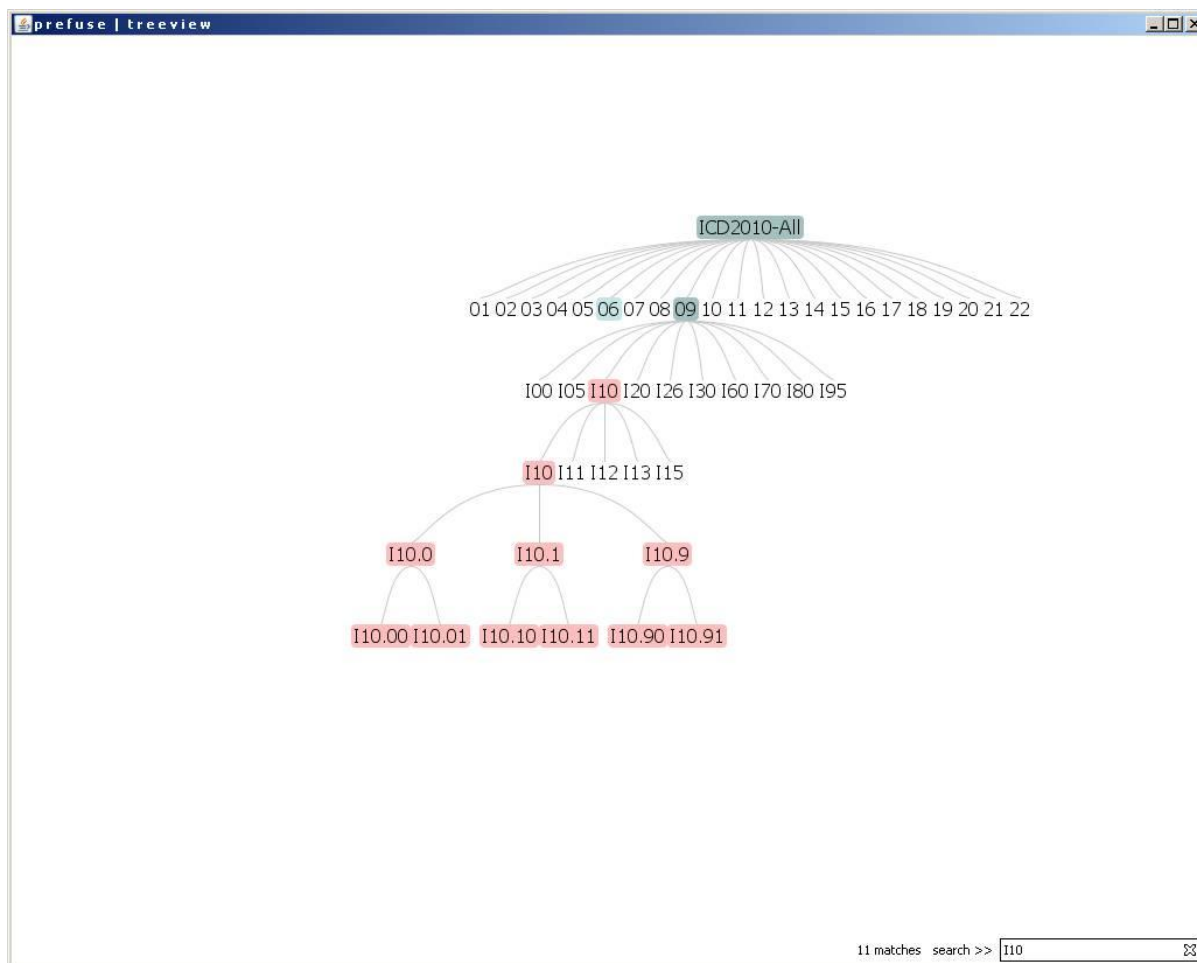


Figure 10 ICD 10 Tree Viewer for identifying the meaning of a code.

4.2.1 Database for Medical Knowledge

The database for medical knowledge holds all data needed for generation of patterns. As shown in Figure 3 this consists of the integration layer with the raw data, the data warehouse with the metadata and the derived data marts. Also the propositionalized relations are stored.

The metadata also holds the ICD10 classifications of the last years together with special transitions rules that enables the visualization of the evolution of the classification. By changing the stored data the evolution of other classifications is possible, as long as it is in the taxonomy structure needed for OLAP structures.

For this database a PostgreSQL DBMS was used. This database is outside the iCARDEA environment and located at the OFFIS. It is specially secured as described in D7.3.1. At Figure 11 an overview of the security implementation for this component is presented.

4.2.2 Database for Pattern

The database for patterns is located directly at the iCARDEA environment. It holds the found patterns of the data analysis in a DACT End user tool compatible format. This database is located directly at the environment, where iCARDEA is used. This database holds the patterns to be shown to the healthcare actors.

