MASTER THESIS

TITLE: Analysis of financial and technical feasibility of a clinicians generated data platform of Fibromyalgia syndrome patients

MASTER DEGREE: Master in Science in Telecommunication & Engineering Management

AUTHOR: Tobias Joschko

ADVISOR: Jesus Alcober

DATE: October 2nd, 2018

Title: Analysis of financial and technical feasibility of a clinicians generated data platform of Fibromyalgia syndrome patients

Author: Tobias Joschko

Advisor: Jesus Alcober

Date: October 2nd, 2018

Abstract

This master thesis analyzes the technical and economical feasibility for a medical database, based on clinically generated data of patients with the fibromyalgia syndrome. The main idea is to collect patient data on a regular basis during standard visiting hours at their doctor. Therefore it is essential to provide a data collection platform that can be simply used by the patient and doctor.

The collected information (no personal data) shall be shared between researchers to enhance collaborative studies, make studies with rare diseases possible as well as to reduce the cost and effort to gather a big enough cohort group for the study.

There are already several medical databases in place that collect and share patient information for research. Yet, despite the significant socioeconomic impact of fibromyalgia, no large database about this disease exists.

An introduction to the fibromyalgia syndrome and its impact on society are given. Furthermore medical database technologies and medical database projects for other diseases are described.

The presented technologies are further analyzed for their usefulness of creating a database to collect information about fibromyalgia syndrome patients and to use it to enhance its research. Additionally the legal requirements for maintaining such a platform as well as the potential cost are examined. Two possible business models to provide such a platform with funding are presented.

Last but not least a possible use case for the collection of patient data via a survey created with REDCap and the integration process into i2b2 has been created and possible suggestions for improvements in the future have been made to bring the platform to a release ready state.

CONTENTS

INTRODUCTION

In our society a healthy life plays an important role. The fibromyalgia syndrome has a strong socioeconomic impact onto our society. It has a severe impact onto people's everyday life and lowers their quality of life dramatically. This is accompanied by a high economic impact as well because it affects the people's ability to follow their jobs. Yet the knowledge about fibromyalgia is fairly limited if it comes to awareness about the disease as well as about the most effective medical treatments. In order to tackle these challenges it is of high importance to collect as much information about the disease as possible and to make it accessible to research.

The general objective is to conduct a technical and business feasibility study for the creation of a database based on clinically generated data of patients with the fibromyalgia syndrome. Such a platform should be used to gather patient data in a simple and secure way according to GDPR and data and privacy regulations of each participating country within the European Union.

The first objective is an introduction to the fibromyalgia syndrome, its medical symptoms, worldwide occurrence and the impact it has on the patient's quality of life as well as its socioeconomic impact on society as a whole. Furthermore a short overview about fibromyalgia research, hence the motivation for creating such a patient database is given.

The second objective is to give an overview about the status of existing fibromyalgia syndrome databases and current state of the art database technology that is available on the market to deal with medical patient and privacy sensitive data.

The third objective is to conduct a feasibility study out of the technical perspective. This entails a thorough analysis of already existing database technologies and how they could be applied for the specific use case of a fibromyalgia patient data.

The fourth objective is to conduct a feasibility study out of an economical and legal point of view. Next to the technical challenges for a fibromyalgia patient database possible business models as well as privacy and data protection rights (e.g. General Data Protection Regulation (GDPR) in the European Union) shall be analyzed.

The fifth objective is to prepare a simplified use case to show how patient data can be collected and stored inside a database. This shall give the proof of concept for such a fibromyalgia patient database.

1. Introduction to fibromyalgia

This chapter gives a basic overview about the fibromyalgia syndrome (from hereon called FMS), its symptoms and the magnitude of impact it has on society. Furthermore, some basic concepts in the medical analysis and research of the fibromyalgia syndrome are introduced to prepare the reader for the motivation of creating such a database platform. Moreover, at the end of the chapter the research objectives of a consortium of organizations whose aim is to tackle several challenges of FMS will be explained.

1.1.What is fibromyalgia

FMS is a disease, which is mainly characterized by chronic pain. The symptoms and signs can vary widely in form and strengths depending on the patient's mental and physical state. The first definition of FMS in 1990 by the American College of Rheumatology (from hereon ACR) was based on ruling out other potential diseases and causes for the patient's condition as well as checking for the presence of specific FMS symptoms.

The treatment commonly consists of spa therapy, aerobic exercise, cognitive behavioral therapy and the drug Amitriptyline[1].

Even though FMS nowadays is understood relatively well the pathogenesis[2] of FMS is still unknown. Furthermore, the diagnosis, classification and treatment of FMS remain controversial.[1]

1.1.1.Overview of the symptoms and diagnosis

FMS is a disease, which has a big variety of symptoms and can affect both the body as well as the mind of a patient suffering from the condition. The symptoms are very diverse and can include restless sleep, headaches and mood disturbances to just name a few. A more detailed description of the symptoms can be found in Annexes A1.

The most common definition of FMS nowadays is the 2010 approved diagnostic criteria by the ACR. The 2010 criteria should make it easier to be used in primary care and introduced a scale for the severity of the symptoms based on

[1] Amitriptyline is a medication which main purpose is the treatment of different mental diseases, such as depressive or anxiety disorders.

[2] Pathogenesis is the biological mechanism of a disease and can also describe its development and origin.

the characteristics of FMS.[2] Therefore the 2010 definition of FMS uses indices and categorical scales. This makes the storage in a database and the processing of this information easier for a mathematical model, because the data variables can be stored in a numerical way of for the symptom description.

In order to evaluate the symptoms of FMS the ACR uses a widespread pain index (WPI) and a severity scale (SS). The WPI is used to describe the pain of up to 19 areas of the body in which the patients experienced pain in the previous week. The WPI is categorized into the category groups 0,1,2 and 3 which each can have a score of 0-3 and a cumulative score of 0-12.[2]

The SS is used to assess the level of the patient's symptoms with a score between 0 and 3 for un-refreshed awaking, fatigue, somatic[3] and cognitive symptoms.

According to the 2010 criteria a diagnosis of FMS is satisfied if the following three conditions are present in the patient [2]:

- (WPI of 3 - 6 **and** SS ≥ 9) **OR** (WPI ≥7 **and** SS ≥ 5).
- The symptoms have to be present for at least 3 months at a constant level.
- No other disorder can explain the patient's pain.

1.1.2. Occurrence and (economic) impact on society

It is assumed by population-based studies that between 7% to 11% of the world's population suffer of chronic widespread pain and about 1% to 5% of FMS. Over 90% of the people affected with FMS are women in the age range of 20 to 50.

FMS impairs the patient's quality of life, causes stress and therefore leads to considerable direct and indirect costs for the disease.[1] It is said that compared to patients of other chronic pain syndromes the health-related quality of life for patients with FMS is lowere. [3]

A study from 2012 analyzed several databases, scientific papers and literature on the prevalence of chronic pain conditions and their related costs within the European Union[4]:

- 20% to 50% of persons with FMS were unable to work.
- 50 % received social payments.
- Of all chronic pain conditions FMS causes the highest rate of unemployment (6%) and the greatest number of days absent from work.

---

[3] Somatic can describe a variety of symptoms related the nervous system. This can include pain, nausea, vertigo, etc.

Experts in general agree that the prevalence of FMS is increasing and also that it is causing an enormous financial burden. Nevertheless, sufficient evidence to support these claims is missing.

A study conducted in Canada, in 1999, yielded that the cost and utilization of medical services is double for patients suffering from FMS compared to patients suffering of widespread pain. [5]

A US study from 2007 showed that the healthcare costs of FMS patients are three times as high as compared to the control group of randomly selected patients from a health insurance database.

A study of Sicras-Mainar and his team compared the incremental cost of FMS patients caused by utilization of healthcare but also non-healthcare resources, comparing these costs to a reference group using European primary care. The incremental costs of FMS patients compared to the reference group is €5000 higher, whereas only 600€ are contributed to the use of healthcare resources. Furthermore, according to the same study, patients suffering from FMS show a higher occurrence of comorbidities, a higher number of average workdays missed as well as six more visits to the doctor per year when comparing them to the reference population group. [6]

A study conducted in the Netherlands calculated the average annually socioeconomic costs caused per patient for different musculoskeletal diseases. It states: €8533 for chronic low back pain, €7813 for FMS and €3205 for ankylosing spondylitis[4]. Central causes for higher costs for FMS and chronic low back pain have been attributed to the lost ability to perform work, aids, adaptations and care-taking. Moreover, FMS and chronic low back pain also score lower in terms of well-being.[7]

When taking these numbers into consideration, assuming that 3% of the population suffers from FMS on average, the total annual cost of FMS would be around €12 billion for a country like Germany with a population of 80 million people and around €6.9 billion for Spain with a population of 46 million people, of which only €960 million and €552 million respectively are contributed by pharmacological treatment. Having in mind that these numbers only represent the additional costs for FMS compared to the populations reference group the total absolute cost of FMS can be assumed to be approximately €18.8 billion for Germany and €10.8 billion for Spain annually.[6] Furthermore, since these studies date back to the years of 2004 and 2009 an even higher number can be expected for the year 2018.

It is apparent that **FMS not only has a big economical impact on society but** also on the quality of life for people suffering from it. Hence, efforts in the medical research for FMS should be directed into decreasing the cost of the FMS treatments and increasing the patties quality of life.

---

[4] It is a form of arthritis accompanied by a long-term inflammation of the spine and the joints.

1.2.Research on FMS

The magnitude of impact of FMS is not yet completely understood in every detail. Sometimes people might not even realize that they suffer from FMS. Therefore research should analyze the socioeconomic impact of FMS as well as to increase the understanding of the symptoms and course of the disease. Furthermore each country with a different healthcare systems should aim to use the latest data available to provide more efficient evidence based treatment for FMS patients.

At the moment not many databases for FMS exist and some of the data they contain is inconsistent. There is a lack in consistency of the assessment tools, deviation in the reference groups and studies with outdated information because the research has been done more than a decade ago. Furthermore, the diversity of medical systems and different political environment in each country needs to be taken into consideration as well.[6]

In most cases the treatment of FMS remains difficult for the physician, and the patients tend to not be satisfied with the outcome. A novel therapeutic concept which could be used for blocking comorbidity is called **altering course** of FMS. In order to apply approaches to alter the course of FMS it is important to perform patient stratification[5]. Patients could be grouped by separating them based on measures for psychological and symptomatic symptoms. Subgroup classifications based on empirical studies are essential so as to provide more specific therapies.[8] Even though there have been prior attempts in order for classification of patients, there are no clinical variables or clear patient subgroups defined yet.[9]

A proper categorization of a disease into different subgroups requires the existence of large database containing sufficient patient data. Such a **database does not exist** at the moment.

1.2.1.Research objectives

A research consortium of different research groups from the Universitat Politècnica de Catalunya (UPC), Spain, Hospital de Terrasa (CST), Spain, the University Hospital of Riga Stradins University (RSU), Latvia and the Universitätsklinikum in Essen (UK Essen), Germany aims to apply a novel approach of altering the course of FMS by pooling the resources of an interdisciplinary collaboration.

The approach of altering the course of FMS aims to mitigate the effects of psychological symptoms and mental disorder comorbidities related to FMS or to

[5] Stratification in a medical research context describes the grouping of patients by a factor, which is different from the current treatment given.

even completely avoid its development. In order to establish this new approach several goals of the project are defined:

1. Quantitative identification of the relationship of psychological symptoms and mental comorbidities with FMS.
2. Utilization of a mathematical model in order to describe the appearance of psychological symptoms, as well as mental disorder comorbidities with FMS.
3. Utilization of quantitative criteria, validated by a database containing measurements of somatic and mental health, to stratify FMS patients.
4. Creation of an anonymized FMS patient database containing variables about the following data types: somatic symptoms, sociodemographics, lifestyle, treatment, data about psychological and mental health, genetics, neuroimaging and neurochemistry.
5. The long-term goal is the provision of individually tailored and effective therapeutic or preventive approaches to reduce the functional impact and improve the mental health of patients suffering from FMS.

The mental disorders and psychological comorbidities with FMS in addition to its insufficiently understood pathogenesis make the prevention and treatment of FMS difficult and problematic.[10] Hence, to reach a better understanding of the pathogenesis and alter the course of FMS, new and innovative methods benefitting from interdisciplinary cooperation, are necessary.

1.2.2.Monte Carlo Analysis

The gathered patient variables in the database will be analyzed regarding their relevance for the inception of FMS and probability distributions will be generated for each variable. A comparison of the derived probability distributions shall be done with relevant literature. Moreover, probability distributions for the previously selected variables will be gathered for a reference group not suffering from FMS. The Monte Carlo Analysis method will be used to create simulated individuals utilizing both probability distributions. In order to describe the impact of each analyzed variable on mental disorder comorbidities, a mathematical model will be proposed. Therefore, it will be possible for each of the simulated individuals to predict an impact score the patient will have for each mental disorder comorbidity.

In addition, the mathematical model shall be used to stratify the patients into different subgroups.

Comparing the statistical distribution for mental disorder comorbidities of the simulated patients against the real patients gathered in the database will perform a validation of the model. [10]

1.3.Motivation to create an FMS database

As of today, no comprehensive database for FMS patient data exists. Researches mostly can only use their individually created and mostly relatively small databases. In order to classify and categorize patients into subgroups for comorbidity studies a much larger and more comprehensive database is needed.

Further the grouping of FMS patients with such a novel method shows potential to be useful for other heterogeneous and chronic psychiatric conditions, such as bipolar disorder, schizophrenia and autism, where the prediction of situations in transition (like the chance to develop a mental disorder) and where development, environmental, genetic and epigenetic elements interact with each other.

Economic benefit

Furthermore, the results of the project, with the aim of altering the course of FMS, could be useful for the scientific community by providing a method, which is able to classify the characteristics underlying the sensitivity for the development of mental disorder comorbidities and psychological symptoms. Moreover, physicians should also be able to benefit by utilizing it as a basis for an individual medical treatment. It should also help in terms of obtaining more knowledge about how the disease runs it course, in terms of mental disorder risks and psychological symptoms, and hence, have a possible positive impact on the direct and indirect costs of FMS.

Benefit for the patient

First of all, through altering the course of the disease by mitigating the effects of psychological symptoms and mental disorder comorbidities related to FMS, or, in the best case scenario, completely preventing the development, the patients quality of life could be dramatically increased.

In addition, another advantage is a system implementation, which is focused more on the patient. This element can help the patient to become more confident and stronger in dealing with a chronic condition that entails heavy economic and social consequences.

Finally, it helps the physician to create a more personalized treatment, which can be adapted to the course of the patient’s disease.

2. Current status of medical database technologies and services

This chapter gives an overview about already existing FMS patient databases and different medical database technology concepts, which could potentially be used for creating a medical patient database. There are already specific database technologies for collecting and storing medical patients' data in such a way that they comply to data protection regulations. Since the FMS patient database should be created inside the European Union (EU) it seems reasonable to start analyzing medical projects and databases in the EU because they have to comply with the corresponding data and privacy protection regulations.

In addition, open-source software platforms appear to be more beneficial not only because of economic reasons but also because they offer more flexibility in the design and additional creation of software modules. Further, when using proprietary database software one becomes dependent on third party software and also might be susceptible to changes in the terms of service.

In the following, a range of database software and medical patient database projects will be described.

2.1.Already existing FMS databases

There are already a few FMS patient databases in place in the Hospital de Terrasa (CST), Spain and the University Hospital of Riga Stradins University (RSU), Latvia. At the moment these databases are very small. But they potentially could be used as a starting point to acquire an overview of the type of data that is stored for FMS patients.

