

---

# Data Integrity and Privacy in an Electronic Case Report Form

---

Master of Science in Technology Thesis  
University of Turku  
Department of Future Technologies  
Software Engineering  
2018  
Esa Aaltonen

Supervisor:  
Ville Leppänen

Health records are used to map the overall health of a patient and most information related to their current status such as existing diseases and medication are usually entered into these records. These records are generally used to provide to the patient the healthcare that they need. Electronic health records (EHR), that are used by healthcare professionals to store the data electronically, are becoming more open in a way that include the patient in some way as well. EHRs that give the patient the chance to view information related to their own health, have become more common in the last few years, especially in Finland where the existence of this type of EHRs was rare just ten years ago.

In addition to using these records in clinical care, they can also be used in healthcare research. Case report forms are used as source documents to gather or answer specific questions related to the research study in question. The data that is produced this way is usually pseudonymized and a lot of it is usually gathered to help with the accuracy of the study.

Usually when gathering this data, there are researchers and other people entering the data electronically first before it can analyzed. So that this data can be used more efficiently and to reduce the risk of inaccurate data, the form where it is entered can be programmed to monitor the entries to help researchers, and make sure the data is more uniform across the dataset.

To achieve this kind of validation, two different techniques are researched and executed in an existing electronic Case Report Form (eCRF) system. The form is changed so that it gives feedback to the data entry person in real time. The form is also looked at from the point of data extraction and if some existing problems can be solved.

The new European General Data Privacy Regulation is also looked at and its impacts on the case eCRF, and what changes must be made to the system for it comply with the regulation.

Keywords: Electronic health records, research studies, General Data Protection Regulation, form validation

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	2
1.2	Research goals . . . . .	3
1.3	Thesis structure . . . . .	3
<b>2</b>	<b>Health records</b>	<b>4</b>
2.1	Electronic health records . . . . .	6
2.2	Finnish EHR systems . . . . .	7
2.3	System integration standards . . . . .	9
<b>3</b>	<b>Using health records</b>	<b>11</b>
3.1	In operative situations . . . . .	11
3.2	In research studies . . . . .	12
3.2.1	Case report forms . . . . .	14
<b>4</b>	<b>General Data Protection Regulation</b>	<b>16</b>
4.1	Biggest changes in GDPR . . . . .	17
<b>5</b>	<b>Detailed description of the problems</b>	<b>21</b>
5.1	Data collection . . . . .	21
5.2	Data extraction . . . . .	26
5.3	Data privacy according to GDPR . . . . .	27

5.4	Summary of problems and concerns . . . . .	28
<b>6</b>	<b>Case eCRF system</b>	<b>30</b>
6.1	P1 - Data collection forms . . . . .	30
6.2	P2 - Data exporting . . . . .	42
6.3	P3 - Data privacy under GDPR . . . . .	45
<b>7</b>	<b>Conclusions</b>	<b>48</b>
7.1	P1 - Data collection . . . . .	48
7.2	P2 - Data exporting . . . . .	49
7.3	P3 - GDPR . . . . .	50
	<b>References</b>	<b>51</b>

# 1 Introduction

Patient records, or health records are collected from patients throughout their lives every time they visit a healthcare facility, pharmacy or other institution that handles information concerning the patient's health. The record contains such information as for example: base information (weight, height, age, etc), existing conditions, prescriptions, visits to a doctor and possible past and future operations. Considering this, these records are sensitive in a way that they usually contain very private information on the patient, and they should be handled with care. The primary use of these records are for the patient's healthcare, and it is in the patients', as well as the healthcare professionals' interest that these records are protected.