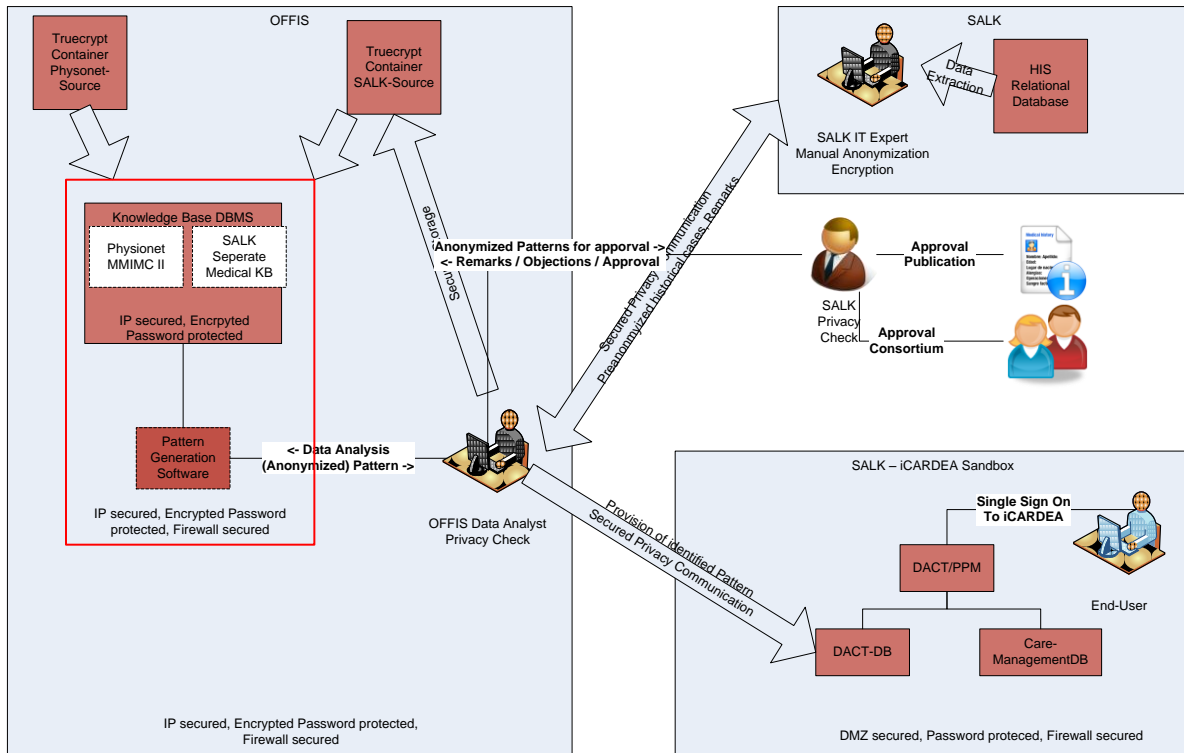


Figure 11 Overview about the Security Concept for Medical Knowledge Databases

5 CONCLUSION

With patient data electronically available more sophisticated opportunities arise to support the healthcare actors at work. In this task an attempt was presented to give healthcare actors suggestions based on previous treated patients. Since the data to be analyzed was from historical cases, a tool was provided to match the semantics of the old classifications to the classification used at the testing environment. Also it was ensured, that at no circumstances non-anonymous patient data can be revealed to unauthorized parties.

The healthcare actors received a tool, where they can have suggestion for further diagnosis or correlations of the patients to older cases based on statically valid patterns, hopefully leading and helping for a better understanding and treatment of the patients.

6 Appendix

6.1 SALK USER FORMS

Here some of the forms are presented, that were used to identify available data at the clinical systems at SALK.

At the document shown in Figure 12 the date of the implementation of the device, the type of device at the leads together with the names of the medical staff could be resolved. These were presented as selectable from predefined sets. Also the date of the implementation is available. The text was individually produced and could not be used for extraction of structured information.

The form presented in Figure 13 shows the available data for a patient. These consist of the patient's birthday, addresses, gender, all family names. For data analysis the gender, birthday and addresses and nationality could be interesting.

The form presented in Figure 14 shows all diagnosis of a patient. This form was very interesting, since it provided a lot of structured information about analysis. This was the ICD10GM code with information about, which department did the diagnosis, was it a first or discharge diagnosis or for treatment and major or minor diagnosis.

At Figure 15 and Figure 16 the discharge summary of a patient is given. These consist of the mayor dates of the operation, special facts about the device and especially the intended medications of the patient. These were identified to be potential useful for data analysis.



UNIVERSITÄTSKLINIK FÜR INNERE MEDIZIN II, KARDIOLOGIE UND INTERNISTISCHE INTENSIVMEDIZIN
VORSTAND: PRIMARIA UNIV. PROF. DR. UTA HOPPE

HERZKATHETERLABOR
Tel.: + 43(0)662/4482-3481
Fax: + 43(0)662/4482-3486

Salzburg, am 02.07.2008
FRDA /

AZ: 1452498008

Zweikammer-ICD Implantationsbericht

EPU-Nr.: [REDACTED]

HK-Nr.: [REDACTED]

Operationsdatum: [REDACTED]

Untersucher: [REDACTED]

1. Assistenz: [REDACTED]

2. Assistenz: [REDACTED]

Beidiens: [REDACTED]

Radiologietechnologe: [REDACTED]

Diagnose:

ischäm CMP

Indikation:

prophylaktische AICD Implantation

Aggregat: Medtronic

Maximo DR [REDACTED]

Rechtsventr. Elektrode: Medtronic

aktiv

Vorhofelektrode: Medtronic

aktiv [REDACTED]

Lokale Infiltrationsanästhesie links subclaviär. Präparation der Schrittmacher-Tasche submuskulär, Punktion der Vena subclavia in Doppeldrahttechnik, fluoroskopisch geführt. Einbringen einer der RV ICD Elektrode über eine 9 F Schleuse. Positionierung im apexnahen Bereich, Einschrauben der Elektrode. Überprüfung von Sensing und Reizschwelle (R-Welle 17,1 mV, RS 0,9 V). Über den verbleibenden Führungsdraht eine 6 F Schleuse für die Vorhofelektrode eingebracht. Einschrauben der atrialen Elektrode im rechten Herzhohr, gutes Sensing und niedrige Reizschwelle (P Welle 2,9 mV, RS 0,6 V). Konnektion aller Elektroden an den Zweikammer-ICD, sodann Einbringen des Aggregates in die Muskeltasche und Annahm des Aggregates am Muskel. Verschluss der Muskeltasche mit resorbierbarer Naht. Versenken der Elektroden subkutan und Wundverschluss mit resorbierbarer Naht, Hautklammern, Kompressionsverband. Der ICD wird entsprechend Protokoll getestet, Induktion von VF durch Schock on T-Wave, 18,1 Joule erfolgreich.