The CST database contains 150 FMS patients, of which 50 patients are in the control group. The database contains socio-demographic information (age, sex, education, premorbid intelligence and cognitive reserve), clinical and functional information (quality of life, sleep quality, emotional symptoms, physical impairment, psychosocial functional capacity, pain intensity), subjective cognitive complaints (memory and executive complain), cognitive assessment (attention and executive function), other psychological characteristics (stressful life events and personality traits) and pharmacological treatment as well as physical activity. [10]

The RSU database contains information of 50 FMS patients that has been used to study neurological indicators, such as thermal quantitative testing, pain and other typical symptoms in FMS. The whole database test group had issues with

their memory recalling and over 80% of them where suffering from anxiety and depression. [10]

2.2.Medical database technologies

In the following section an overview of medical database technology concepts of *Integrating Biology and the Bedside,* Shared Health Research Information Network and the European Society for Immunodeficiencies are given. The focus hereby is on the technology used to create the databases.

2.2.1.Integrating Biology and the Bedside (i2b2)

Integrating Biology and the Bedside (from hereon called i2b2) is a National Center for Biomedical Computing located at Partners HealthCare System in Boston, Massachusetts and is funded by the NIH[6]. It has been established in 2004 according to a Roadmap Initiative of the NIH.

The i2b2 provides an open source scalable informatics platform for clinical researchers in the field of human health. One of the i2b2 core fields is the development and testing of new patterns in computation and methodologies for several diseases (type 2 diabetes mellitus, Huntington's Diseases, airways disease, rheumatoid arthritis, multiple sclerosis, inflammatory bowels disease, major depressive disorder).

Furthermore, the i2b2 Center supports a group of more than 250 academic users and "sponsors annual shared tasks for challenges in Natural Language Processing for Clinical Data". [11]

Next to the database software framework the i2b2 also provides a cooperative organizational infrastructure to help researchers in general to enhance the overall results and insights from their clinical studies. It provides a set of software packages to collect data but also to manage a clinical research project and run analysis with the data.

Since i2b2 is an open source project the whole source code is available on GitHub (https://github.com/i2b2) and also, a development contribution to the project is possible.

Figure 21 gives an overview about the available software modules from the i2b2 framework. The i2b2 modules are structured into five different main groups of cells: The i2b2 Core Cell, the i2b2 Optional Cell, Workbench/Plug-in, Web Client and CRC-Plugin. A detailed explanation for each module as well as the i2b2

---

[6] NIH stands for National Institutes of Health. It is the major United States government agency for research in biomedical and public health.

Workbench Client Software and complete Client & Server software packages are available on the i2b2 Foundation homepage[7].

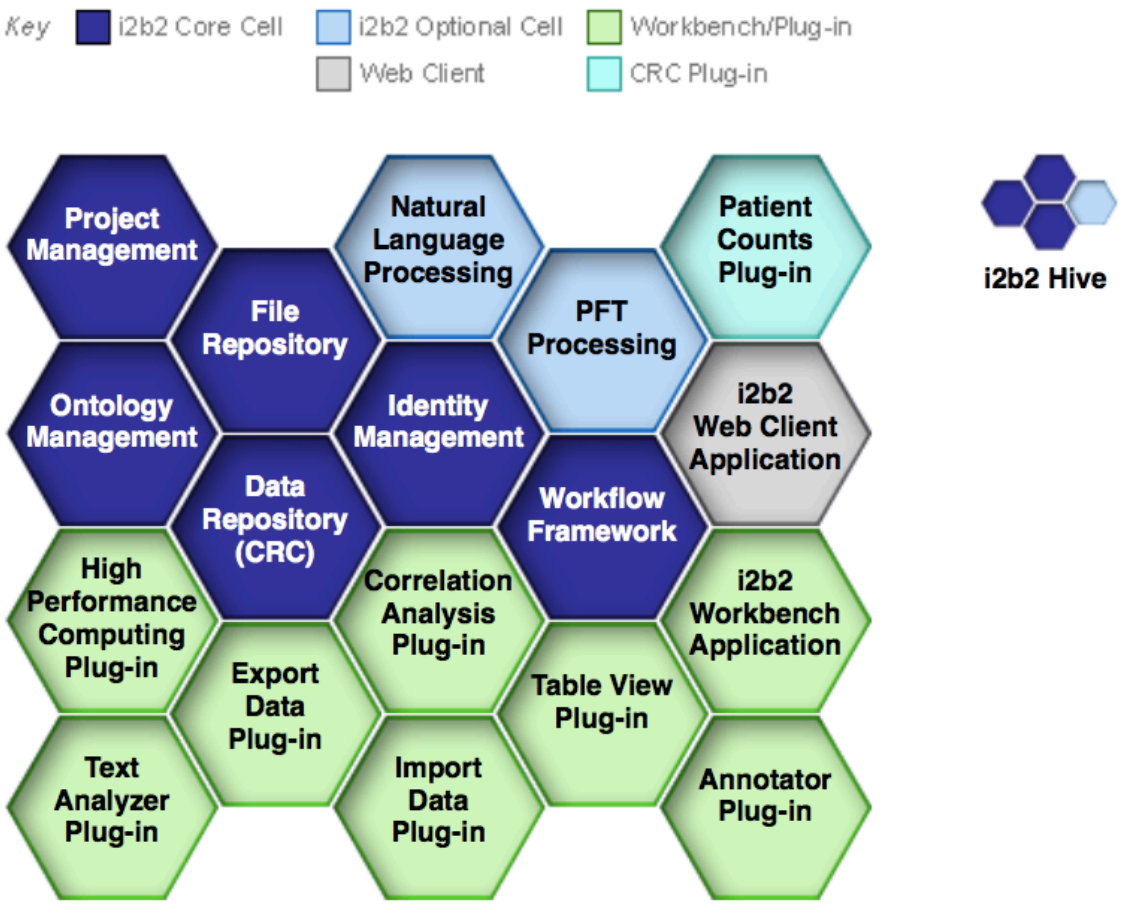


Figure 21: i2b2 available modules.

The whole cluster of software modules (i2b2 “cells”) is also referred to as the i2b2 “hive”. It has some core components like the “i2b2 Core Cells”, which provide the main database core running on the server side. But the i2b2 hive is an extensible software architecture to which developers can add new cells (software modules), which functionalities are provided through plug-ins for the i2b2 Workbench and i2b2 Web interface. The communication between the Workbench and cells is realized through a web-service based on XML messages. This setup of the software architecture makes i2b2 very customizable and flexible for various kinds of clinical research projects.

The Workbench is a compilation of the available client-side modules in form of Java plug-ins which can communicate with the other i2b2 cells and gives the user more possibilities in terms of analyzing, querying and displaying the data of the i2b2 hive in comparison to the web client. The Workbench is available as an executable for macOS or Windows operating system.[11]

[7] https://www.i2b2.org/software/index.html , accessed August 7th, 2018

The i2b2 user interface

In the i2b2 web interface (Figure 22) the user can drag terms from an ontology[8] tree (top left) into a query tool (top right) where they can be used to create Boolean expressions to find the desired patients. The query tool is split into three panels, which are logically AND'ed with each other (logical conjunction). The items, which can be dragged into each panel, are logically OR'ed (logical disjunction) and can be negated by clicking onto the "exclude button". This gives the user the possibility of creating complex Boolean algorithms to filter for the desired patient data. Furthermore, a minimum number of occurrences can be specified in a panel and a time range can be given. In the "previous queries" window (bottom left) results from earlier queries are saved and can be dragged into one of the panels to combine it with the current search.[11]

All in all, this provides a very powerful tool to make detailed search requests even for users, which do not have any experience of data base programming and SQL queries.

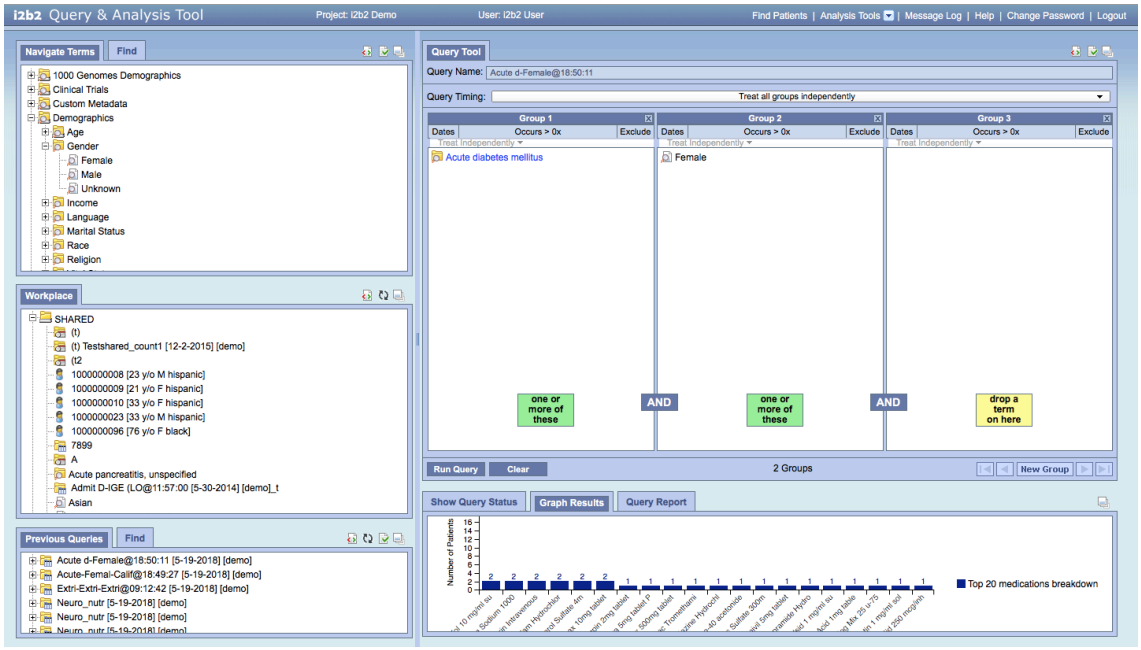


Figure 22: i2b2 web interface.

A very detailed user guide for the web client and desktop client can be found on the i2b2-community-website[9].

[8] In this context ontology mainly describes the grouping of medical symptoms or facts within a hierarchy. They are subdivided according to differences and similarities.

[9] https://community.i2b2.org/wiki/display/mi2b2/mi2b2+User+Documentation , accessed September 18th, 2018

2.2.2.Shared Health Research Information Network (SHRINE)

The Shared Health Research Information Network (SHRINE) is a general-purpose clinical querying protocol that can be adapted to many different types of data repositories.

The first prototype of SHRINE has been developed in 2008 together with different Institutional Review Boards (IRBs) of the three biggest health centers associated with Harvard University. These health centers allowed the usage of their data and the Harvard Medical School IRB approved to create a Query Aggregator Interface. This interface can send out requests concurrently to each hospital database and get an accumulated result of the number of matching patients.

SHRINE enables collaborative research work across different institutions, which have independent database systems. While each institution maintains their own database with individual patient privacy regulations and access policies, different financial funding plans and own research teams, and so forth, SHRINE allows queries to all the databases in the network through a standardized web interface.

The first SHRINE prototype was setup by Mr. Weber and his team. They faced complications with regulatory, political and technical issues during this process. One can assume that such issues might occur again while setting up another SHRINE system, hence the previous mentioned work could be used as a guideline, which highlights problems and their possible solutions.

SHRINE was created with the success of i2b2 and SPIN[10] in mind. It should allow researchers to analyze electronic health records of patients across independent platforms (initially it was the three participating Harvard hospitals for the prototype project).

Due to the limited timeframe of only 6 months for the prototype project some exclusions had to be made:

- Only the three largest databases of Harvard's health centers were queried.
- Limitation to only aggregate queries and a limited user group.
- The query interface only allowed searching for the two parameters patients' demographics and diagnosis.
- The technical architecture of the prototype has to be redesigned to make it scalable for utilization with a large number of health centers.

[10] The Shared Pathology Information Network (short SPIN) provides a solution for sharing data between institutions over a peer-to-peer network while letting them keep autonomy over their data and protect the patients' data privacy.

SHRINE System

The user interface for SHRINE is a web-based user interface very similar to the one of i2b2 (see Figure 23). Like the i2b2 interface the user can drag terms from an ontology tree (top left) into a query tool (top right) where they can be used to create Boolean expressions, as in the i2b2 interface.

The query status window (in the prototype) shows the execution time of the query and the number of matching patients at each of the three hospitals, which have been part of the SHRINE prototype project.

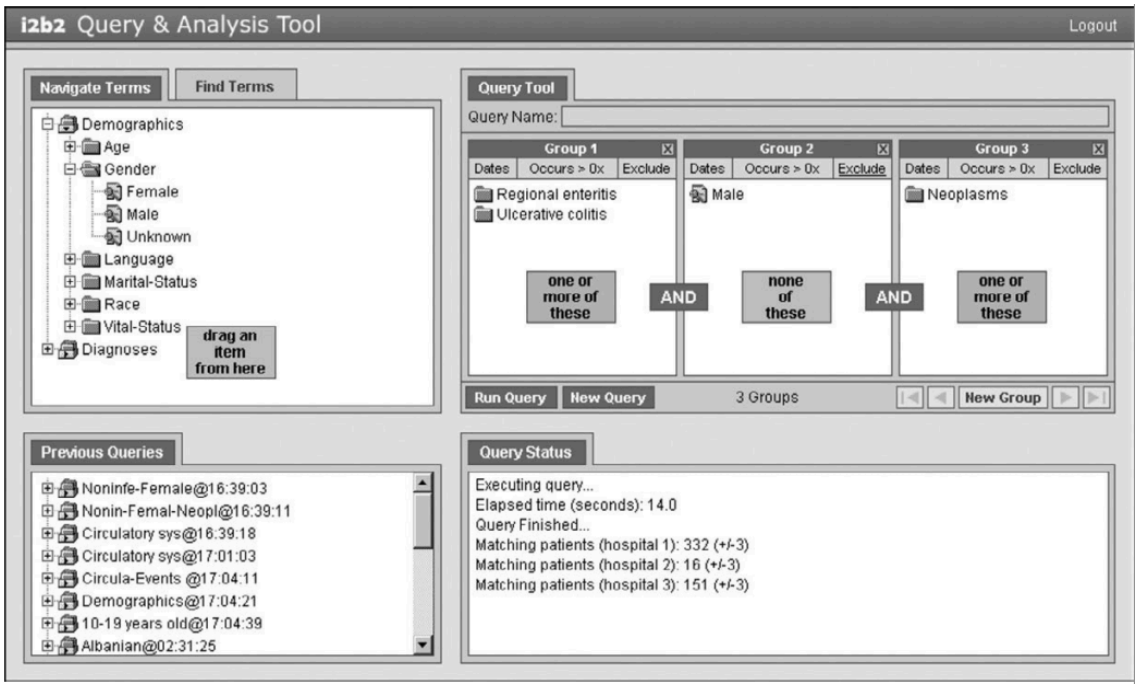


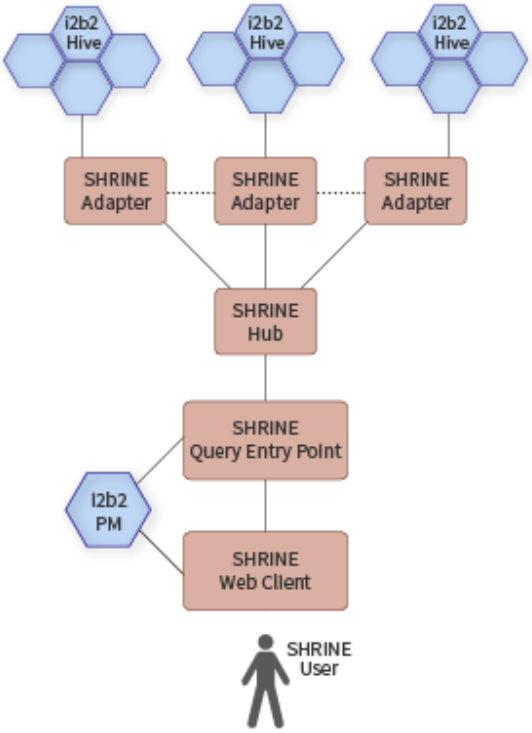
Figure 23: SHRINE User Interface, similar to the web interface of i2b2.