The medical records are mostly in electronic form nowadays, known as Electronic Health Records (EHR). Storing and handling the records electronically provides many advantages over paper records, such as better accessibility, data redundancy in the form of back-ups and better availability for secondary use. For secondary uses the data is used more widely, and usually in bigger volumes, comprised of even thousands of patients' records. Since it is not directly used for medical care of the patient in these cases, it needs to be properly made pseudonymous so that a patient cannot be singled out from the data. General Data Protection Regulation (GDPR) is a new European regulation on how individuals private information is supposed to be handled when stored in on-line databases. While it is not specifically about health records, the same rules apply for them as well.

One purpose for secondary use of these records is for healthcare studies. The data is collected with Case Report Forms (CRF) using specific questions relating to the study. When enough data is collected, it is analyzed to see how it relates to the study question.

This thesis will look into how better data integrity could be achieved when collecting these records into an electronic Case Report Form (eCRF) to be used in various healthcare studies. It will also look into how the new privacy regulations could affect collecting these records. Both of these points will be looked at from a general perspective and from the perspective of an existing eCRF system.

## **1.1 Motivation**

This thesis was inspired partly by an existing medical research data collecting and distributing system and some of the problems that even such a simple system sometimes faces. It is beneficial to study these problems, not only to help the existing system to solve or at least ease with some integrity issues of the data it handles and ensure that the data can be used in future researches properly, but also so that in the future when making similar systems some of these problems could be expected and hopefully avoided.

Data integrity is very important, since the data that this existing system is used to collect is used in academic studies that aim to develop better healthcare methods and study effects of different medications. The aim is for the information gathered and then passed on to be as correct as it can be. The case system that is talked about later in the thesis is our primary motivation when exploring different methods for achieving the data integrity that is necessary for the specific studies that this system is used for. It has been found that the system in question needs some improvements when it comes to collecting and extracting the research data.

Recent concerns and regulations that protects the patient's anonymity in the research

database also need to be addressed and possibly make changes based on the conclusions that studying these regulations might arouse. Because the data entered into the case system is supposed to be made pseudonymous beforehand, the new regulations should not affect the case system, at least not directly. It is still necessary to explore how much and what kinds of data can be collected under the GDPR, and whether it has an effect on existing eCRFs. The sanctions for not following these regulations in the case of a breach can be severe.

## **1.2 Research goals**

How to help with gathering and transferring of health record data between systems to eliminate human error. Formatting the data so that it can be properly used with the statistical tools used to analyze the data for study purposes.

Look into research data in the case system, especially privacy and information security and how they comply with the new GDPR, and what possible changes need to be made.

## **1.3 Thesis structure**

In Chapter 2 it is explained what health records are and how they are formed, and also how electronic records help with some problems. Chapter 3 explores how those records are used in operative situations, and how they are used in studies, especially in those that the case system is a part of. After that, in Chapter 4, we look at the new GDPR and what changes these regulations impose. In Chapter 5, we detail problems with health data collection and how these problems manifest in our example environment and what possible effects the GDPR has on it. Chapter 6 deals with the actual fixing and research of the problems listed in Chapter 5. Last is Chapter 7 with conclusions.

## 2 Health records

This chapter looks into health records, how they are formed, their possible flaws and how those flaws can be mitigated with electronic health records.

A health record is basically different types of information gathered from a patient which are then traditionally stored within the organization that gathered it, a hospital for example. These records can be comprised of variety of information, such as:

- Identifying information, for example: name, address and social security number
- Patient characteristics, for example: height, weight, age and gender
- Existing diseases
- Current medication and prescription information
- Laboratory results
- Doctors' diagnoses
- Follow-up informations

This information is used to map the patient's overall health and to help with more specific diseases or ailments. Medical records are also useful, not only with patient's current problems, but also with any future medical concerns they might have. All this data leads to a better understanding of the patient's health and medical condition. Therefore proper storage and care of this data is very important.



In the past, these records were mostly collected and stored manually, using paper forms and such, but there are numerous problems and inconveniences involved with these methods, such as physical storage, data longevity, accessibility and integrity.