Vermerk: Seitens der Landeskliniken besteht kein Einwand, die im Entlassungsbrief angeführten Arzneimittel gegen wirkstoffidentie Generica auszutauschen.

Gemeinnützige Salzburger Landeskliniken Betriebsges. m.b.H. | Landeskrankenhaus Salzburg
A-5020 Salzburg | Müllner Hauptstraße 48 | Telefon +43(0)662 4482-3427 | Fax +43(0)662 4482
www.saik.at | UID-Nr. ATU57476234 | DVR 0512915 | Landesgericht Salzburg | FN 240632 s

Figure 12 Protocol of Implementation – Kind of Device and leads



LANDESKRANKENHAUS SALZBURG
UNIVERSITÄTSKLINIKUM
DER PARACELSDUS MEDIZINISCHEN PRIVATUNIVERSITÄT



UNIVERSITÄTSKLINIK FÜR INNERE MEDIZIN II, KARDIOLOGIE UND INTERNISTISCHE INTENSIVMEDIZIN
VORSTAND: UNIV.PROF. DR. MAXIMILIAN PICHLER

Herrn



Salzburg, am [REDACTED] 2010
DAHA / DAHA

[REDACTED]	AZ: [REDACTED] SVNR: [REDACTED]
------------	------------------------------------

ICD-Ambulanz

Sehr geehrter Herr [REDACTED]

wir berichten über [REDACTED] der sich am [REDACTED] 2010 zur ambulanten Behandlung in unserer Klinik befand.

Diagnosen

Zweikammer-ICD-Implantation [REDACTED]; Z.n. Reanimation bei Kammerflimmern
Ischämische Kardiomyopathie bei koronarer 2-Gefäßerkrankung
Z.n. Hinterwand-MCI 1997
Risikofaktoren: Art. Hypertonie, Hypercholesterinämie

Aggregat

[REDACTED] current DR, SN [REDACTED]

Sonden A [REDACTED]

RV [REDACTED]

Spontane Arrhythmien

Keine Episoden im Speicher.

Modus

DDD

EKG

SR, f 60/min, durchgehend Vorhofstimulation, QRS 100 ms, unauffällige Repolarisation

Messwerte

Batteriestatus	3.2V	Ladezeit	11.10Sek.	Schockimpedanz	40 Ohm
RV Sensing	7.8mV	A Sensing	4.0mV		
RV RS	0.5V/ 0.5 ms	A RS	0.6V/ 0.5 ms		
RV Impedanz	460Ohm	A Impedanz	350Ohm		

Procedere

Termin am [REDACTED] 2011

Gemeinnützige Salzburger Landeskliniken Betriebsges. m.b.H. | Landeskrankenhaus Salzburg
A-5020 Salzburg | Müllner Hauptstraße 48 | Telefon +43(0)662 4482-3427 | Fax +43(0)662 4482
www.salk.at | UID-Nr. ATU57476234 | DVR 0512915 | Landesgericht Salzburg | FN 240832 s

Figure 15 Discharge summary at the ambulance – Page 1 CIED Data

Therapievorschlag

Medikamente	Früh	Mittag	Abend	Nacht
Nebivolol 5mg	1			
T-ASS 100mg		1		
Plavix 75mg		1		
Lisihexal 10mg <i>Lisinhexal</i>	1			
HCT Beta 25mg <i>HCT Beta 25mg</i>	1/2			
Simvastatin 20mg			1	

Vermerk: Seitens der Landeskliniken besteht kein Einwand, die im Entlassungsbrief angeführten Arzneimittel gegen wirkstoffidentische Generica auszutauschen.

Dieses Schriftstück wurde elektronisch validiert und ist auch ohne handschriftliche Abzeichnung gültig.

Mit freundlichen Grüßen

Primar

Oberarzt

Assistenzärztin

Figure 16 Discharge summary at the ambulance – Page 2 Medications

6.2 RESULTS OF OLAP CUBES

At this section another chart showing the results of the OLAP analysis is given. The Figure 17 shows the distribution of patients as described in section 3.4.1.1. In this chart it is also calculated and plotted, what the average distribution of the patient by gender is. Then the other plots present the distribution in 5year age bins. It seems, that females, who get a device, appears at younger years and male people more around the age of 57.