SHRINE architecture

The main architecture of the prototype SHRINE is constructed by a Query Aggregator (hosted on servers of the Harvard Medical School) and SHRINE Adapters placed at every hospital.[12]

The aggregator consists of the frontend web interface for the user access, the Query Entry Point (QEP) that is the point of entry for the SHRINE network and of the HUB. The QEP authenticates all incoming requests and later forwards them to the HUB for broadcasting. Furthermore, when it receives cumulative results back from the HUB it aggregates them into a single message for the client (web or desktop client). The high-level architecture of SHRINE is shown in Figure 25.

The HUB broadcasts the requests from the QEP to the adapters and buffers the adapters responses before sending the collected results back to the QEP.



itecture of

The SHRINE adapter is the point of entry for the database servers of the institutions. It converts the query into a format that is compatible with the format of the institution. In theory there is no limitation on how many databases can be added to the SHRINE network. Even though the adapters need to accept the standardized XML query format, the adapters backend can be modified without restrictions in order to connect it to a health centers database server utilizing a different technology. This gives flexibility of connecting different database technologies to the SHRINE network.

2.2.3.European Society for Immunodeficiencies (ESID)

ESID is a database for primary immunodeficiencies (PID). PID are rare diseases, therefore a cooperation between health institutions, also over national borders, is important to gather enough information for medical studies. ESID was originally founded in 1994. In 2004 a web-based ESID Online Database was developed by the Center for web-based Research and Patient Databases (CwebRD), a core facility of the IT Center of the University Hospital Freiburg, Germany.[13]

The ESID Online Database is designed for data entry, reporting, data storage and the import of already existing data sources. All these features are available via a web interface (as shown in Figure 25), which is secured with login and password credentials. For a party to participate in the project an agreement must be signed with the ESID and an application has to be made to receive the login credentials. Without specific permissions each user only has access to its own patient data sets. To receive access for another party's data set permission must be requested. Furthermore, if patient information is shared it can only be done in a de-identified manner.[14]

In summary the database is custom build of components by the Java Platform family such as the Java Platform Enterprise Edition (J2EE) and the Java Database Connectivity interface JDBC™ for database access. The data is stored within a MaxDB™ SQL relational database management system.[13]

Another central component is called the Enterprise Integration and Development Platform Application (EIDP). It is supposed to help with data protection, integrating solutions for clinical databases and to structure research

networks. The system is developed by the company Toolwerk GmbH. [13] Even though Toolwerk GmbH claim to be open source no access to the source code is give and it also seems that the ESID database has been their only project. [15]

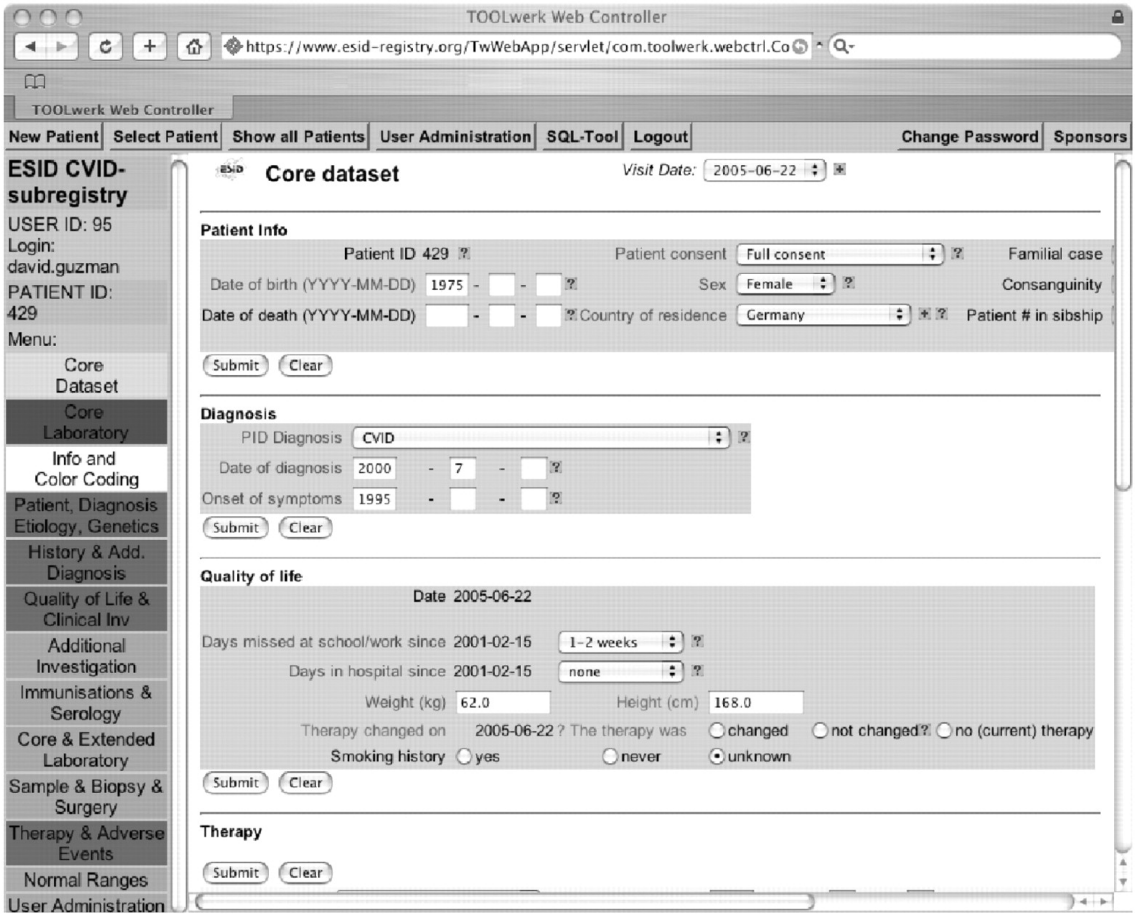


Figure 25: The ESID web platform for database access and patient entry.

2.3.Medical Patient Database Projects

In Europe as well as the United States, several projects and feasibility studies dealing with the topic of sharing and re-using Electronic Health Records (from hereon called EHRs) have already been conducted. Their main objective is to foster research and make existing EHR data available to a broader user group.

This chapter will give an overview about the projects, which deemed most promising to be useful for the goal of this master thesis, on how to tackle the technical and possible economic challenges of creating a medical database. The Electronic Health Records for Clinical Research project, the Strategic Health IT Advanced Research Project and a German feasibility study for the usage of i2b2 will be explained briefly and their results analyzed.

2.3.1. Electronic Health Records for Clinical Research (EHR4CR)

The Electronic Health Records for Clinical Research (EHR4CR) project was conducted over 4 years (2011 – 2014) with the goal of fostering clinical research on a European level by analyzing the possible re-use of already existing Electronic Health Records (EHR) at distributed locations. A total of 33 partners, coming from academia, pharmaceutical industry, as well as mid-size and small businesses, contributed to the project.

During the EHR4CR project, a protocol feasibility platform was developed in order to send uniform query requests to databases across national borders. Furthermore, the feasibility protocol was evaluated in terms of fulfillment of necessary technical and governance requirements for a successful execution.

The developed EHR4CR platform consists of three parts:

- **Workbench**: Used to create the query criteria via drag-and-drop with a graphical user-interface based on a comprehensive query model.[16]
- **Orchestrator**: Its function can be seen as the broker for the queries. It receives the queries from the workbench and knows where they should be redirected. It buffers the queries until the endpoint sends a request.
- **Endpoint**: The connection point to the database. It polls for new queries on a regular basis and executes them to aggregate patients out of the anonymized patients' data. Presently two types of Clinical Data Warehouses (CDW) are supported, the i2b2 and an own EHR4CR schema warehouse.

Outcome

Queries have been sent to different countries within the EU, therefore it was important to obey the data protection regulations in each country. Due to the right of 'informational self-determination' in Germany for example, the burden for re-use of patient data is substantially higher there. On the contrary, in the United Kingdom the creation of databases is quite easier, and the Clinical Practice Research Datalink represents the biggest collection of anonymized primary care patient data in the world, covering about 9% of the UK population, which makes up about 5.5 million people.[17]

Despite all the obstacles, the EHR4CR project presented successfully that it is possible to use one software module, after some necessary modifications, in different hospital networks (in different countries) to collect aggregated information. It was possible to send one request to be executed at 11 different locations in 5 different countries.

Next to the technical feasibility study the project also displayed that the **legal obstacles** can be tackled successfully but also **should not be underestimated** in terms of time management.

Finally, the EHR4CR project was conducted with a business model in mind of providing large sets of EHR data to medical research project in order to help them meeting their project deadlines in exchange for a usage fee. The EHR4CR project analyzed that most research projects in Europe are delayed because they cannot recruit or gather enough EHR data in their specified deadline, which cause major economic losses for them.

2.3.2.Strategic Health IT Advanced Research Project (SHARP)

The Strategic Health IT Advanced Research Project (SHARP) was established in 2010 by the Office of the National Coordinator for Health Information Technology, which belongs to the United States Department of Health and Human Services.

It was created with the following objectives in mind:

1. Receive data of EHR in different source formats.
2. Process written text to create structured data.
3. Utilization of common clinical models to create normalized EHR data.
4. The data should be accessed and classified by a algorithm for the identification of cohort groups for a specific research purpose.

Generally speaking it should help to standardize EHRs from different health care organizations and providers for secondary usage (e.g. by research institutes).[18]

Comparing the SHARP with the EHR4CR project they are similar in the sense that both try to reuse EHR data. Whereas SHARP aims to store the anonymized patient data in a central location in contrast to EHR4CR that is storing the data only in distributed CDWs. Furthermore, SHARP applies Natural Language Processing (NPL) to extract information from clinical free text and create structured data whereas EHR4CR only relies on structured data as input.

2.3.3.The German i2b2 experience

A consortium of centers for medical information and medical university centers in Germany conducted a practical feasibility study about the usability of i2b2 to perform networked medical research in 2010. The general objective was to unlock existing patient information, which is collected in different databases on a routinely basis during primary care, for scientific research.[19]

During the study, four different already existing medical databases were used for integration into an i2b2 database. The process & workflow of data migration into i2b2, the applicability and the performance of the i2b2 database were analyzed. The integrated databases had the following four database technology types:

- A CDW by Cognos BI™, IBM, at Erlangen University Hospital. The data has been converted from a star-based scheme into the ontology format of i2b2.
- A documentation about prostate cancer in the Electronic Medical Record system by Soarian™ , Siemens Inc. at the Erlangen University Hospital.
- A local electronic data capturing system, which metadata has been transferred manually into a Microsoft Excel™ spreadsheet and later per script-based conversion into the ontology format of i2b2.
- A trial database (Competence Network for Congenital Heart Defects) using the electronic data capturing tool SecuTrial™ (iAS GmbH). Data has been exported and transformed into SQL statements with a specifically developed tool (Java), which are compatible for the import into i2b2.

For de-identification of patient data, if necessary, the pseudonymization service for medical research data[20], [21] provided by the German Technology and Method Platform for Networked Medical Research (TMF) has been used.

Furthermore, a tool for the setup and administration has been developed to make the installation of i2b2 onto a virtual machine from source code faster and simpler as well as to help with the whole process of setting up and loading an i2b2 database. The tool has been published on the TMF website and the i2b2 Academic User Group[11] to make it available to be used by the public. Further research yielded that the original setup and administration tool has been developed further and was integrated into a project called the Integrated Data Repository Toolkit (IDRT).[22] IDRT entails an i2b2 installation wizard, a GUI based IDRT-Import-Tool to start the ETL[12] process and a GUI based i2b2-Ontology-Editor to implement own terminologies. The IDRT-Import-Tool allows the import of data from standard formats, such as CSV, CDISC ODM[13] and SQL databases, into i2b2.

Since 2015 IDRT is available in the English language as well and the installation wizard can be downloaded on the IDRT homepage[14].

A more detailed description of the used hardware and software tools can be found in the case report of the study. [19]

Conclusion

[11] https://community.i2b2.org/wiki/display/i2b2wizard/Home , accessed June 20th, 2018

[12] ETL stands for Extraction, Transformation and Loading. First information is extracted from a data source, after the data is transformed into the data format of the target system (e.g. i2b2) and later loaded into the target database.

[13] CDISC ODM stands for Clinical Data Interchange Standards Consortium Operational Data Model and provides a platform-independent global standard for medical data.

[14] http://idrt.imise.uni-leipzig.de/IDRT-II/#Contact , accessed June 29th, 2018

The study demonstrated that i2b2 is a feasible tool to enhance networked medical research. Implementation of the TMF privacy related tools was possible without problems or reduced performance. Therefore, all privacy guidelines could be fulfilled for the usage of i2b2.

Compared to native SQL queries i2b2 had a lower reaction time but offers advantages in terms of an easy to use GUI to create queries for medical users without detailed knowledge in database structures.

2.4.Summary

The presented projects and database technologies demonstrated that a successful setup of a medical databases for EHRs is a viable task. Technical as well as legal obstacles can be solved. Furthermore, the EHR4CR project also gave a guideline in terms of how to build a business concept for such a medical database.

The extensive literature research showed that a common denominator in a lot of projects within the EU is the support or usage of the database system i2b2. The aforementioned TMF fosters the dissemination of i2b2 and is cooperating with the i2b2 community.

Furthermore, next to the technical challenges the legal obstacles in terms of privacy regulations that can be different in each country should also be taken into consideration because they are also very important in order to setup a database with sensitive patient information.

3. Technical feasibility

In this chapter a technical feasibility analysis of creating a database for FMS patient data is performed. As a guideline regarding the technical requirements for such a database the aforementioned research consortium with the aim of altering the course of FMS and ideas from EHR4CR will be applied to the use case of a FMS patient database.

First the technical **system requirements** for the input and collection process of patient data as well as the requirements to the system for data access and utilization of analysis methods and tools in the field of FMS are be laid out.

Based on the defined system requirements an **initial project proposal** will be given, utilizing the available database technologies mentioned in the previous chapter. The practicability with existing software and the necessity of additional development to reach the system requirements targets are analyzed.

3.1.System Requirements

In order to be useful for FMS research, the database system needs to fulfill and support a data collection process that is feasible for medical practitioners in primary care. Hence, it should be easy to use for the patient as well as for the medical practitioner. Further, typical analysis methods in the field of the FMS research should be supported to create additional value.

3.1.1.Data acquisition requirements

According to the aforementioned research consortium proposal [10] with the aim of altering the course of FMS a group of 80 patients should be used in order to create an improved mathematical model with a Monte Carlo Analysis. These 80 patients as well as the already existing patient data at CST (150) and RSU (50) should be integrated into a database and used for validation of the previously created mathematical model.

The database system should follow the international health data standard HL7 Fast Healthcare Interoperability Resources (FHIR). This standard defines the exchange of EHR in terms of an Application Programming Interface (API) as well as data format and elements. Furthermore, the database should be accessible for software to perform and apply the Monte Carlo mathematical model.

The data entry of EHRs should be done via a mobile application. Patient data needs to be saved anonymous and according to the data protection regulations in the participating countries and the EU. Data entry should be possible in two ways:

- Via the patient by answering a questionnaire in order to generate the psychological variables for the database.
- Via the clinical practitioner through a special interface to enter patient data.

Furthermore, a mathematical mode should be integrated into the database in such a way that it fills in missing patient variables that could not have been recorded during the survey or by the doctor. This mathematical model should be derived from a Monte Carlo Analysis on a verified patient set.
This is because each practitioner can only assess aspects corresponding to his field. By adding simulated variable entries, every patient will be completely defined. Therefore, most patients will consist of a combination of real and simulated variables. Simulated variables should be clearly marked and assigned with an uncertainty rating.

In Figure 31 a schematic flowchart of the desired patient data input process into the i2b2 database is shown.