Data longevity could be shortened simply by paper deteriorating over time. Paper is also flammable and hard to backup, which can result in things such as natural disasters or accidents leading to destruction of records.

Accessibility is a problem as the records will take space and have to be stored somewhere, usually in filing cabinets and rooms. Fetching these records then takes resources, such as manpower and time. Sometimes the records can also be lost due to filing errors or similar mistakes that lead the data to be somewhere else than where it is supposed to be. Transferring the records between institutions, different hospitals, private clinics, would also lead to similar problems, but transferring it would take even more time and resources. It would also be a hard to collect there records for secondary use.

Data integrity is a concern when it comes to collecting the data and expecting it to be collected right, especially with physical mediums as there usually are no checks to ensure this. Errors can manifest easily when the information is written down.

The storing and availability of medical records is not only concern for the patients themselves, but also for future researches that this data could be used for.

We can conclude from these different problems that physical records, or any information in physical form, face with collection, storage and transportation, that all these aspects are done more efficiently electronically. Data collection can be improved with electronic forms and other techniques, data storing and longevity problems are solved by storing them in databases and having regular back-ups done and accessibility is enhanced through the use of internet and networks.

## 2.1 Electronic health records

Since the invention of computers and databases, hospitals and other healthcare institutions have used electronic means to collect and store most of the data they need. Especially due to the problems with physical records. Still it is not uncommon to make paper notes first and then create or update the records in digital form. Recently the increasing popularity and implementation of cloud services has also helped with some accessibility problems.

There does not exist a precise description or standard on what information is included in electronic health records (EHRs). The term is also sometimes used interchangeably with personal health record (PHR). These records could potentially include all or some of these informations [1]:

- Electronic clinical records (medical records, laboratory values and all others listed at the beginning of Chapter 2)
- Information on past and future appointments to the doctor or tests
- Prescription information, past and present
- Communications between patient and healthcare professionals
- Data bank on current diseases for better self-treatment

PHRs are usually meant to mean records that the patients themselves can access and fill, and through this access have a more active role in their own healthcare. Older people might have more problems in adopting and using such services. [2]

While it might be obvious that healthcare providers have already moved to using EHRs, these personal records are still somewhat of a service which is not available everywhere. Their development and implementation is usually a concern for third-party service providers.

In the context of this paper and the healthcare research we will mention, we are mostly interested in the first item on that list which is clinical records and lab results and such, however it is worth looking into what kinds of information can be stored and accessed online by healthcare professionals and patients. In the future there might be a chance to use more of the information for those studies, when patients can participate in providing this information. It is also interesting to see that in Finland there has only recently been a centralized and national system for patients to access their information through the internet. Similar systems have existed in some other countries for well over ten years. [1]

Another advantage of electronic records is the case of natural disasters and other such events that might destroy physical records. Since these records are vital to patients, great care should be taken in preserving them. [2]

## **2.2 Finnish EHR systems**

There are many different EHR systems in Finland that are used to manage patient records and data. These systems are developed for healthcare professionals and are used to fetch patient information when needed, for example when the patient visits or calls a doctor, when storing investigation results or medication information. The reason why there are many different systems is that the records are not, or at least were not in the past, managed on national level. It was up to the hospitals, towns or regions to procure and pay for their own system. The system could have been an existing system that was modified to meet the organizations needs, or a completely new system made specifically for the organization. Whatever the case, it would have been up to the organization to choose the supplier for the system and there are some different aspects that would have affected the decision. Pricing, location, customization and experience of the supplier and feedback from other regions could have been some of those aspects.

There have been studies made on these EHR systems especially on the subject of

how they rank between each other in usability, technical reliability, data availability and communicating between different systems. The recent study from year 2014 indicates that with every system studied there are problems in many different areas, and with all of them it takes too much time to get records from different organizations [3]. This would indicate difficulties in how these systems communicate with each other, which could be because the systems are made by different companies, for different organizations with different needs. There exists standards on how especially medical documents should be formatted so that they can be transferred and interpreted correctly. One of these standards is Health Level 7 (HL7).