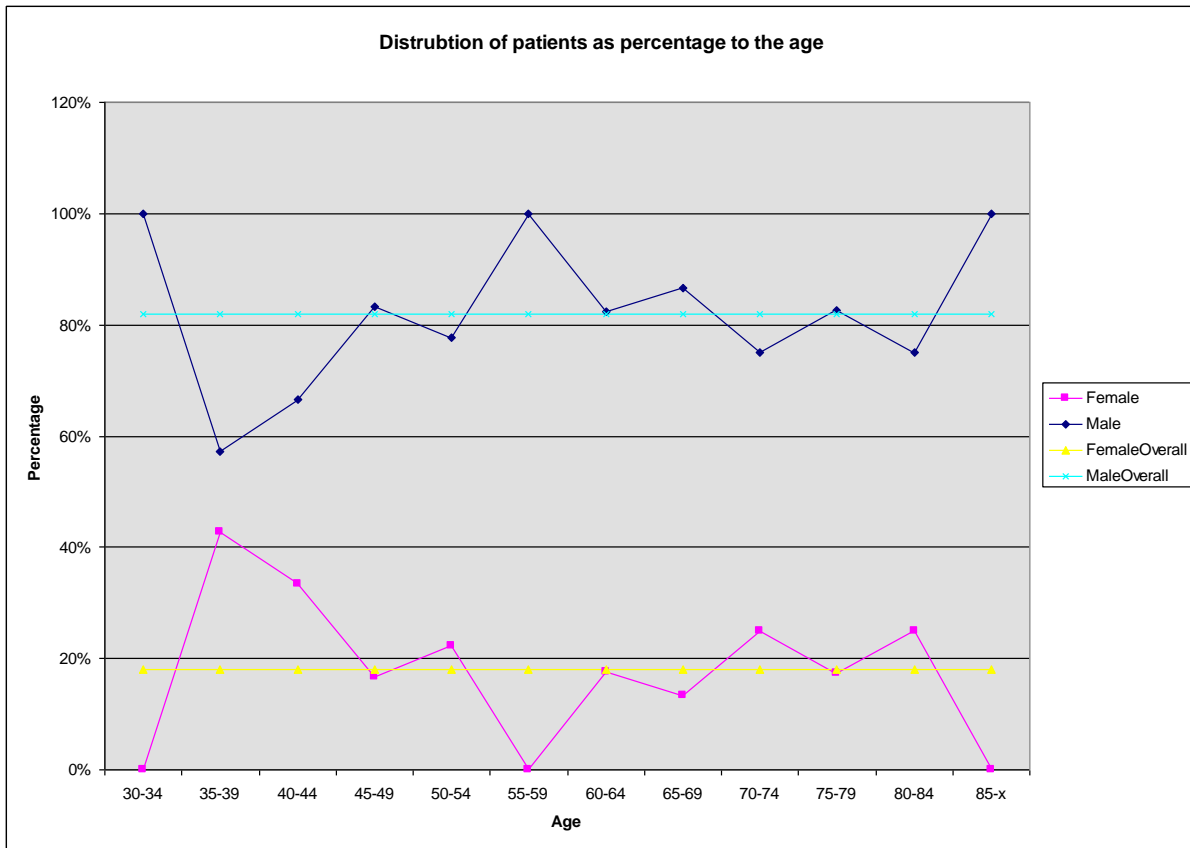


Figure 17 OLAP Chart showing distribution of Patients

6.3 PHYSIONET DESCRIPTIONS

This section holds a document of the Physionet Description, created when exploring MIMIC II for potential usefulness and also some results from the preanalysis study conducted after integrating the data. The comments to usefulness are from the data analysis experts before discussing the issues with the clinical professionals.

6.3.1 Research of useful MIMIC II Data

This section provides an overview about the relations provided by MIMIC II, the short description and the amount of data items that where integrated at OFFIS for the relations. The descriptions were not modified and are provided from Physionet. The fields in the column 'Table' and 'Amount' contains links to more technical descriptions on the Physionet webpage.

Tables:

Definitions:

Table	Description	Amount
MIMIC2V26.D CAREGIVERS	Care givers...	11015
MIMIC2V26.D CAREUNITS	Care units...	22
MIMIC2V26.D CHARTITEMS	Items which can be entered on a patient's chart...	4832

MIMIC2V26.D_CODEDITEMS	Items with an attached code; may be used in DRGEVE...	3339
MIMIC2V26.D_DEMOGRAPHICITEMS	Demographic items: items that can be entered into ...	88
MIMIC2V26.D_IOITEMS	IO items. Patient fluid input/output records...	6808t
MIMIC2V26.D_LABITEMS	Laboratory tests....	713t
MIMIC2V26.D_MEDITEMS	Medication items. Possible medications which can b...	405
MIMIC2V26.D_PARAMMAP_ITEMS	Types of parameter mapping provided. The actual ma...	2
MIMIC2V26.PARAMETER_MAPPING	Mappings between parameters which do not fit in th...	33024t

Figure 18: Table of MIMIC II Definitions

Content:

Table	Description	Amount (DDL)
<u>MIMIC2V26.ADDITIVES</u>		<u>149212</u>
<u>MIMIC2V26.ADMISSIONS</u>	The record of patient admissions to the hospital...	<u>31459</u>
<u>MIMIC2V26.A_CHARTDURATIONS</u>	Chart event durations...	<u>3196426</u>
<u>MIMIC2V26.A_IODURATIONS</u>	Duration of IO events...	<u>342680</u>
<u>MIMIC2V26.A_MEDDURATIONS</u>	Duration of medication events...	<u>1464553</u>
<u>MIMIC2V26.CENSUSEVENTS</u>	Events which record ICU transfers...	<u>44615</u>
<u>MIMIC2V26.CHARTEVENTS</u>	Events which occur on a patient chart....	<u>19242247</u>
<u>MIMIC2V26.COMORBIDITY_SCORES</u>		<u>31347</u>
<u>MIMIC2V26.DB_SCHEMA</u>	Database schema information....	<u>DDL script</u>
<u>MIMIC2V26.DELIVERIES</u>		<u>79919</u>
<u>MIMIC2V26.DEMOGRAPHICEVENTS</u>	Events indicating demographic information about a ...	<u>180019</u>
<u>MIMIC2V26.DEMOGRAPHIC_DETAIL</u>		<u>31459</u>
<u>MIMIC2V26.DRGEVENTS</u>	Events indicating DRGs recorded for a patient...	<u>31378</u>
<u>MIMIC2V26.D PATIENTS</u>	Patient table. Contains patient specific informati...	<u>27903</u>
<u>MIMIC2V26.ICD9</u>	ICD9 code records for each hospital admission....	<u>267150</u>
<u>MIMIC2V26.ICUSTAYEVENTS</u>	Unique ICU stay definitions per patient...	<u>35075</u>
<u>MIMIC2V26.ICUSTAY_DAYS</u>		<u>216888</u>
<u>MIMIC2V26.ICUSTAY_DETAIL</u>		<u>35075</u>
<u>MIMIC2V26.IOEVENTS</u>	Fluid input/output (IO) events....	<u>10581079</u>
<u>MIMIC2V26.LABEVENTS</u>	Laboratory tests...	<u>16820162</u>
<u>MIMIC2V26.MEDEVENTS</u>	Medication events....	<u>3784614</u>
<u>MIMIC2V26.MICROBIOLOGYEVENTS</u>	Events indicating microbiology tests taken from a ...	<u>327878</u>
<u>MIMIC2V26.NOTEEVENTS</u>	Note events. Patient nursing/doctor notes, dischar...	<u>1091092</u>
<u>MIMIC2V26.POE_MED</u>	Medications which make up a POE_ORDER. For each en...	<u>1769005</u>
<u>MIMIC2V26.POE_ORDER</u>	POE Orders. Each row in this table indicates a POE...	<u>1471262</u>
<u>MIMIC2V26.PROCEDUREEVENTS</u>	Events indicating procedures performed on a patien...	<u>133062</u>
<u>MIMIC2V26.TOTALBALEVENTS</u>	Total fluid balance events....	<u>1891215</u>

Figure 19 Table of Mimic II Contentrelations

These tables were all integrated and the data analyzed for correctness and usefulness.

6.3.2 Definition-Tables:

At this section the personal remarks on the integrated data are given. This means, these are comments to the relations from the data analyst expert when the actual data of the tables was examined. Therefore sometimes the significance may be not clear to the reader. At this section for the definition tables.

6.3.2.1 [MIMIC2V26.D CAREGIVERS](#)

- Links a CaregiverID to the profession, like Co-Worker, Medical Doctor or Unit A. Intended to be a catalogue.
- It holds 237 different labels. Not in good quality (There are a lot of repeating content, different extensions (same semantic, different syntax)) and not useful for analysis.
- Integrated since needed for Foreign Key constraints.
- To interpret the syntax, following URL could be useful: http://en.wikipedia.org/wiki/List_of_abbreviations_for_medical_organisations_and_personnel

6.3.2.2 [MIMIC2V26.D CAREUNITS](#)

- Addresses the CareunitID.
- Intended to be a catalogue
- It holds 22 different Labels, all distinct but with no official standard for meaning.
- Integrated since needed for Foreign Key constraints.

6.3.2.3 [MIMIC2V26.D CHARTITEMS](#)

- Addresses the Items which can be entered on a patient's chart, name of the parameter.
- Intended to be a catalogue
- It holds 4832 different labels. Not in good quality (There are a lot of repeating content, different extensions (same semantic, different syntax)) and not useful for analysis.
- Integrated since needed for Foreign Key constraints.
- Maybe it could be used for identifying patients with _VT_ or AT. VT could be easily spotted, but AT was not

6.3.2.4 [MIMIC2V26.D CODEDITEMS](#)

- Addresses the meaning and interpretation depend on the category; not necessarily unique.
 - DRGEVENTS, MICROBIOLOGYEVENTS, or PROCEDUREEVENTS
- The procedures may be interesting for data analysis
- "PROCEDURE" and "HFCA_DRG" has a description like "LOBECTOMY OF LUNG"
- "MICROBIOLOGY" has no description but has category and label
- Integrated since needed for Foreign Key constraints.

6.3.2.5 [MIMIC2V26.D DEMOGRAPHICITEMS](#)

- Addresses 6 different topics of patient demographics:
 - "MARITAL STATUS"
 - "ETHNICITY"
 - "ADMISSION SOURCE"
 - "OVERALL PAYOR GROUP"

- "ADMISSION TYPE"
 - "RELIGION"
 - Could be interesting. But due to European laws it is prohibited or complicated to make medical research over religion / ethics combined with medical data.
 - Integrated since needed for Foreign Key constraints and the data is interesting.
 - Not to be used for data analysis due to legal / ethical constraints
- 6.3.2.6 [MIMIC2V26.D_IOITEMS](#)
- Addresses Fluids that where put into the patient.
 - Has 19 different categories (Feeding, Nutirition, Infusions)
 - Seems not to be really sorted
 - Integrated since needed for Foreign Key constraints
 - Maybe Category could be used for data analysis
- 6.3.2.7 [MIMIC2V26.D_LABITEMS](#)
- Addresses possible LAbtest and results. Also provides LOINC Codes for some of them.
 - Has three categories:
 - "CHEMISTRY"
 - "HEMATOLOGY"
 - "BLOOD GAS"
 - 139 items has no LOINC code and 574 has one
 - Could be used for assessing the influence of lab results or drugs
 - Integrated since needed for Foreign Key constraints
 - Use has to be discussed with medical partners
- 6.3.2.8 [MIMIC2V26.D_MEDITEMS](#)
- Addresses the possible medications which can be administered to a patient.
 - It has 405 items.
 - Can be used for data analysis. Quality seems very good.
 - Only drawback, that no dimension-metadata (structure) is provided
 - Used in MIMIC2V26.ADDITIVES
- 6.3.2.9 [MIMIC2V26.D_PARAMMAP_ITEMS](#)
- No content. Not usefull.
 - Only explains that it was used for Mapping between Subjext_id and old PIDs
- 6.3.2.10 [MIMIC2V26.PARAMETER_MAPPING](#)
- The real Mapping of Subject_ID to PID or CASE_ID.
 - Not to be used directly at data analysis, but could be needed for associating dataitems together.