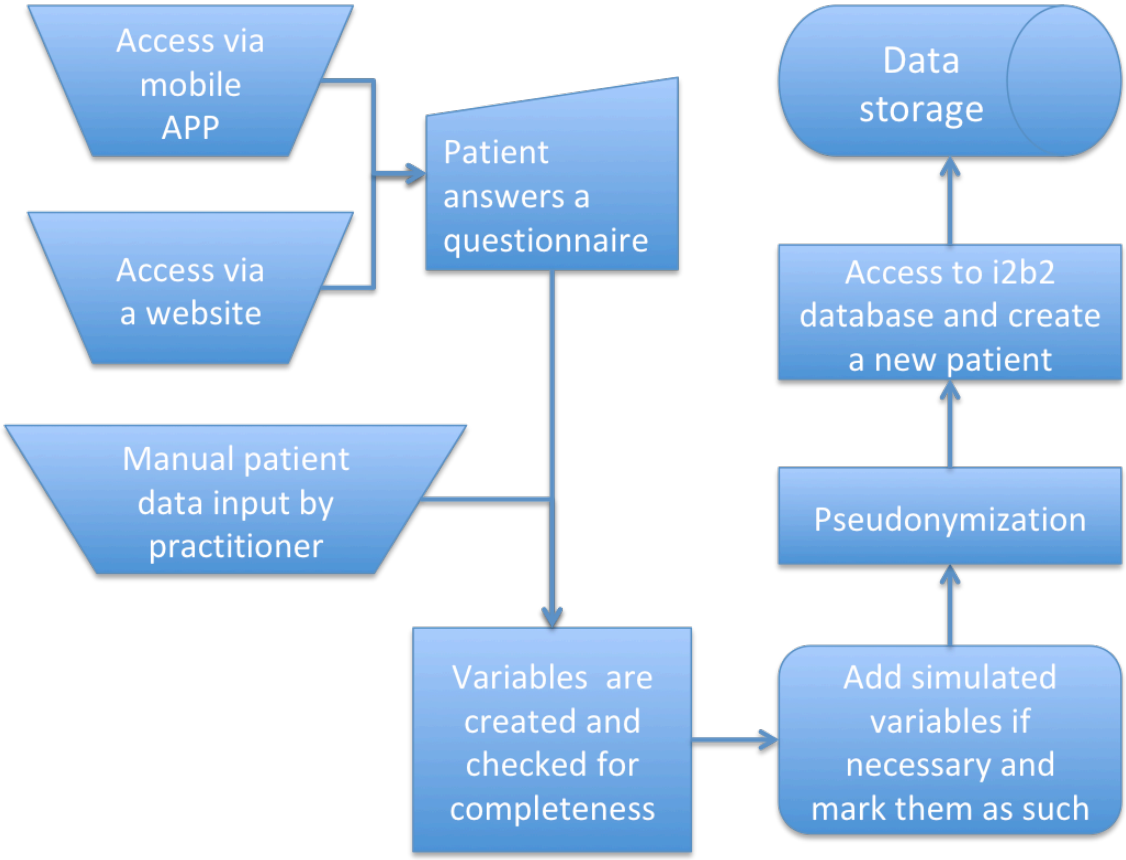


Figure 31: Required patient input process to the database.

3.1.2.Data analysis requirements for the database

The data analysis goal for the FMS patient data is to stratify the patients into different subgroups depending on their symptoms evaluated in the database.

A study conducted in 2008 [23] used a clinical tool, called the Fibromyalgia Impact Questionnaire (FIQ), to stratify patients into two subgroups, because these groups were fitting their data profile the best:

- FM-Type I: Patients having the lowest levels of anxiety, depressive and morning tiredness symptoms.
- FM-Type II: Patients characterized by higher levels of fatigue, morning tiredness, anxiety, etc.

The FMS is a heterogeneous disease with a variation of symptoms and their intensity for each patient. According to this study the FIQ represents a good tool to assess a high number of clinical characteristics related to the FMS. The study has been conducted on two different days separated by two weeks between each other.

The first day was a one-hour interview to obtain demographic information, identification of FMS characteristics and principal complaints, etc.

The second day was used to evaluate the patients cold pain sensitivity and to field out several questionnaires.

By using this study as an example for the analysis and stratification of FMS patients one can derive a set of requirements the database has to fulfill: Either the database itself has to contain the corresponding analysis tools or it needs to give the ability to export the data for utilization of further analysis tools, which are necessary. Therefore, this chapter will also analyze the capability of i2b2 for patient data analysis and stratification.

3.2.Project Proposal with i2b2

In this section, a project proposal and its feasibility analysis on a technical level for the utilization of the i2b2 software to create a database of EHRs for FMS patients will be done. In the previous section, mentioned system requirements will be taken into consideration. Since i2b2 already offers a great variety of available software modules for clinical research it will be analyzed to what extent “out-of-the-box” software modules can be used to fulfill the requirements and which additional software modules have to be developed in order to meet the project requirements.

3.2.1.System Architecture

The primary proposal is to create an i2b2 database server at the locations already having a set of FMS patients, such as CST and RSU. The i2b2 server can be installed on a virtual machine with a Linux distribution supported by the i2b2 Wizard (recommended is Ubuntu 14.04 (32 bit)). The i2b2 Wizard, of the IDRT, offers a simplified way to install the i2b2 server. The different server

locations should be connected utilizing SHRINE to give all participants access to the complete EHRs.

Pros	Cons
An **advantage** of this approach would be that each participating entity keeps the property rights and the data in its own designated location. This could potentially simplify obstacles of different data privacy regulations of different countries within the EU. In addition, a distributed architecture can help to distribute the server load for a growing database in the future.	As a **disadvantage** could be seen that this approach has a more complex technical architecture and that it could be more expensive in terms of maintenance and additional hardware that is needed to establish several server locations.

An alternative proposal is to create an i2b2 database server at one central location that has to be chosen beforehand and to which all research participants have access to and are also willing to save the patient data at.

Pros	Cons
Compared to the first proposal an **advantage** could be a simpler and more cost-effective solution in terms of maintenance and hardware, because less servers and no SHRINE architecture are needed.	A **disadvantage** could be more possible obstacles regarding data protection regulations and less scalability for a growing patient database.

From a technical and financial perspective, a central server solutions has more advantages. Hence, it seems more reasonable to proceed with this approach. In addition SHRINE networks (in the USA and Europe) so far only have been used within national borders[24]. The previously described ESID project can be used as an example for a centralized database. Their servers are located at the IT department of the University Hospital in Freiburg, Germany.

Last but not least, an expansion with SHRINE can be realized at a later time after more experience has been gained and the database would start to grow at a fast pace.

3.2.2.Database Input

The initial content for the database can come from the integration of the patient data at CST and RSU. The data needs to be analyzed and converted into the

ontology format of i2b2. Several ETL tools, such as the previously mentioned IDRT-Import-Tool, are already available to help with the automation of the import process. Furthermore, i2b2 as well as SHARP offer natural language processing functionalities. This can be beneficial for the integration of EHR or questionnaires that are not available in a structured data format or only exist in analogue. Some of the databases, which have been mentioned in the previous chapter, successfully managed a complete import of other databases and different data formats into i2b2.[19]

The main source of FMS patient data should be the ongoing input of data during primary care. Ideally through answering of questionnaires from the patient itself or through the medical practitioner via a special interface. Therefore, an interface for the practitioner and the patient needs to be created.

The first step is the creation of a web-interface with a mobile version to make it accessible for tablets and smartphones. The next step is the creation of a mobile App to function as the database interface.

The patient interface needs to guide the patient through the answering process of the questionnaire and generate data variables from the answers. After that the patient profile will be analyzed and for missing variables simulated data will be derived from the Monte Carlo Mathematical Model. Simulated data has to be marked as such. Further, the patient data has to be pseudonymized according to EU legal privacy standards before imported to the database. For this purpose the pseudonymization service for medical research data[20], [21] provided by TMF can be used. The compliance of this service with the new GDPR regulation (which are applicable since May 25th 2018) needs to be checked.

Furthermore, the interface for the medical practitioner needs to allow him or her to create new patient and to modify the patient variables directly.

A straightforward approach to create a data input interface would be to tap into already existing Electronic Data Capture (EDC) tools. Such an EDC system, which has a worldwide user community in medical research is REDCap (Research Electronic Data Capture). The Vanderbilt University has done the original development and rollout. A big international community of non-profit organizations now supports it. The REDCap software is **free** of charge **for non-commercial** research purposes.[25]

REDCap User Interface

The REDCap software can be used to create online surveys and databases without having extensive technical knowledge. In REDCap the user can create projects. A project is a website used for the entry of data. The tool is suitable for almost any type of data and a multitude of purposes, because the user can create each survey for the data entry individually and therefore REDCap offers a lot of flexibility.

Under the “Project Setup” Tab the user can build his or her own survey that can contain the following formats:

- Text and notes boxes for alphanumeric text, which optionally can use validation to ensure a specific input format, such as a date for a birthday.
- Calculated fields that are automatically filled based on a predefined formula and another data field.
- Multiple choice and checkbox fields to select from a set of answers.
- Branching logic can hide or reveal fields based on previous answers.

A data record (e.g. information about a patient) can be constructed of several input questionnaire forms. For organizational reasons the status of each form can be marked as complete, unverified, incomplete and incomplete (no data saved. This helps to get a quick overview whether information in certain forms are still missing.

Furthermore, REDCap offers an export functionality for a variety of formats: CSV (raw data or with labels), CDISC ODM (XML), SPSS Statistical Software, SAS Statistical Software, R Statistical Software and Stata Statistical Software.

REDCap offers many useful features to support the data collection process throughout a project. The REDCap project homepage has many detailed explanatory videos[15] regarding its functionality and after joining the REDCap consortium one can profit from the support of a big REDCap community.[25]

REDCap Infrastructure Requirements

Regarding computing power, memory and hard drive space REDCap does not have any specific requirements but it recommends 10 GB free disk space on both the web server and the database server. REDCap is supported on different operating systems (Linux, Unix, Windows, Mac). Furthermore a smtp webserver is needed to send emails from REDCap and an optional fileserver for the storage of files uploaded to REDCap.

On the REDCap project homepage, a detailed documentation about the technical overview and requirements is given.[26]

3.2.3.Data Analysis

The primary i2b2 web-client and workbench interface that have been presented in chapter 2.2.1 offer the capability of aggregating a patient cohort group for a further study. This gives the user the possibility to filter patients based on desired symptoms or other available features.

As an example, if a researcher is looking for a patient group that has mental and behavioral disorders and/or a defect with its nervous system and is older than 34 years.

---

[15] https://projectredcap.org/resources/videos/ , accessed August 17th, 2018

To find this patient group, the web client will be utilized: The ontology tree of the i2b2 demo version[16] of the web client has diseases categorized according to ICD10[17], which makes the search easier because it is following a standardized scheme. Within ICD10 diseases and symptoms are categorized by letters and numbers, hence the researcher can select from an alphabetically and numerically ordered list. ICD10 is only an example of a standardized and international classification format in i2b2. A variety of different classifications are already available in the ontology tree of i2b2 and the user can create individual ontologies (e.g. for FMS symptoms). In the case of FMS for example one might want to classify the patient by FM-Type I or FM-Type II with a specifically created ontology for this use case. Such new classifications can be used to perform further analysis or to do more specific grouping of the patients.

In the specific case of the mentioned example the ontologies “Diseases of the nervous system (g00-g99)” and “Mental and behavioral disorders (f01-f99) are selected in the first group (left column). One or more of the selected ontologies must be present in the patient. The second group (middle column) excludes patients with an age between 0 and 34, therefore the selected patients have to be minimum 35 years of age. The third group (right column) is left empty for this query, but additional parameters could be added here. The constructed query is shown in Figure 32.

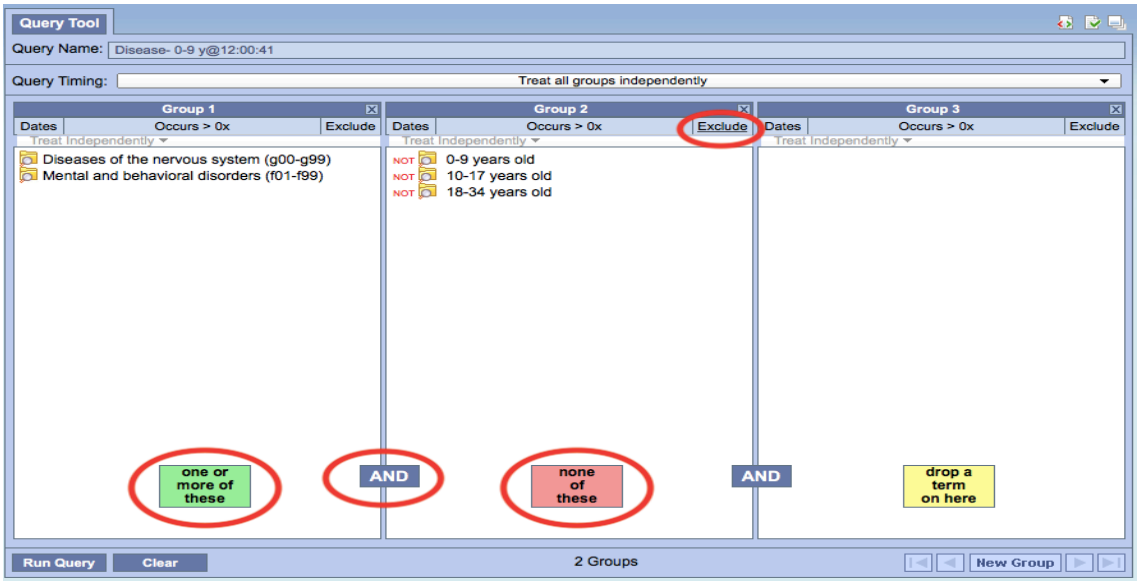


Figure 32: Query for patients with mental diseases and/or mental and behavioral disorders and more than 34 years of age.

After executing the query, the obtained results can be seen in Figure 33. In the i2b2 demo database 74 patients with the specified query parameters were found. Furthermore, we can see an age distribution and the breakdown into the

[16] https://www.i2b2.org/webclient/ , accessed August 5th, 2018

[17] ICD10 stands for International Statistical Classification of Diseases and Related Health Problems and is the most important and worldwide recognized classification system for medical diagnosis published by the WHO. It gets updated on a regular basis.

top 20-diagnosis of the patients. It is possible to choose a variety of additional query results types, such as “Top 20 diagnosis breakdown” or “Length of stay breakdown” just to name a few. A very detailed explanation of the query generator is available directly in the web client under the button “help” in the top right corner. Hence going here into more detail would exceed the scope of this thesis.