Since the study came out however, some of the systems have been integrated together and some have been renewed. Another study could see some improvements in usability and communications [4].

For patients there have been less options for viewing and managing their records and other medical data electronically, as these are traditionally something that health-care professionals handle. In recent years there have been improvements regarding that as well, with electronic prescriptions and the Omakanta-service from where users can view, among other things, their doctors visits, medications and lab results. Bringing patients immediate and more information about their health could make them more interested in their healthcare and generally improve their health. It is also more convenient than visiting or calling the hospital about such things as lab results.

In the year 2010 there was a study done on how common electronic services were in healthcare, in Finland. Different hospital districts and health centers were researched on what different electronic services they provided. Informing the public about services on their respective websites is something that existed in all of these organizations, and electronic making of appointments, guidance and evaluating the need of treatment were something about half of the organizations had. For patients, modifying their own personal information was possible in 6% of health centers and viewing and saving other informa-

tion related to their health was possible in only 1-2% of health centers and in one hospital district. It was also noted that there were multiple plans or projects underway in many of these categories. [5]

## 2.3 System integration standards

Transferring health records, and any sub-records or other information that could later be included in a health record, between different organizations and institutions and between different electronic systems can be challenging as many of these organizations might have their own customs on how the information is formatted within them. Another thing that can bring difficulty is different conventions and different character sets between different countries, and even within a country.

To get all these different parts to be able to communicate with each other, one would try to reach a so-called 'semantic interoperability', which is defined as such: "Ability for data shared by systems to be understood at the level of fully defined domain concepts" [6]. For something like that to be possible, every message these subsystems produce and communicate would have to be carefully structured based on instructions, standards or specifications agreed upon by all those organizations involved.

HL7 is a set of standards developed by the Health Level Seven corporation. These interface standards or specifications are for health records and other medical data and how they are formatted electronically. This formatting is meant to make it easier to transfer that data between different systems, especially new systems, and add new functionality to existing systems. There are other such standards developed by different organizations. [6]

The reason for these sorts of standards is that different organizations can have different ways of collecting and storing information, which leads to the data being formatted differently. It becomes difficult for those systems to communicate between each other

without any kind instructions on how their messages are supposed to be structured.

Since there are many different standards and organizations making those standards, it should become somewhat impossible to attain perfect interoperability. Making these standards and then teaching people how to use them are also a business for these organizations, and thus developing a single standard is not likely to happen, at least not without interference by a regulating body.

## **3 Using health records**

Primarily health records are used first hand to improve patient healthcare because they form a bigger picture of the patient's health. They can also be used as a source of information in medical research by combining many records, and using them to answer questions with a specific research question as the subject. This chapter looks into how health records are used, especially in research studies. Case report forms are also looked at and how they are used to collect and use health records in clinical studies.

### **3.1 In operative situations**

Health records are used in patient healthcare to ensure the patient gets the best care they can presently and in the future. As explained earlier, electronic records increase the availability of those records so that doctors and nurses have the information readily available that they need to provide care to the patient.

From patient's records, doctors can view for example possible allergies to different medications, and use that information to prescribe medication. More advanced usage of patient's records could be for long-term care and monitoring of specific diseases. As individuals, the patients sometimes need individual care and more conclusive records give a better chance for the patient to receive the care they need.

Another situation for the need of health records in operative situations is when there is an emergency. Sometimes in these situations the patient might be unconscious or oth-

erwise unable to answer any questions the healthcare personnel might have, and they can then use the records to fetch relevant information.