6.3.3 Content:

At this section, as the previous, the personal remarks on the integrated data are given. This means, these are comments to the relations from the data analyst expert when the actual data of the tables was examined. Therefore sometimes the significance may be not clear to the reader. In this section the real patient data is observed.

6.3.3.1 [MIMIC2V26.ADDITIVES](#)

- Contents the given Medication of a patient based on the Meditems.
- Very useful for data analysis.
- It has also the amount and units of the dose given. Also 8 different types of route is stated
 - "IV Drip"
 - "Gastric/Feeding Tube"
 - "Nasogastric"
 - "By Mouth"
 - "Intravenous"
 - "Intravenous"
 - "Intravenous Push"
 - "Drip"
- It could be used to make an own dimension, which will probably correlate with the medications.

6.3.3.2 [MIMIC2V26.ADMISSIONS](#)

- The record of patient admissions to the hospital
- Only holds the stay of patients.
- Could be useful for timeseries.
- The year is totally faked!
- Not useful for intended data analysis, but maybe for time series analysis / temporal data analysis.
- It could be also useful for : Had one stay, had multiple stays

6.3.3.3 [MIMIC2V26.A.CHARTDURATIONS](#)

- Duration of Chart events. These are connected to the Care Unit and the hospital stay.
- Only interesting, if Charts can be analyzed.
- Over this relation, the chartitems can be accessed. This means, via this relation it can be decided which patients had a VT (and maybe AT) chart event.

6.3.3.4 [MIMIC2V26.A.IODURATIONS](#)

- Duration of IO events. Connections the IOITEMS to the patient and its stay
- Useful for data analysis since this gives access to d_ioitems

6.3.3.5 [MIMIC2V26.A.MEDDURATIONS](#)

- Duration of medication events. Connects D_Meditems to a patient. This is also done by additives.
- Useful for data analysis. The timestamps are not needed and also not useful, since they are mixed up by the anonymization done by Physionet.

6.3.3.6 [MIMIC2V26.CENSUSEVENTS](#)

- Events which record ICU transfers: Transfers from the originating unit to the destination unit and the length of stay at ICU unit.
- Not interesting for data analysis at all

6.3.3.7 [MIMIC2V26.CHARTEVENTS](#)

- Maybe useful to connect the D_Chartitems.
- But this could be also done by a_chartdurations a lots faster and with only 20MB datasets and not 18GB!
- But according to a selection: Charthevents has more patients with VT
- There is a big difference between the patients for Charthevents and chartdurations.
- Only the first 2000 patients where integrated, first patient has 68315 records, and this isn't the biggest.
- But seems not useful: 41% of the 2000 patients had any kind of VT and 40% of spontaneous VT.

6.3.3.8 [MIMIC2V26.COMORBIDITY_SCORES](#)

- Holds a lot of scores for different types.
- Maybe the people with the following conditions could be interesting. This should be decided by a doctor
 - Score for congestive heart failure
 - Score for cardiac arrhythmias
 - Score for alcohol abuse
 - Score for drug abuse
 - Score for hypertension (combined complicated and uncomplicated)
- The relation is already a binary propositionalization fact table.
- 27385 Patients.

6.3.3.9 [MIMIC2V26.DB_SCHEMA](#)

Only technical information about the database. No patientdata

6.3.3.10 [MIMIC2V26.DELIVERIES](#)

- Contains information about deliveries to the patient including dosage. Makes a connection to D_IOITEMS
- Therefore it is integrated.
- For 17829 Patients

6.3.3.11 [MIMIC2V26.DEMOGRAPHICEVENTS](#)

- Events indicating demographic information about a patient, recorded upon hospital admission.
- This connects the Admissions and D_Demographiocitems to a patient.
- The first one should not, and the second one is not allowed to be analyzed.
- Therefore the data is not used for data analysis. But integrated for potential data linking needs
- 27442 Patients

6.3.3.12 [MIMIC2V26.DEMOGRAPHIC_DETAIL](#)

- This holds the demographic data in one relation: Martial Status, Ethnic, Payor Group, Religion, Admission Type.
- Could be funny to analyze, but is also legal problematic.
- Therefore the data is not used for data analysis. But integrated for potential data linking needs
- 27442 Patients

6.3.3.13 [MIMIC2V26.DRGEVENTS](#)

- Events indicating DRGs recorded for a patient. Connect D_Codeitems to a Patient. It is also connected to Admissions
- DRG seems to be similar to Germany: Diagnosis related Group
- Should be used, if D_Codeitems is interesting

6.3.3.14 [MIMIC2V26.D PATIENTS](#)

- Patient table. Contains patient specific information such as DOB, DOD, and sex. Date of death is obtained by 2 methods: 1. If the patient died in the hospital, the died_in_hospital flag is 'Y' and the date of discharge is used as the date of death. 2. Patient social security numbers are matched to the US death records to obtain out of hospital date of death.
- Very important. Central class for Patient-data links
- 27903 Patients where integrated

6.3.3.15 [MIMIC2V26.D WAVEFORM_SIG](#)

- Describes the signal class and the type of signal.
- Not integrated, since it only says something about the technical staff

6.3.3.16 [MIMIC2V26.ICD9](#)

- ICD9 code records for each hospital admission.
- There are 7612 different CODE-Description items and only 5392 codes.
- Integrated und should be used.
- Based on a catalogue and structured. Perfect for data analysis. Only drawback: Different definitions.