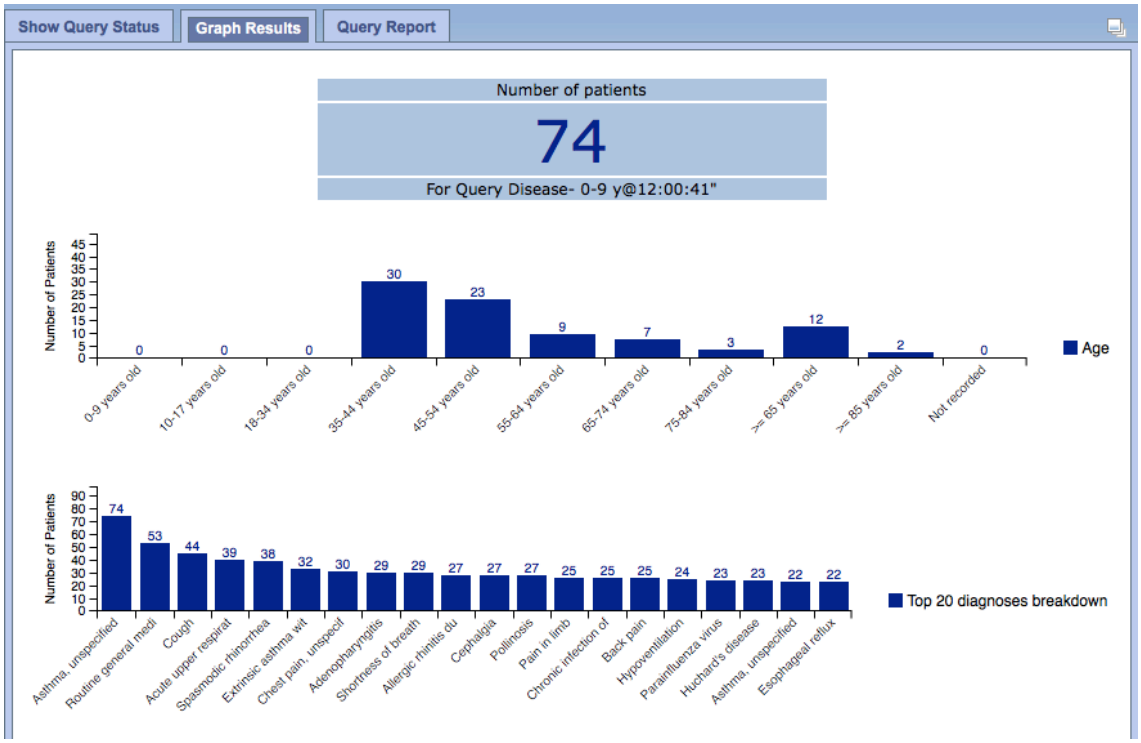


Figure 33: Patient cohort results with age and top 20-diagnosis breakdown.

The standard web-client already offers some analysis tools for simple patient analysis with de-identified data. To obtain more detailed and identified patient level data other existing medical databases, such as the one at the UMass Medical School, implemented a process to create an access request that has to be approved by the Institutional Review Board (IRB).[27]

To enhance the client’s capabilities additional tools can be installed into the web client as plugins. Within the i2b2 community[18] several community projects are available to extend to capability of workbench and web client.

i2b2/tranSMART

Another useful extension is i2b2/tranSMART[19]. It is an application layer for i2b2, which allows the same functionality as i2b2 plus the capability to perform complex statistical analysis and loading data from a simple excel sheet or other different sources into the database.

---

[18] https://community.i2b2.org/wiki/ , accessed August 6th, 2018

[19] http://transmartfoundation.org , accessed July 10th, 2018

The Global Rare Diseases Patient Registry Data Repository by the NIHs National Center for Advancing Translational Sciences (NCATS) in conjunction with the Harvard Medical Schools department for Biomedical Informatics uses i2b2/tranSMART. They structured the access rights into a two-level approach:

- Level 1 access: Only provides aggregated data with no IRB approval required.
- Level 2 access: Provides full patient level data but needs a user with IRB approval.

Furthermore, the Harvard Medical School provides a demo version[20] of i2b2/tranSMART (with level 1 data access) to get acquainted with the user interface and some of its analytic capabilities. This database is created from the National Health and Nutrition Examination Survey (from hereon NHANES) with around 41,000 patients and around 2,000 environmental patient level variables.[28]

The user interface of i2b2/tranSMART is similar to the one of i2b2. As shown in Figure 34 an ontology tree is present on the left side. On the right-side patient subsets can be created.

In the following usage example of i2b2/tranSMART the goal is to check the hypothesis that there is a correlation of patients that are smoking and therefore are more likely to have been diagnosed with cancer. As a starting point two subsets filtering for the diagnosis (cancer) are created. For subset 1 “any cancer”, “blood cancer” and “leukemia” are selected. For subset 2 “lung cancer” and “mouth cancer” are selected.

First one needs to drag the desired terms from the ontology tree on the left into subset 1 and subset 2 on the right side. Equivalent to the i2b2 query creator the terms in one box are logically OR’ed with each other and the boxes logically ANDed (as highlighted in Figure 34). After the desired terms have been selected clicking the “Generate Summary” button can start the patient aggregation process.

[20] https://nhanes.hms.harvard.edu/transmart , accessed July 10th, 2018

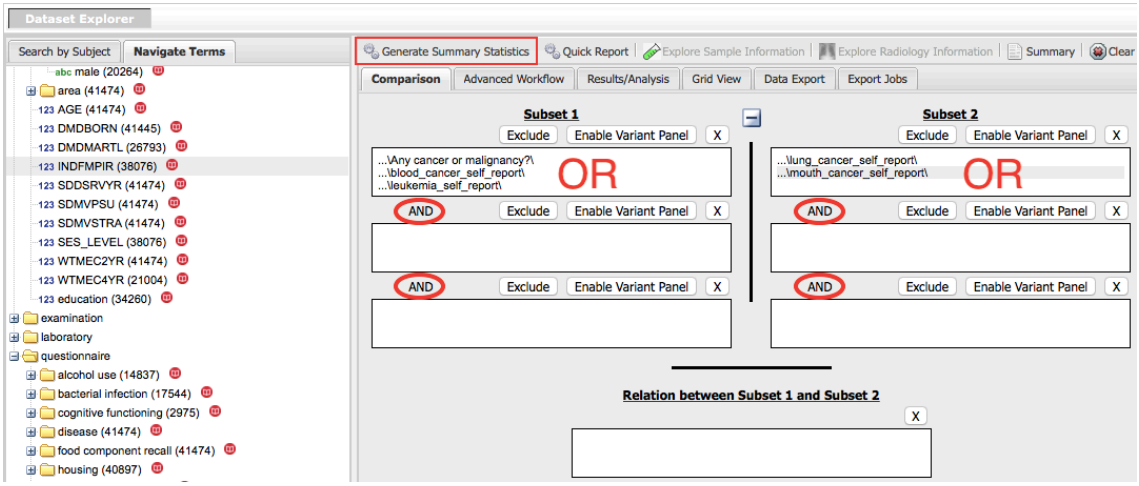


Figure 34: i2b2/TranSMART subset selection.

A summary statistic can be seen in the “Results/Analysis” tab (Figure 35). In the summary a histogram and comparison of the patients age of both subsets, a gender distribution, as well as a race distribution are given.

Figure 35: Results/Analysis tab with Summary Statistics

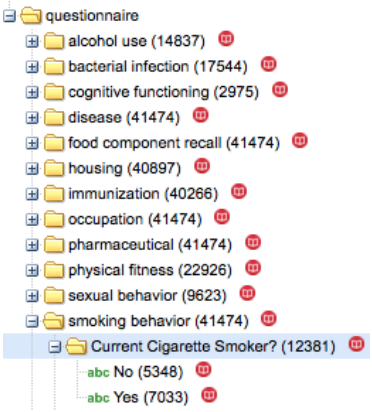


Figure 36: Smoking behavior.

Since the goal is to analyze the association of cancer with the smoking behavior of the patients the term “Current Cigarette Smoker?” from the ontology tree (Figure 36) can be moved per drag & drop into the “Results/Analysis” tab.

After the smoking behavior term has been dropped into the Summary Statistics, the association will be calculated. The result (as shown in Figure 37) is a new bar graph for each subset displaying the association with the current smoking status. One can see that for patients diagnosed with both types of cancer the share of patients that are currently smoking is higher. Furthermore, one can see that the result is marked as significant at a 95% confidence level.

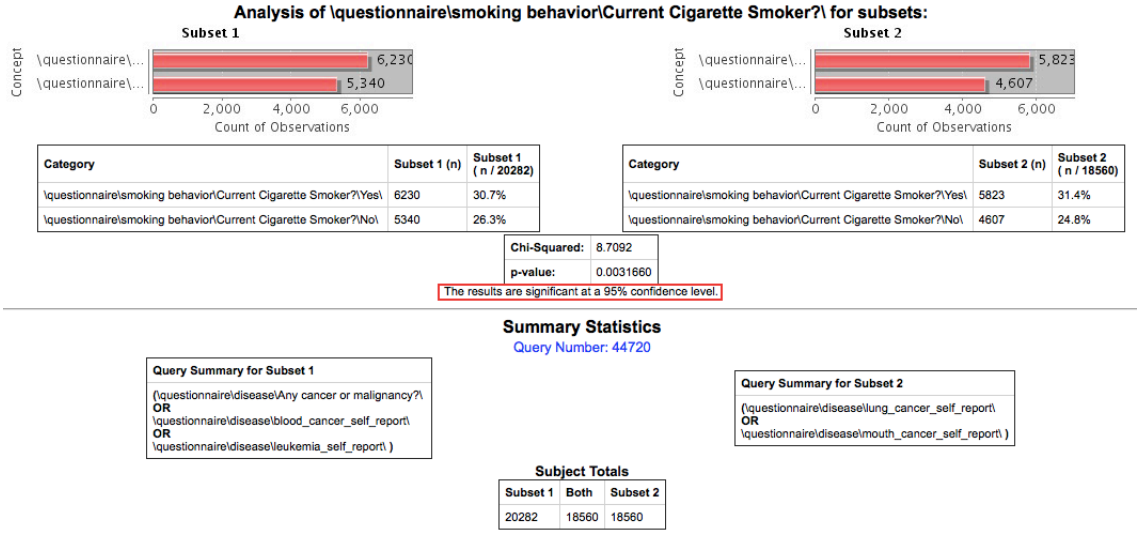


Figure 37: Association of current smoking behavior with cancer diagnosis.

This demonstration is just a simple analysis with a limited demo database and only level-1 access rights. But nevertheless, it demonstrates in a good way how i2b2/tranSMART works and can be used to create a patient group and initial hypothesis for a more detailed analysis.

To perform a deeper analysis more detailed patient level data is necessary (level-2 access in the case of NCATS) to make use of more powerful statistical tools such as the R software tool. The group of Chirag Patel at the Bioinformatics Department at the Harvard Medical School gives a hands-on step-by-step explanation for the utilization of the R software tool for medical analysis of the NHANES patient level data.[29]

Furthermore, there are plugins already available for i2b2, such as GIRI[21] that offer a generic integration of the statistical R software directly into i2b2. GIRIs approach is to make the integration of one or more desired add-on functionalities (also called “scriptlet”) as simple as adding a R script and an optional XML configuration.[30] The GIRI developers aim was to simplify the usage of R within i2b2 as compared to other available solutions (“R Engine Cell”[22] and “rgate HERON/i2b2”[23])

If functionalities that are important for FMS database are not already available in the community, there is always the possibility of creating a new i2b2 hive cell suiting the specific use case.

After analysis of a multitude of projects performed on i2b2, thorough utilization of the available demo client and examination of the user guides it can be concluded that i2b2 offers a good basis for the EHR data analysis of FMS

---

[21] https://community.i2b2.org/wiki/display/GIRI/GIRI+Home , accessed July 12th, 2018

[22] https://code.google.com/archive/p/i2b2-r-engine-project/ , accessed July 12th, 2018

[23] https://informatics.kumc.edu/work/wiki/HeronStatsPlugins , accessed July 12th, 2018

patients. An initial cohort analysis with aggregated patient data can be performed directly in the software and for further analysis with external tools the data can be exported after an approval is granted.

Overall i2b2 seems to offer the capabilities to perform similar studies to the one performed by De Souza et al. [23] for the grouping of FMS patients.

Furthermore, i2b2 offers the possibility to enhance collaborative research by allowing other researchers to make a cohort analysis for their studies but it also ensures the patients privacy by only displaying aggregated data without further approval.

3.3.Next Steps and Summary

An analysis of the tools i2b2, the i2b2 add-on i2b2/TranSMART and REDCap resulted in the following:

i2b2 is great tool to foster collaboration and sharing of information between different medical centers. It lets the researcher do initial research to allocate a patient cohort group that fits its desired criteria in a simple and privacy conform way. Furthermore, with the right tools (which are available within the i2b2 community) data from various sources can be integrated. In general, i2b2 has a big open source community that already offers a big variety of additional software tools and packages. One being i2b2/TranSMART, which offers more complex analysis capabilities to create an initial hypothesis.

Nevertheless, i2b2 does not seem suitable for the collection of patient data with questionnaires on a regular basis during patient visits. For this the REDCap tool seems more suitable. It has been specifically developed for the purpose of patient data collection. REDCap can be used to create and schedule individual online surveys. Those surveys can either be filled out during the patient visit on a tablet/computer or at home by access through a link.

Therefore, the next steps are for one to obtain a license for REDCap to install an instance on a server to proceed with the creation of FMS specific surveys. The other one is to create an i2b2 installation to import the data from the REDCap surveys and already available FMS patient data.

4. Business feasibility

In this chapter the business feasibility of a FMS EHR database is conducted. This entails an evaluation of the legal feasibility as well as an expense estimation and possibilities for income generation.

First the legal obstacles in terms of data protection regulations in the EU (especially the GDPR) are analyzed. Further the cost for necessary investments for server hardware and expert knowledge (IT and legal) for the setup and maintenance of a medical database are analyzed. Last but not least a possibility for income generation with a cost and profitability analysis is performed.

4.1.Legal feasibility

Another very important step in the process of a database setup that collects personal information of individuals is to be aware and adhere to the legal obligations and data protection regulations of the countries it will collect data in. Of specific importance within the EU is the GDPR, which is a regulation in the EU for data protection and privacy for individuals residing in the EU.

The regulation has been enforced on May 24th, 2016 and is applicable in the member states since May 25th, 2018.

Overview of the GDPR:

The following paragraphs shall give an overview about the GDPR and comprise a summary of the GDPR that deemed most important for the context of this thesis. A detailed overview of the GDPR can be found on the website of the European Commission[24].

A company needs to apply to the GDPR if the core business is the processing of personal information of EU residents, regardless whether the processing is done inside or outside of the EU. Furthermore, it is irrelevant whether the service charges a fee or is offered for free.[31]

The conditions for consent of the user have been strengthened. A request for the consent must be made in an understandable and easily accessible form. Clear and simple language shall be used, and it shall be distinguishable from other matters. Furthermore, it must be as simple to withdraw as to give the consent.[32]

---

[24] https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en , accessed August 12th, 2018

A specific purpose for the collected data has to be given for what and why this specific data is collected. It is only allowed to use the data for this specific purpose. The data shall not be stored longer that it is necessary. Technical and organizational measures have to put in place to ensure the security of personal data.[33]

A company needs to assign a Data Protection Officer (DPO) if its main business involves the monitoring of individuals in a large scale, regular and systematic manner. The DPO can be employed by the organization itself or contracted externally. For example, **a DPO is necessary** in a hospital **processing large sets of sensitive data**.[34]

For scientific research, the GDPR offers some simplifications regarding to how specifically and granular the consent must be given. When researchers are not able to completely identify all the purposes for the data processing, individuals can be asked to consent to research projects or certain scientific research fields.[35]

Impact on this project:

First, after consulting the GDPR it becomes evident that a database project collecting health data of individuals on a regular basis needs to take the GDPR into consideration and follow its regulations. For the purpose of scientific research some simplifications have been made in terms of obtaining consent from the patient.

Nevertheless, the GDPR puts an additional burden for the successful accomplishment of a FMS patient database project. An external consultant, such as a lawyer, should be used for reference to make sure that the database project is setup in a GDPR compliant manner. Last but not least most likely a DPO has to be hired or at least externally contracted for the purpose of this project.

This means that for such a small project extra costs due to legal counseling are created for the initial setup and most likely also on a regular basis in form a permanent legal advice through a DPO.

4.2.Expenses

First of all, a very important point for a successful database setup is reliable server hardware to install the necessary software. In this case that means the REDCap database and the i2b2 database software. The recommendation is to have different servers or at least different virtual machines for the different platforms.

Initial test installations showed that one powerful server should be sufficient to host several virtual Docker[25] containers for the REDCap and the i2b2 installation. By using Docker containers, the physical hardware can be utilized in a more efficient way compared to renting several virtual private servers (VPS).

For the search of a standard VPS a few minimum requirements (100GB storage, 8GB RAM, 4 CPU vCores and backup functionality) have been set to make the comparison of different providers easier

A comparison of different service providers yielded a VPS that meets the minimum requirements are in the range of 5€ - 12€ per month. Most providers have bigger VPS packages available, which should be taken into consideration to be prepared for an expansion of the database.