## 3.2 In research studies

Healthcare studies or clinical research aim to study and answer a specific research question or hypothesis, usually with the aim to find out whether some practises in healthcare can be improved. Clinical research questions can include:

- How treatment and practises of care for different illnesses could be improved to provide better care for patients
- How advancements in technology could aid in patient care and how current practises could be changed to use these technologies
- Finding out how current medications compare to newer medications in treating different illnesses. Also how medications work in different situations and how they work together with other treatments, their effectiveness and possible adverse effects

There are always improvements that can be made to patient care and pharmaceutical companies are usually also interested in these studies.

Valid and accurate patient records can be invaluable information for healthcare institutions and medical companies performing researches in various medical fields. Data collecting and especially big data can be used for different kinds of studies. The more data there is, the more accurate are the results or observations based on it. That is of course if the data is collected and processed properly. A local study might have around 100 suitable patients that the data can be collected from, more or less depending on the scope of the study. It makes sense then to try and get more hospitals or centers to participate in the data collection so that more accurate results can be produced.



The data collected for a study is usually at least somewhat prepared for third party use, meaning it is pseudonymized and selected so that it can be used for a specific study question. Pseudonymization in this context means that from the data collected for research, individual patients can not be singled out, but are given an identifier so that the researchers can still update the data if there are changes. It is in fact illegal in many countries to pass on patient records to third parties, or use them in such a way, without de-identifying it first. The data is collected with the patient's consent.

EHRs help with these studies from the perspective of availability. In Table 1 there are listed some different sources that researchers might have available for them get EHRs or other for their studies. [7]

From Table 1 we can see that for example if the study is held in a single organization, the data is much easier to use, but might not be diverse or abundant enough. Then again using a databank for the data, there is more data and therefore the results should be more accurate, however the data might be older and needs more work before it can be used.

Table 1: Data sources for clinical studies

EHR sources	Advantages	Disadvantages
Data from or within a single organization	Less problems with data management, specifications and consents	Might not have enough data for important studies. Possible lack of research tools
National or regional data-banks or disease registers	Records from many sources. Larger amounts of data. Better structuring of data	Need to apply for permission to use the data. Data might not be most current. Additional steps needed to obtain data and verify it
Study specific eCRF	Structure and data is managed locally and offers better management	Harder to set up and more expensive. Data has to be transferred to the system
Federal database of EHRs	Large amount of records	Consents more difficult to manage

### 3.2.1 Case report forms

Case report forms (CRFs) are used to collect specific information about patients and in the context of this paper they are used mainly in clinical studies. CRFs are usually structured with a specific study question in mind. The questions in the form are selected and the form structure is defined so that the answers to these questions provide the researchers enough data to analyze it and hopefully get an answer to the research question.

Gathering this data is much more convenient when done electronically, though some hospitals might still gather the data on paper and transfer it to digital format later. To collect this data on computers, electronic case report forms (eCRFs) are used. They are also more convenient when collecting data from many different organizations, as then

they can all use the eCRF to fill in their data and it does not need to be transferred by other means.

Later on in this paper the term eCRF is used quite often and 'the eCRF' is used to refer to the case system which houses an eCRF. Even though there exists some free forms that can be used the same way as the eCRF, the case system was build so that it can be modified to specific needs more easily. It also becomes cheaper and faster to maintain a custom simple eCRF than to have an actual company do it.

## 4 General Data Protection Regulation

General Data Protection Regulation (GDPR) is new European Union regulation which imposes new data protection handling requirements for companies and other organizations. These regulations mostly concern transferring and storing personal data, and procedures after a possible breach has happened.

Personal data meaning any data that can be used to single out a person, including such data as: names, addresses, usernames, passwords, social security numbers and photos. There have been other laws and regulations concerning storing private information into second and third party systems. These new regulations are meant to unify and solidify these customs, over country specific laws. There have been many cases in recent years when private information about customers and other users have been leaked due to insufficient privacy customs and oversights in organizations. Also because of the continuing increase in storing information online and possibly even selling that information to third parties, there needed to be stricter and more solid regulations and procedures for this.