6.3.3.17 [MIMIC2V26.ICUSTAYEVENTS](#)

- Gives an overview of ICU stays: First careunit of patient, last one, length of ICU stay and time of stay.
- Only interesting if length of stay is important.

6.3.3.18 [MIMIC2V26.ICUSTAY_DAYS](#)

- Gives the time as days. No more information as icustayevents.
- Totally useless for data analysis. More a data preparation step.

6.3.3.19 [MIMIC2V26.ICUSTAY_DETAIL](#)

- Also seems as if data prepared: but very interesting:

- Holds DOB, DOD (Date of birth, date of death), Gender as M or F, height weight,
- Very good for data analysis to decide about the kind of patients.
- Calculation of DOD – DOB gives age in Days

6.3.3.20 [MIMIC2V26.IOEVENTS](#)

- Fluid input/output (IO) events. Who gave what amount at what unit into the patient at which stay
 - Connects the d_ioitems to a patient (itemid and altid). The difference between them is not clear: All datasets as itemid, but not all altid. IT seems, altid is a more general term. It seems that some are very frequent together

6.3.3.21 [MIMIC2V26.LABEVENTS](#)

- Just connects the patient to there labresults of d_labitems.
- Also Admissions is connected by hadm_id. But not available for all entries. Maybe this is for labresults after the patient was discharged or for follow visits
- It has a date, and a flag for abnormal, delta and Null.
- Maybe this could be useful for data analysis

6.3.3.22 [MIMIC2V26.MEDEVENTS](#)

- Connects the patient with medication events. Includes the timestamp, the medicament from d_meditems.
- Not available for every patient. List Care giver, volume / dosage of med, route
- Route "Intravenous" "IV Drip" "Drip"
- 12 different dose-units

6.3.3.23 [MIMIC2V26.MICROBIOLOGYEVENTS](#)

- Events indicating microbiology tests taken from a patient.
- What was tested, and what are the results. Connects to d_codeitems
- Also has a result tag: The interpretation of the test: R, P, I, or S (or null when not available). This has to be explained, but could be interesting for data analysis

6.3.3.24 [MIMIC2V26.NOTEVENTS](#)

- Note events. Patient nursing/doctor notes, discharge summaries and reports such as echo, ecg, radiology:
- Not for data analysis.
- Has plain text of the discharge summary or other CDAs.
- The four report categories are: "RADIOLOGY_REPORT" "MD Notes" "DISCHARGE_SUMMARY" "Nursing/Other"

6.3.3.25 [MIMIC2V26.POE_MED](#)

- Medications which make up a POE_ORDER. For each entry in the POE_ORDER table there will be 1 or more rows indicating the medication administered
- Has type of drug and drugs name. Also generic name is available, but not often used.

6.3.3.26 [MIMIC2V26.POE_ORDER](#)

- POE Orders. Each row in this table indicates a POE order with start and stop dates and instructions for administration. The POE_MED table contains the individual medications which comprise the order.
- It has also medication: Seems better (generic) as the poe_med. But also not always available. Also doses are per 24 hours.
- Could be used for data analysis.

6.3.3.27 [MIMIC2V26.PROCEDUREEVENTS](#)

- Events indicating procedures performed on a patient. Connects a patient to d_codeitems. The procedures are chronologically ordered.
- The linkage is to d_codeitems.
- To be used, if d_codeitems should be used.

6.3.3.28 [MIMIC2V26.TOTALBALEVENTS](#)

Total fluid balance events. Has to be explained by medical professionals.

6.3.4 Keyword for cardiac Patients

The following table provides the comments of the medical professionals of HCPB to the keywords at the column 'label' which seemed promising to identify cardiac patients for data analysis. The 'X' in the column 'Use less' indicates, that this keyword should not be used for identifying patients. The comment, also provided by the medical experts, gives the potential meaning of the label.

Physionet itself did not provide information about the meaning of the labels.

label	Use less	AF	AT	VT	SVT	Comment
Atrial Sens mV/mA						These are programmed parameters of pacemakers
Atrial Sens/Capture						These are programmed parameters of pacemakers
Atrial Threshold						These are programmed parameters of pacemakers
AtrialmA/Sensitivity						These are programmed parameters of pacemakers
BIPAP - Est. Vt	x					It could be volume tidal of a ventilator
Cardiac consult	x					It could be or not a patient with cardiac problems, for example it could be a control consult or test
Cardiac Echo	x					It could be or not a patient with cardiac problems, for example it could be a control consult or test
Cardiac Index	x					It could be or not a patient with cardiac problems, for example it could be a control consult or test
cardiac index o	x					It could be or not a patient with cardiac problems, for example it could be a control consult or test
Cardiac Murmur	x					It could be or not a patient with cardiac problems, for example it could be a control consult or test
Cardiac Rhythm	x					It could be or not a patient with cardiac problems, for example it could be a control consult or test
Education Response	x					