Hence the biggest cost factor will be the staff, which needs to be hired for the setup and maintenance of the database. The experience during the feasibility study showed that additional time for the database setup should be taken into consideration in order to account for unexpected obstacles. An initial and completely running system setup (including REDCap and i2b2 database) could be expected within a 3 to 6 month timeframe done by a system administrator working fulltime. The system administrator should ideally have knowledge with database systems and server security. After the initial setup the maintenance effort should decrease, depending on the customer needs.

Next to the system administrator a legal counselor for the role of DPO is needed. Most likely during the initial setup phase higher workload is expected for legal counseling but should reduce after that. There the most reasonable decision is to hire external legal counselor for the setup phase and reduce is to an on demand basis afterwards. The DPO needs to have expert knowledge about data protection law and additionally understand the IT infrastructure and technology to ensure its regulatory compliance. As explained in the previous subchapter according to the GDPR regulations it is necessary for a company that processes patient data to appoint a DPO.

A system administrator should be hired full time to take care of system maintenance, data import and customer requests. To support in the initial setup phase of the database and to ensure that a functioning import routine from REDCap to i2b2 can be implemented an IT expert on SQL databases should be contracted for a 6-month period (assuming a cost of 20 000 €). For the system admin we assume a cost of 30 000 € annually. Since the GDPR just came into place in March 2018 it is hard to estimate the workload, but it should be less than a half time position. Assuming 35 000 € for a legal expert and the need of his consultation for 1/5 out of the year (which roughly would be 8 hours per week) it would amount to 7 000 € for external legal consultation.

---

[25] Docker is software that virtualizes applications to create several instances of an operating (called containers) isolated from each other sharing the same resources on the same host system. https://www.docker.com/ , accessed September 18th, 2018

Necessary hardware (computer, cellphone, etc.) and office space is only considered for the system administrator who would work full time, at a rate of 1200 € annually. For the external IT expert and DPO a bit higher fee is assumed for working out of their own premises. A co-working space[26] in Barcelona can be rented from as low as 150 € monthly including all costs and amenities such as Internet, meeting rooms, etc.

In the second year only, the system admin and the extern DPO will work and in the third year one (or more) additional system administrator(s) should be hired depending on the amount of work and customers.

The in Table 1 presented expenses shall only demonstrate the bare minimum expected for running such as database service. Additional 10 000 € to 20 000 € should be added for the firsts year to accommodate possible marketing or travelling expenses.

Table 1: Minimum expenses to run a medical database platform.

Years	Year 1	Year 2	Year 3
Expense statement			
Staff	€ 57000	€ 37000	€ 67000
Co-working space	€ 1800	€ 1800	€ 3600
Office supplies	€ 1200	€ 1200	€ 2400
Safety buffer	€ 10000	€ 15000	€ 20000
Total operating expenses	€ 70000	€ 55000	€ 93000

4.3.Incomes

In order to generate incomes, two different business models seem possible:

The **first business model** is the provision a complete service line for a data collection and storage concept for EHRs. This includes the REDcap tool for survey creation and data collection on a routine basis as well as the import of this collected data and already existing data (at the hospital or research organization) into the i2b2 database. To each customer a personal account shall be given, which only allows him the full access to his EHRs. Since i2b2 shall foster the research and identification of cohort patient groups for medical research, it shall be possible to receive aggregated patient counts for characteristics through the i2b2 search interface. Full access to the data shall

[26] https://www.lavacacoworking.com/ , accessed September 20th, 2018

only be given if the owner authorizes the access (following all legal requirements).

Furthermore, with this concept the EHR data will be stored in a central location. This approach has the **advantage** of saving cost and effort in terms of maintenance of the server and database systems by implementing a **scalable system**. Ideally the service cost per customer should decrease the more customers join the platform. Furthermore, the customers can benefit from a growing number of EHRs they can use for their own research similar to the ESID (Ch. 2.2.3) and EHR4CR (Ch. 2.3.1) projects described before. For some diseases such as FMS it is very hard and costly to find sufficient patient data for a research project.

Disadvantages
Scaling of the database only works if there are enough customers to share the resources with. Moreover, some customers could have doubts about storing privacy sensitive data in an external database. Applying all necessary security and privacy measures necessary can mitigate this doubt to give the customer enough confidence about the security of the database service.

The **second business model** could lay in the provision of consulting services for the setup of medical databases on the customer's premises. Actually, the first and second models do not have to exclude each other. If a customer does not want to outsource its patient data to an external provider, consulting services can be delivered for the database setup to provide an individual solution adapted to the customer needs with a regular care and support plan.

The main **advantage** of the second model for the customer is to have full control about the data residing in its own hardware. Another advantage could be the possibility to modify the database specifically for individual needs of the customer.

Nevertheless, this model has **disadvantage** in terms of gathering patient data for research and fostering the collaborative research work between different institutions. Furthermore, additional staff would be needed to accommodate for the necessary consulting servers and is hard to estimate at this moment.

Revenue generation

The estimated minimum cost to setup and maintain the database service are around 70 000 € for the first year. If one would start with 10 hospitals the shared costs would be 7000 € per institution annually to maintain the service. To obtain more customers, the goal is to offer a lower price and try to share the cost between more customers. Studies in this thesis showed that patient databases could store data for several hundred thousand patients and several million records in some cases. The Hospital de Terrasa and the University Hospital of Riga Stradins University only hold around 50 to 200 patients at the moment.

Assuming 100 to 300 patients per institution it seems possible to accommodate up to 1000 institutions with one database, by simplifying and not taking the server load and usage profile into account.

The estimation is to start out with 10 institutions in the first year and the aim of serving 25 institutions in the second year and 50 institutions within 3 years. Assuming that one can acquire the aforementioned amount of institutions and that the institutions are willing to pay a usage fee of 2 500 € annually a breakeven point could be reached in the second year if the workload can be handled by one system admin in the first and second year. By the third year an additional system admin would be hired. The expected income statement is shown in Table 2.

Table 2: Expected income statement for the first 3 years.

	Year 1	Year 2	Year 3
Income statement			
Net sales	€ 25 000	€ 62 500	€ 125 000
Expense statements	€		
Operational expenses	€ 70 000	€ 55 000	€ 93 000
Net income	- € 45 000	€ 7 500	€ 32 000

Without a thorough business plan, which would include a more detailed market analysis it is hard to estimate how many research institutions and hospitals would participate in such a database and what amount they would be willing to contribute. As explained in the previous chapters gathering a big enough cohort group for a study most of the time requires a great effort and financial resources. Having access to a big database could dramatically reduce this effort and therefore save costs. Hence it can be assumed be assumed that research institutions and maybe even pharmaceutical companies would be willing to pay for such a service (as mentioned in 2.3.1 by the EHR4CR project).

Due to the time constraints of this thesis a complete business plan cannot be performed, and simplified estimations are made.

5. A use case with REDCap and i2b2

This chapter describes a use case of how patient data could be collected via surveys created through REDCap for a regular patient visit and later stored in i2b2. Entities with access to the database can run anonymous queries with their personal criteria to search for a patient cohort for their potential studies. If a cohort has been identified a patient data set can be exported.

5.1.Data collection with REDCap

As the license for REDCap for UPC is in progress, a trial version is used that has the full functionality but is limited to a time period of one week.

To simplify the use case a simple survey has been used to gather the contact and basic information about a patient (as shown in Figure 51). Two sample patients have been created to show an exemplary import to i2b2.

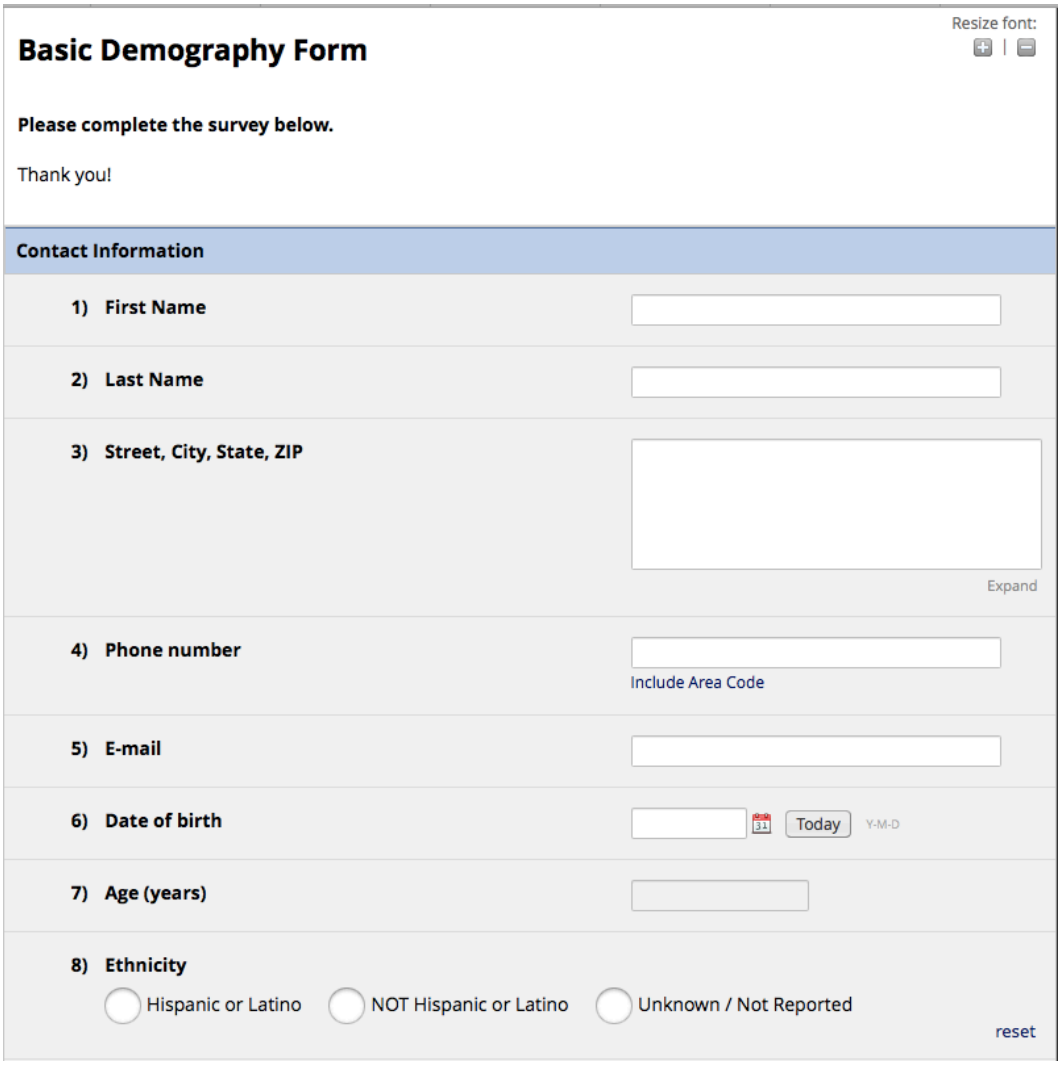


Figure 51: Basic patient information survey with REDCap.

The collected information can be viewed inside the REDCap web based user interface (see Figure 52) and exported to a CSV based file.

Study ID record_id	Repeat Instrument redcap_repeat_instrument	Repeat Instance redcap_repeat_instance	Survey Identifier redcap_survey_identifier	Survey Timestamp demographics_timestamp	First Name first_name	Last Name last_name	Street, City, State, ZIP address	Phone number telephone	E-mail email	Date of birth dob	Age (years) age	Ethnicity ethnicity	Race race	Gender sex	Height (cm) height	W
1	Basic Demography Form	1		09-20-2018 13:02	To	J	Barcelona	(223) 456-7890	tobias.joschko@gmx.de	1989-08-20	29	NOT Hispanic or Latino (1)	White (4)	Male (1)	170	7
2	Basic Demography Form	1		09-20-2018 15:29	James	Bond	London	(222) 345-1882	james@bond.uk	1960-09-05	58	NOT Hispanic or Latino (1)	White (4)	Male (1)	180	8

Figure 52: Example view of 2 patients in REDCap.

To start the data collection links to the survey can be send to the patient to fill them out at home or they could be using a tablet or computer while in the doctors office to fill out the survey. The information is collected anonymously unless it is asked for personal information. This use case contains fields for personal information and email address. If necessary, these fields can be left out and the patient can only be identified by a patient number. In the ESID database for example the patients personal information are not stored in the database and only the practicing doctor can map the patient id to the real patient.

5.2.Storage and queries with i2b2

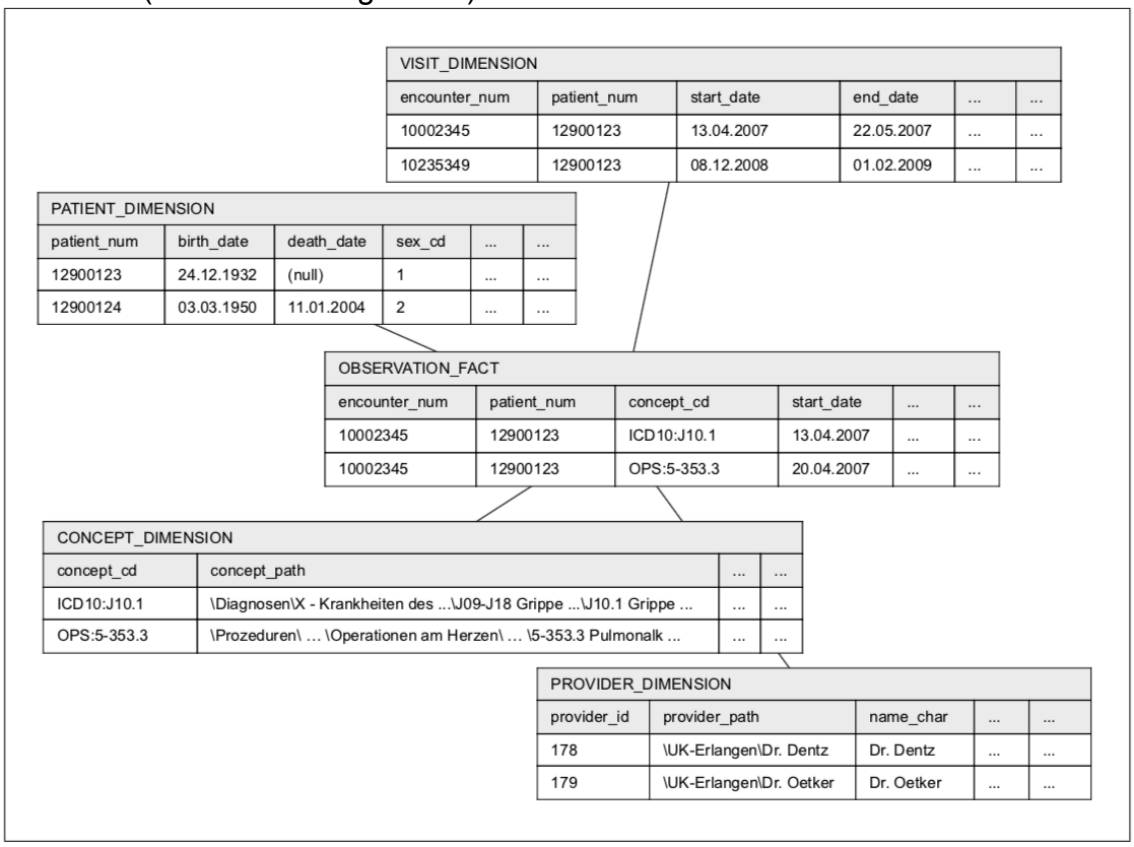
In order to import the data into the i2b2 database one needs to understand the way in which i2b2 is structuring the data. The Data CRC Cell of i2b2 takes care about the Data Mart and stores the data in a star format. The OBSERVATION_FACT is the central fact table of the database. Data in i2b2 is stored in an Entity-Attribute-Value (EAV) format. The entity in this case is the patient who is identified through the columns patient number (“patient_num”) and encounter number (“encounter_num”) as it can be seen in Figure 53. The attribute can be a classification (e.g. ICD-10) or a numeric value (e.g. age).