GDPR in its current form was approved in May 2016 and enacted in May 2018, after a two year transition period [8]. The effects were widely seen as most web pages in Europe were forced to inform their users about how their data is used. Even many websites not physically located in the European Union forbid access from the EU, at least until they could implement a sort of disclaimer for EU users. It is also possible that some sites restricted access altogether. Some of the biggest visible changes to these websites were the inclusion of a pop-up window which detailed what the user's data is used for and

asked for consent from the user for the purpose of storing private data. The users would also see an influx of e-mail messages, provided that they had given their address to the website or service provider. These e-mails informed the user about the privacy changes the service had adopted.

There is another proposal issued for more regulation dealing with electronic marketing, called e-Privacy Directive. This regulation would apply to e-marketing and more specifically requiring consent for commercial communications, email advertisements, and in other ways dealing with unwanted communications. [9]

## 4.1 Biggest changes in GDPR

The biggest changes in the new regulation [10]:

- Consent

Personal data processing requires consent from the subject or data owner, and must be given freely without pressure, so that there is a real choice for the user. The choice to give consent must also be specific and informed, which means that the user must be informed on service provider's identity, how the data will be used and what data will be collected and processed. Consent must also be given clearly, so that there is no doubt whether or not the consent was given.

- Personal data

Personal data means any data that is related to a directly or indirectly identifiable person. There are many different identifiers for direct identification, such as name, location data, online names, addresses, physical conditions, etc. Basically anything that is assigned to a person. IP addresses can also be considered personal data in some cases, such as if the provider is required by law to hand over those addresses to a party which can link them to real people. However storing such data as IP

addresses is allowed for a limited time when it is used to ensure security in the system [11].

- Right to get data

The user has the right to request the provider to inform them and produce for them all the data they have collected and stored about the user. Within a certain time period after the request for the information is received, they must reply.

- Right to be forgotten

The data subject can withdraw their consent or the original purpose for which the data was collected can become no longer valid. Basically if the need or consent to store the data is no longer, the data must be erased. This does not concern every registry however, such as some government registries.

- Increased reach of regulations

The new regulations concern every organization and establishment that collects or handles personal data of subjects residing in EU.

- Penalties

Penalties can involve a fine in addition to requirements for fixing the violations, other instructions or even a ban on data processing. Monetary amount of the possible fines are determined so that they will be effective.

- Notification of breaches

The users must immediately be notified of any data breaches that occur in the system that handles the data, right after it has been discovered. They must also be informed on the severity of the breach with information such as how their private data is affected and what data the attackers were possibly able to access.

- Privacy by design

Data privacy must be taken into account when designing and developing the system. The legislation however leaves it open on what measures must be taken to achieve this.

- Data protection officer

In case the main business of the organization requires handling sensitive personal data, especially in large scale, there must be an appointed Data Protection Officer (DPO). The duties of the DPO include making sure the organization and systems under it are following the privacy regulations, increasing awareness to other employees and training them.

This list was a short summary on the biggest changes that are in the regulation. It is clear that they are mostly there to protect the rights of individual subjects and adopting these changes was financially heavy for corporations, but they could also be seen as protecting corporations from possible lawsuits. Overall there is now a bigger financial incentive for companies to protect their user's data.

The regulation also does not explicitly disallow the usage of personal data for the purpose of analyzing or other secondary uses which might be worthwhile, but it does require protection of said data. The regulation suggests using pseudonymization and de-identification of the data to comply with some of the regulation. [12]

The registry controller, that is the organization that collects the private information, should use the information they collect only in a way that they have received consent for. It is also the responsibility of the controller to oversee that the data is processed according to the regulation. They are also responsible for applying necessary protection measures to comply with data protection policies. [13]

Processors are those who actually use the information the controller collects. They receive instructions from the controller on how the data can be processed, and are under confidentiality agreement. They are also under the same obligations concerning the

privacy of the data subject, and are required to delete or return the information to the controller when required by them. [14]