epicardial pacemaker				
exh. vt	x			It could be volume tidal of a ventilator
HIGH VT	x			It could be volume tidal of a ventilator
ICP-ventriculostomy	x			It is a cardiac surgery
ICP ventricle	x			It a cardiac surgery
insp. vt	x			It could be volume tidal of a ventilator
irrigate pericardial		x		It is an ablation procedure
LeadII cardiac strip				
Low Exhaled Vt	x			It could be volume tidal of a ventilator
low Vt	x			It could be volume tidal of a ventilator
LOW VT	x			It could be volume tidal of a ventilator
Motor Response	x			It is a neurological patient/problem
Pacer Wires Atrial				
Pain Level/Response	x			It is a neurological patient/problem
PCV Exh Vt (Obser)	x			It is a programmed parameter of ventilator
PCV Insp Vt (Obser)	x			It is a programmed parameter of ventilator
Pericardial Drainage	x			It is a cardiac surgery
Pericardial flush	x			It is a cardiac surgery
Pericardial Pressure	x			It is a cardiac surgery
Pupil Response R/L	x			It is a neurological patient/problem
Resp Rate (Spont)	x			It is a parameter of pneumonological patient
Responds to Stimuli	x			It is a neurological patient/problem
Response To Stimuli	x			It is a neurological patient/problem
Spon Minute VE L/min	x			It is a parameter of pneumonological patient
Spon RR (Mech.)	x			It is a parameter of pneumonological patient
Spon. Vt (L) (Mech.)	x			It is a parameter of pneumonological patient
sponge bath	x			
Spont Minute vols	x			It is a parameter of pneumonological patient
Spont Resp rate	x			It is a parameter of pneumonological patient
spont tidal volumes	x			It is a parameter of pneumonological patient
spont Tidal volumes	x			It is a parameter of pneumonological patient
SPONT TV'S		x	x	It could be a SVT or respiratory parameter
spont Ve	x			It is a parameter of pneumonological patient
Spont VT	x	x	x	It could be a SVT or respiratory parameter
SPONT VT	x	x	x	It could be a SVT or respiratory parameter
spont Vt's	x	x	x	It could be a SVT or respiratory parameter
Spont. Resp. Rate	x			It is a parameter of pneumonological patient
Spont. Tidal Volume	x			It is a parameter of pneumonological patient
Spontaneous Movement	x			It is a neurological patient/problem
spontaneous vt		x		It could be a VT or a respiratory parameter
spontaneous VT		x		It could be a VT or a respiratory parameter
Spontaneous VT		x		It could be a VT or a respiratory parameter
svt			x	
Tidal Volume (Spont)	x			It is a parameter of pneumonological patient
TOF Response/Twitch	x			It is a parameter of pneumonological patient
Vd/Vt Ratio (40-60%)	x			
Vd/Vt:	x			
ventriculostomy icp	x			It is a cardiac surgery
Verbal Response	x			It is a neurological patient/problem
Vt		x		It could be a VT or a respiratory parameter
VT		x		It could be a VT or a respiratory parameter
Vt [Spontaneous]	x			It is a parameter of ventilator
Vt [Ventilator]	x			It is a parameter of a ventilator

Figure 20 Table of potential Keywords and the clinical rating of usefulness

6.4 LITERATURE

- [Azevedo2008] Ana Azevedo, Manuel Filipe Santos: KDD, SEMMA and CRISP-DM: a parallel overview In Ajith P. Abraham(Eds): Proceeding of IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands, July 24-26, 2008. ISBN 978-972-8924-63-8
- [Bauer2008] Andreas Bauer, Holger Günzel (Eds.): Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung 3rd ed. dpunkt Verlag, November 2008, ISBN 3-89864-540-1
- [Fayyad1996] Usama Fayyad, Gregory Piatetsky-Shapiro, Smyth Padhraic (1996): From Data Mining to Knowledge Discovery in Databases In AI Magazine, American Association for Artificial Intelligence, California, USA
- [Han2006] Jiawei Han, Micheline Kamber: Data Mining: Concepts and Techniques, 2nd ed. The Morgan Kaufmann Series in Data Management Systems, Jim Gray(Eds) Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6
- [ICD2005]DIMDI - Deutsches Institut für Medizinische Dokumentation und Information: ICD-10-GM 2005 Systematisches Verzeichnis. Systematisches Verzeichnis zur Internationalen statistischen Klassifikation der Krankheiten und verwandter Gesundheitsprobleme - German Modification. Deutsche Krankenhaus Verlags-Gesellschaft, (2004)
- [ICD2006] DIMDI - Deutsches Institut für Medizinische Dokumentation und Information: ICD-10-GM Version 2006. Systematisches Verzeichnis. Deutsche Krankenhaus Verlags-Gesellschaft, (2005).
- [ICD2007] DIMDI - Deutsches Institut für Medizinische Dokumentation und Information: ICD-10-GM Version 2007. Band I: Systematisches Verzeichnis. Deutsche Krankenhaus Verlags-Gesellschaft, (2006)
- [ICD2011]DIMDI - Deutsches Institut für Medizinische Dokumentation und Information: ICD-10-GM Version 2011: Band I: Systematisches Verzeichnis. Deutsche Krankenhaus Verlags-Gesellschaft, (2010)
- [Luepkes2011] Christian Lüpkes: Ad-hoc Datentransformationen für Analytische Informationssysteme. Gassler, Wolfgang; Zangerle, Eva; Specht, Guenther (Eds): Proceedings of the 23rd GI-Workshop Grundlagen von Datenbanken 2011.