Other dimension tables provide information about hospital visits (VISIT_DIMENSION) or additional patient information (PATIENT_DIMENSION), such as gender, age, race, etc.[36]

So in order to import the information from REDCap one needs to export and convert it into the format of i2b2 first. This process is also known as Extraction, Transformation and Loading as described in chapter 2.3.3. For the sake of this use case only two patients have been created and their age and gender has been manually transformed into a format that can be imported to the i2b2 database. Figure 54 shows the transformation of the REDCap output table into the OBSERVATION_FACT and PATIENT_DIMENSION tables.

Once the data is stored in an i2b2 compatible format it can be directly imported into the PostgreSQL database. For this use case a CSV file with the data is

imported via the pgAdmin web based user interface of the PostgreSQL database (as shown in Figure 55).



VISIT_DIMENSION

encounter_num	patient_num	start_date	end_date	...	...
10002345	12900123	13.04.2007	22.05.2007	...	...
10235349	12900123	08.12.2008	01.02.2009	...	...

PATIENT_DIMENSION

patient_num	birth_date	death_date	sex_cd	...	...
12900123	24.12.1932	(null)	1	...	...
12900124	03.03.1950	11.01.2004	2	...	...

OBSERVATION_FACT

encounter_num	patient_num	concept_cd	start_date	...	...
10002345	12900123	ICD10:J10.1	13.04.2007	...	...
10002345	12900123	OPS:5-353.3	20.04.2007	...	...

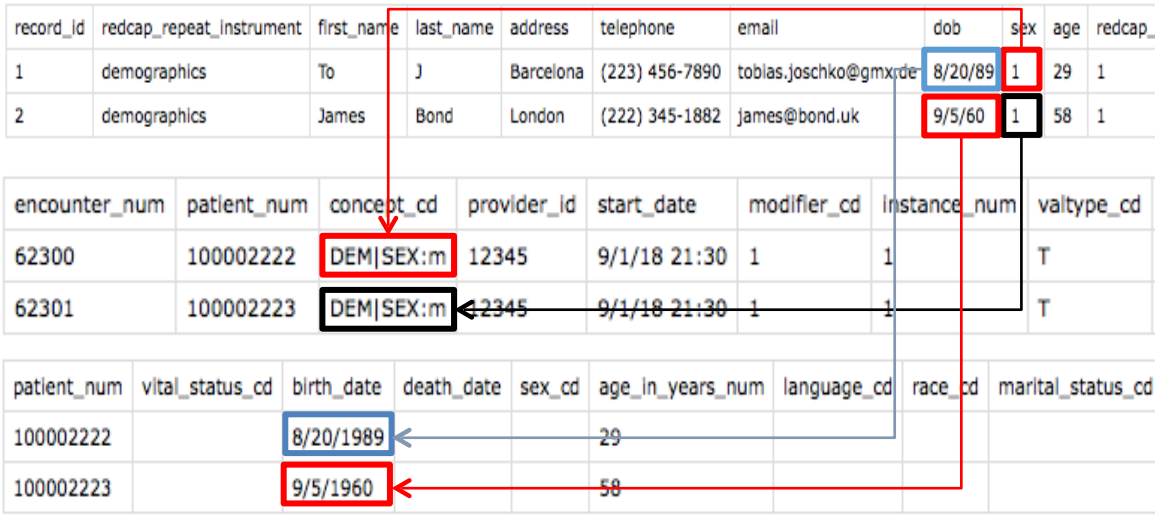
CONCEPT_DIMENSION

concept_cd	concept_path	...	...
ICD10:J10.1	\Diagnosen\X - Krankheiten des ...\J09-J18 Grippe ...\J10.1 Grippe ...	...	...
OPS:5-353.3	\Prozeduren\ ... \Operationen am Herzen\ ... \5-353.3 Pulmonalk ...	...	...

PROVIDER_DIMENSION

provider_id	provider_path	name_char	...	...
178	\UK-Erlangen\Dr. Dentz	Dr. Dentz	...	...
179	\UK-Erlangen\Dr. Oetker	Dr. Oetker	...	...

Figure 53: The CRC Cell with its star schematics, filled with exemplary entries [37].



record_id	redcap_repeat_instrument	first_name	last_name	address	telephone	email	dob	sex	age	redcap_
1	demographics	To	J	Barcelona	(223) 456-7890	tobias.joschko@gmx.de	8/20/89	1	29	1
2	demographics	James	Bond	London	(222) 345-1882	james@bond.uk	9/5/60	1	58	1

encounter_num	patient_num	concept_cd	provider_id	start_date	modifier_cd	instance_num	valtype_cd
62300	100002222	DEM|SEX:m	12345	9/1/18 21:30	1	1	T
62301	100002223	DEM|SEX:m	12345	9/1/18 21:30	1	1	T

patient_num	vital_status_cd	birth_date	death_date	sex_cd	age_in_years_num	language_cd	race_cd	marital_status_cd
100002222		8/20/1989			29			
100002223		9/5/1960			58			

Figure 54: Transformation of the REDCap output (top) to the OBERSERVATION FACT (middle) and PATIENT DIMENSION tables.

Once logged into the pgAdmin interface the prepared CSV files can be directly imported into the corresponding tables of the database. In order to test the

successful import to the database a simple query for “male patients” with a patient age breakdown has been executed before and after the import.

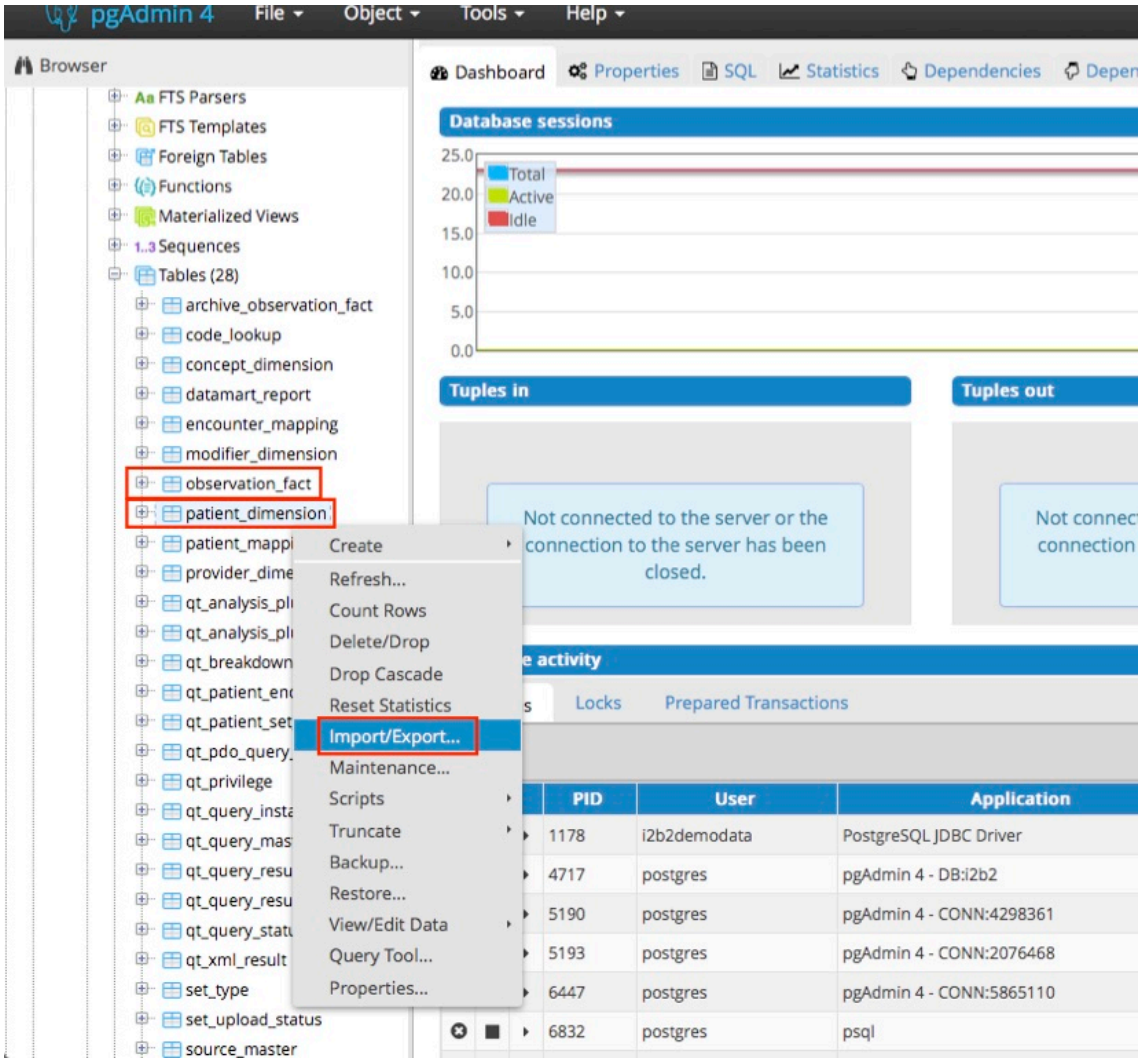


Figure 55: Screenshot of the import via the web based pgAdmin interface of the PostgreSQL database.

The query result for male patients is shown in Figure 56: **Male patient query before data import.**Figure 56. There are 82 male patients in the database. There are 28 patients in the range of 18-34 years and 8 in the range of 55 – 64 years before the data import.

After the import (Figure 57) we can see that the number of male patients has increased to 84 as well as each aforementioned age range (18-34 and 55-64 years) has increased by one.

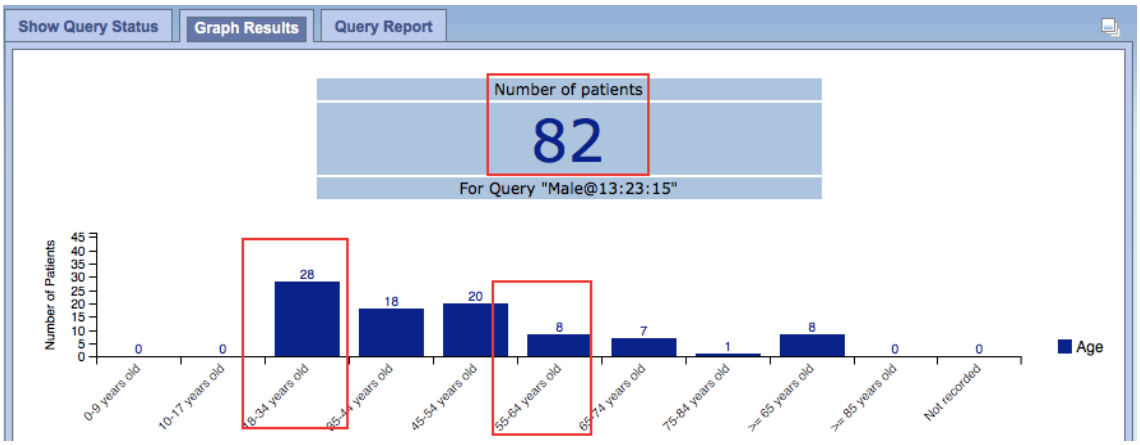


Figure 56: Male patient query before data import.

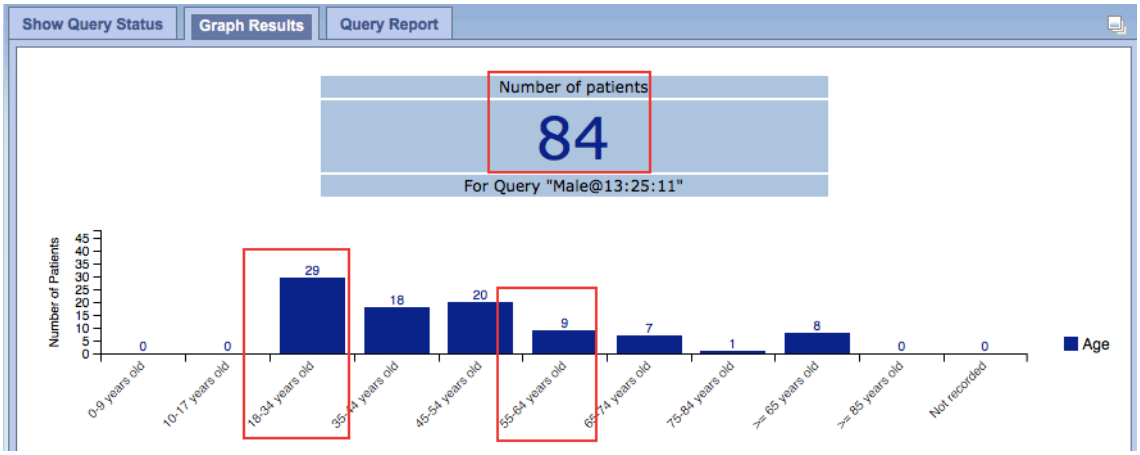


Figure 57: Male patient query after the import.

The results of this use case have been as accepted and it was possible to import two patients into the database and run a query for them successfully. More patient data (e.g. race, diagnosis, medication, etc.) can be imported into the database in the same way.

5.3.Use case insights and future development

The presented use case exemplifies a proof of concept that it is possible to collect patient data in an automated fashion via REDCap surveys and import them successfully into the i2b2 database.

It should be mentioned that the installation process was more time consuming than expected because some of the tools and tutorials did not work as expected. A more detailed description can be found in Annex **A2 – Installation process of i2b2** and hopefully gives helpful guidelines for future i2b2 implementations.

Furthermore, in order to use these tools for a large-scale collection of patient data on a routine basis the import process to i2b2 needs to be automated. Different approaches seem possible to tackle this issue: One and most likely the easiest would be to make use of an available plugin within the i2b2 community for the import of REDCap data into i2b2. The feasibility still needs to be evaluated for this use case. Another solution could be to develop a custom SQL script to directly import the data into the database.

Furthermore, another important aspect of the import routine is to develop a proper patient de-identification concept together with real world data. An evaluation of existing de-identification tools should be done.

CONCLUSIONS

Overall the objectives of this thesis have been reached. An introduction to FMS and its impact on society has been given. Possible database technologies for the setup of an FMS patient database have been researched and the most feasible technology has been selected.

Further a theoretical concept for the data collection, storage and analysis process of FMS patient information has been created. I2B2 offers a great possibility to enhance collaborated research. The graphical user interface makes the identification process of a cohort group simple, even for people who do not have a technical background. However, i2b2 turned out not to be the most optimal solution to gather patient on regular basis. Therefore, the data collection process with surveys should be done with REDCap, because it offers a simple and effective solution to create surveys and gather patient data. The collected data later can be imported to i2b2 foster research collaboration.

The created use-case with REDCap as a data collection system and i2b2 as the main storage system showed that the creation of a FMS patient database is a viable task from a technical perspective.

The business and legal analysis yielded into a not so simple answer. The expenses to maintain such a platform could be estimated and the legal obstacles evaluated. It remained difficult to estimate a correct number of potential customers and how much they would be willing to pay for this service. Legally it should be possible to run this platform, but the newly released GDPR increased the regulations and efforts to maintain it (e.g. the necessity of a DPO).

Nevertheless, to create and properly role out the FMS patient database additional steps need to be taken into consideration:

1. A concept for the whole ETL process remains to be created. Ideally an existing plugin for the import of REDCap data into i2b2 can be used but it needs to be thoroughly evaluated whether it complies with the data privacy regulations of the GDPR and whether it is a feasible solution for data collection on a regular basis. That means that doctors might want to add additional data for an already existent patient in the data and the patient id needs to be mapped with its actual patient.
2. A REDCap license needs to be obtained. Also it is important to decide whether the organization will be run as a for profit or non-profit, because the REDCap license is only free of charge for non-profit organizations.
3. A data security concept for i2b2 and REDCap is needed. All data has to be saved on the server encrypted and it should be evaluated whether the necessary standards for the medical field are met.

4. The legal regulation for data protection, especially with the GDPR must to be taken into consideration. Therefore an external expert in the field of data protection and GDPR can be consulted.
5. Even though the cost for providing such a server has been calculated it is still unclear whether hospitals and research institutions are willing to pay for such a service and whether it would be enough to cover the expenses. Hence, an in depth market analysis is necessary on whether the service can be maintained with generated revenue or whether it will be constantly depending on public funding.

Sustainability considerations

The creation of a FMS patient database can have a highly positive economical and social impact on society. Such a central database can help to foster the research in the field of FMS. If research institutions and hospitals start sharing their data it would have a benefit for all participants. First of all they could save the effort of maintaining their own database locally and increase efficiency in collecting the patient data. In addition everybody's research data would increase because one can tap into a network of medical patient data and enable research, which would have not been possible with a limited data set. A better research on FMS potentially helps to mitigate the impact of FMS on society and increase the efficiency of its treatment. Therefore a FMS patient database has more of an indirect impact on society as a whole, but can make a direct impact to medical research.

This being said, the impact factor of the database highly depends on the acceptance of the users. The more users, therefore more data the higher the impact factor on research.

Ethical considerations

From an ethical the thesis does not violate any human rights. It rather helps with regard of the human right for health. The mass data collection of personal information could raise privacy issues. Since the collected information contains a person's medical detail it is a very delicate matter. Therefore it is even more important to ensure the highest security standards possible and to only store de-identified information. It has to be guaranteed that due to the participation no harm is inflicted on any person.

ACRONYMS

ACR	American College of Rheumatology
API	Application Programming Interface
CDW	Clinical Data Warehouse
CDISC	Clinical Data Interchange Standards Consortium
CST	Hospital de Terrasa
DPO	Data Protection Officer
EAV	Entity-Attribute-Value
EDC	Electronic Data Capture
EHR	Electronic Health Records
EIDP	Enterprise Integration and Development Platform Application
ESID	European Society for Immunodeficiences
EU	European Union
FHIR	Fast Healthcare Interoperability Resources
FIQ	Fibromyalgia Impact Questionnaire
FMS	Fibromyalgia Syndrome
GDPR	General Data Protection Regulation
I2b2	Integrating Biology and the Bedside
IDRT	Integrated Data Repository Toolkit
IRB	Institutional Review Board
NCATS	National Center for Advancing Translational Sciences
NHANES	National Health and Nutrition Examination Survey
NIH	National Institutes of Health
NPL	Natural Language Processing
ODM	Operational Data Model
QEP	Query Entry Point
RSU	Ridan Stradins University
SHRINE	Shared Health Research Information Network
SHARP	Strategic Health IT Advanced Research Project
SPIN	Shared Pathology Information Network
SS	Severity Scale
TMF	German Technology and Method Platform for Networked Medical Research
UK Essen	Universitätsklinikum in Essen
VPS	Virtual Private Server
WPI	Widespread Pain Index

REFERENCES

[1] W. Häuser, W. Eich, M. Herrmann, D. O. Nutzinger, M. Schiltenwolf, and P. Henningsen, “Fibromyalgia syndrome: classification, diagnosis, and treatment.,” *Dtsch. Arztebl. Int.*, vol. 106, no. 23, pp. 383–91, Jun. 2009.

[2] F. Wolfe, D. J. Clauw, M. A. Fitzcharles, D. L. Goldenberg, R. S. Katz, P. Mease, A. S. Russell, I. J. Russell, J. B. Winfield, and M. B. Yunus, “The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity,” *Arthritis Care Res.*, vol. 62, no. 5, pp. 600–610, 2010.

[3] J. Lee, K. Lee, D. Park, S. Kim, S. Nah, H. Lee, S. Kim, Y. Lee, S. Hong, H. Kim, S. Lee, H. A. Kim, C. Joung, S. Kim, and S. Lee, “Determinants of quality of life in patients with fibromyalgia : A structural equation modeling approach,” 2017.

[4] R. M. Leadley, N. Armstrong, Y. C. Lee, A. Allen, and J. Kleijnen, “Chronic diseases in the European Union: The prevalence and health cost implications of chronic pain,” *Journal of Pain and Palliative Care Pharmacotherapy*. 2012.

[5] K. P. White, M. Speechley, M. Harth, and T. ØStbye, “The London fibromyalgia epidemiology study: The prevalence of fibromyalgia syndrome in London, Ontario,” *J. Rheumatol.*, 1999.

[6] M. Spaeth, “Epidemiology, costs, and the economic burden of fibromyalgia,” *Arthritis Res. Ther.*, vol. 11, no. 3, p. 117, Jun. 2009.

[7] A. Boonen, R. van den Heuvel, A. van Tubergen, M. Goossens, J. L. Severens, D. van der Heijde, and S. van der Linden, “Large differences in cost of illness and wellbeing between patients with fibromyalgia, chronic low back pain, or ankylosing spondylitis.,” *Ann. Rheum. Dis.*, vol. 64, no. 3, pp. 396–402, Mar. 2005.

[8] M. Rodriguez, J. Sempau, and L. Brualla, “PRIMO: A graphical environment for the Monte Carlo simulation of Varian and Elekta linacs,” *Strahlentherapie und Onkol.*, vol. 189, no. 10, pp. 881–886, Oct. 2013.

[9] R. P. Campos, M. I. Vázquez Rodríguez, and M. I. R. Vázquez, “Health-related quality of life in women with fibromyalgia: clinical and psychological factors associated.,” *Clin. Rheumatol.*, vol. 31, no. 2, pp. 347–55, 2012.

[10] ERA-NET NEURON Joint Call, “PANDORA: Project Proposal for Transnational Research Projects on Mental Disorders. (Internal Report),” 2018.

[11] “i2b2: Informatics for Integrating Biology & the Bedside.” [Online]. Available: https://www.i2b2.org/about/index.html. [Accessed: 04-May-2018].

[12] G. M. Weber, S. N. Murphy, A. J. McMurry, D. MacFadden, D. J. Nigrin, S. Churchill, and I. S. Kohane, “The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories,” *J. Am. Med. Informatics Assoc.*, vol. 16, no. 5, pp. 624–630, Sep. 2009.

[13] D. Guzman, D. Veit, V. Knerr, G. Kindle, B. Gathmann, A. M. Eades-Perner, and B. Grimbacher, “The ESID Online Database network,” *Bioinformatics*, vol. 23, no. 5, pp. 654–655, Mar. 2007.

[14] “ESID - European Society for Immunodeficiencies.” [Online]. Available: https://esid.org/. [Accessed: 01-Jun-2018].
[15] Toolwerk GmbH, “Toolwerk GmbH Home Page.” [Online]. Available: http://www.toolwerk.de. [Accessed: 18-May-2018].
[16] R. Bache, S. Miles, and A. Taweel, “An adaptable architecture for patient cohort identification from diverse data sources,” *J. Am. Med. Informatics Assoc.*, vol. 20, no. E2, 2013.
[17] A. R. Tate, N. Beloff, B. Al-Radwan, J. Wickson, S. Puri, T. Williams, T. Van Staa, and A. Bleach, “Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface,” *J. Am. Med. Informatics Assoc.*, vol. 21, no. 2, pp. 292–298, Mar. 2014.
[18] S. Rea, J. Pathak, G. Savova, T. A. Oniki, L. Westberg, C. E. Beebe, C. Tao, C. G. Parker, P. J. Haug, S. M. Huff, and C. G. Chute, “Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project,” *J. Biomed. Inform.*, vol. 45, no. 4, pp. 763–771, Aug. 2012.
[19] T. Ganslandt, S. Mate, K. Helbing, U. Sax, and H. U. Prokosch, “Unlocking Data for Clinical Research – The German i2b2 Experience,” *Appl. Clin. Inform.*, vol. 2, no. 1, pp. 116–127, Dec. 2011.
[20] K. Pommerening and M. Reng, “Secondary use of the EHR via pseudonymisation,” in *Studies in Health Technology and Informatics*, 2004, vol. 103, pp. 441–444.
[21] K. . Helbing, S. Y. . Demiroglu, F. . Rakebrandt, K. . Pommerening, O. . Rienhoff, and U. . c Sax, “A data protection scheme for medical research networks review after five years of operation,” *Methods Inf. Med.*, vol. 49, no. 6, pp. 601–607, 2010.
[22] “IDRT - The Integrated Data Repository Toolkit for i2b2.” [Online]. Available: http://idrt.imise.uni-leipzig.de/IDRT-II/#Contact. [Accessed: 22-Jun-2018].
[23] J. B. De Souza, P. Goffaux, N. Julien, S. Potvin, J. Charest, and S. Marchand, “Fibromyalgia subgroups: Profiling distinct subgroups using the Fibromyalgia Impact Questionnaire. A preliminary study,” *Rheumatol. Int.*, vol. 29, no. 5, pp. 509–515, 2009.
[24] J. Doods, R. Bache, M. McGilchrist, C. Daniel, M. Dugas, and F. Fritz, “Piloting the EHR4CR feasibility platform across Europe,” *Methods Inf. Med.*, vol. 53, no. 4, pp. 264–268, Jan. 2014.
[25] Vanderbilt University, “REDCap.” [Online]. Available: https://www.project-redcap.org/. [Accessed: 16-Jul-2018].
[26] “REDCap Technical Overview.” [Online]. Available: https://projectredcap.org/wp-content/resources/REDCapTechnicalOverview.pdf. [Accessed: 16-Jul-2018].
[27] “UMass Medical School - Clinical Data Portal - Information Technology.” [Online]. Available: https://umassmed.edu/it/cdp/. [Accessed: 02-Jul-2018].
[28] “NHANES- National Health and Nutrition Examination Survey.” [Online]. Available: https://www.cdc.gov/nchs/nhanes/index.htm. [Accessed: 11-Jul-2018].
[29] C. J. Patel, N. Pho, M. McDuffie, J. Easton-Marks, C. Kothari, I. S. Kohane, and P. Avillach, “A database of human exposomes and

phenomes from the US National Health and Nutrition Examination Survey," *Sci. Data*, vol. 3, 2016.

[30] "GIRI Home - GIRI (Generic integration of R into i2b2) - i2b2 Community Wiki." [Online]. Available: https://community.i2b2.org/wiki/display/GIRI/GIRI+Home. [Accessed: 04-Jul-2018].

[31] European Commission, "Who does the data protection law apply to?," 2018. [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/application-regulation/who-does-data-protection-law-apply_en. [Accessed: 02-Aug-2018].

[32] European Commission, "What information must be given to individuals whose data is collected?" [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/what-information-must-be-given-individuals-whose-data-collected_en.

[33] European Commission, "What data can we process and under which conditions?" [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/what-data-can-we-process-and-under-which-conditions_en. [Accessed: 02-Aug-2018].

[34] European Commission, "Does my company/organisation need to have a Data Protection Officer (DPO)?" [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/data-protection-officers/does-my-company-organisation-need-have-data-protection-officer-dpo_en. [Accessed: 06-Aug-2018].

[35] European Commission, "How is consent for processing in scientific research obtained? | European Commission." [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/grounds-processing/how-consent-processing-scientific-research-obtained_en. [Accessed: 07-Aug-2018].

[36] i2b2, "i2b2 User Guide Import Data View."

[37] S. Mate, "Evaluation von i2b2 am Universitätsklinikum Erlangen," no. August, 2009.

Annexes

A1 – List of possible FMS symptoms

The symptoms of FMS are very diverse and incorporate restless sleep, fatigue, irritable bowel syndrome, cognitive disturbance, skin tenderness, post-exertional pain, stiffness, irritable bladder syndrome, headaches, parenthesis, restless legs, mood disturbances, dizziness and fluid retention.

A2 – Installation process of i2b2

In the following the experience and issues occurred during the setup of i2b2 shall be described and offer a potential guideline during the another installation of i2b2.

The installation process of i2b2 on the UPC server turned out to be significantly more time consuming than it was expected.

Two Docker installations from Docker Hub have been tested. One Docker image contained the i2b2 Wizard[27] (as described in 2.3.3) utilizing an Oracle database as the underlying SQL database. The other Docker installation contained an i2b2 installation with an underlying PostgreSQL database and a WildFly[28] application server[29].

The automatic installation of the i2b2 Wizard Docker image was not working because it could not connect to the Oracle database. Therefore, the steps of the manual quick start guide[30] from the i2b2 community have been followed. After about two days trial-and-error the i2b2 Wizard ran through successfully. The web client login worked but it was not possible to access the admin menu, to connect to the i2b2 Hive via the Workbench and to access the i2b2demodata on the Oracle database.

The second Docker image caused less effort to execute. Three Docker images (i2b2/i2b2-web, i2b2/i2b2-wildfly and i2b2/i2b2-pg) for the web-interface WildFly application server and PostgreSQL database, from the i2b2 DockerHub, were used. After it was possible to log into the web client and admin menu. Nevertheless, a connection to the Workbench still was not possible.

---

[27] https://hub.docker.com/r/cyberseb/i2b2/ , accessed July 20th, 2018

[28] http://wildfly.org/ , accessed July 20th, 2018

[29] https://hub.docker.com/u/i2b2/ , accessed July 20th, 2018

[30] http://community.i2b2.org/wiki/display/IDRT/260.+i2b2+Wizard+quick+start, accessed June 4th, 2018

In a conversation with Sebastian Mate (the developer of the i2b2 Wizard) he confirmed that the version of the i2b2 Wizard has not been updated for some time and is not completely compatible with the latest i2b2 software. Furthermore, he provided additional information about his experience with the i2b2 setup and in general information about i2b2 and its data structure.

Hence, the decision was made to continue with the second Docker image, which caused less trouble during the installation process and allowed more access to i2b2.

The connection with the Workbench was tested from different computers with different operating systems (Windows and a provided Ubuntu Virtual Box image from Mac OS X). The connection the to installed version on the UPC server as well as to the test server provided by i2b2 was unsuccessful.

On the official website of i2b2[31] the client software, a Client & Server VMWare image and the i2b2 source code are provided. In order to rule out connection issues, locked ports or a firewall blocking access, the next step was to utilize the Client & Server VMWare image provided by i2b2. The guidelines recommendation is to use VMware server 1.0.5 or higher. When using the referenced link once becomes to know that the VMware server product line does not exist anymore. VMware has only a few products to offer that are free of charges and others in the range of 500 – 3000 USD per year. Testing on Windows, MAC OS X and Ubuntu with the available VMware products and the virtualization software VirtualBox were not successful.

Since i2b2 is an open-source system there is no dedicated support team available to assist with i2b2 issues. Nevertheless there is a JIRA Bug tracking system[32] to report issues (which eventually somebody will fix), there is a Google discussion group[33] to help with the installation of i2b2 and there is the i2b2 community wiki[34]. So with some time and dedication it is possible to gather the information needed to get the system running.

The login issues for the Workbench have been fixed, but after that the import plugin for i2b2 did not work. Since at this point more understanding about the database structure of i2b2 was present, the decision was made to directly import the patient data into the PostgreSQL database. It also makes more sense to directly work with the database because to (later) provide an automated system for import of data into i2b2 either a personally developed SQL script or an existing ETL solution should be used with some adaption.

---

[31] https://www.i2b2.org/software/index.html , May 3rd, 2018

[32] http://community.i2b2.org/jira/browse/ , September 10th, 2018

[33] https://groups.google.com/forum/#!forum/i2b2-install-help , September 10th, 2018

[34] https://community.i2b2.org/ , July 10th, 2018

To summarize, it is good to know that there is a group discussion and issue tracking system present, which can help with the installation of i2b2. Furthermore, one should also bring some patience, because with an open-source system not everything is always working as expected and some sources might be not available temporarily.