Each country under the GDPR has to appoint one or more authorities whose responsibilities are monitoring that the regulation is followed. These authorities will act independently in enforcing the regulation. However in the case there are multiple authorities, there will be a supervisory authority who will represent and monitor the other authorities. [15]



# 5 Detailed description of the problems

## 5.1 Data collection

Since there is still a decent amount of data collecting and processing done by hand using paper forms and other types of mediums that require transferring the information to an electrical form, it is necessary to look into how that information can more reliably be entered and submitted. Problems arise because there are many different ways the information can be collected and entered wrongly, and usually by many different researchers and organizations that can have different methods and customs. Problems can also manifest after the data has been collected and submitted to the eCRF system.

While it can be inevitable that mistakes will happen when collecting data by hand and later digitizing it, the likelihood of those mistakes happening can be lowered. The same goes for errors later when the data is exported from the system.

The focus especially in this case is data gathered at mostly Finnish hospitals and entered into the case system. This data consists mostly of information gathered from patient records, which have been pseudonymized for research uses, mainly clinical studies in these examples. Because the data is used in such studies it becomes even more important for the data to be as correct as possible. After all it is possible that the studies based on the data can be the basis for changes in some healthcare or medication methods. For example the eCRF system could be used to collect information on patients that are using a new drug of some sort, the study could be focused on how the new drug compares to the old

one. The result of the study could have an effect on how the patients are medicated.

The data is collected in different towns and sometimes even different countries and the customs for collecting and storing the data can be quite different. For unity in collecting this data, there would be an eCRF data collection manual that outlines what information is collected and how.

The information in patient records come from many different sources that when put together form a better picture of the patient's health. Such information and sources include patient baseline information like height and weight and laboratory results. More examples of possible sources can be seen in Figure 5.1. This information can usually be collected locally with reasonable ease, as these sources usually work together and the infrastructure is in place.



Figure 5.1: Researcher collecting data from different information sources

Considering the Figure 5.1, one can see that the information comes from many different sources and it would not be strange for there to be some problems when there are many types of sources for the information. Eventually, at least locally, these problems would be solved. Recently the case system has been used in the background while the

researcher accesses different databases to get the information they need about a patient, and then enters that information straight into the case system, section by section.

Receiving patient records from other hospitals, as seen in Figure 5.2, can introduce more problems with data transfer and integrity. More so if the data travels through multiple institutions. With the example system discussed later in this thesis, every hospital enters their own patient records to the system, so that the records are collected in one place, ready to be analyzed eventually. This way data entry can at least be controlled somewhat, by formatting the eCRF.

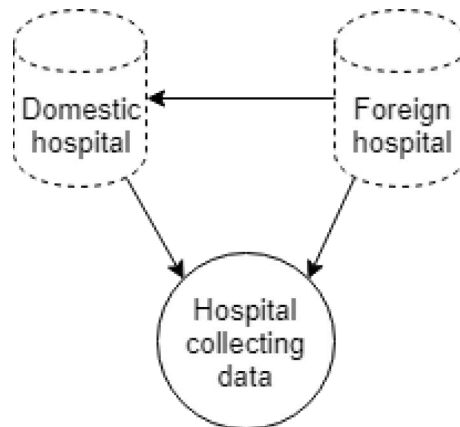


Figure 5.2: Records from different hospitals

As discussed earlier patient records are comprised of a good deal of different information, results, diagnoses etc. Report forms used for research studies are usually made custom for the given study, having only fields or information that the researchers deem relevant to the study. For example if the study was concerning the effects of some specific medication in relation to a disease or operation, most of the questions would be about that.

Seen in Figure 5.3 is an example of an eCRF structure. It might not seem too complicated, but in reality it could have around 200-300 fields of different questions, results and other information, per patient. If one collects information from 100 patients, usually there

is a lot more, one would already have around 3000 fields of all sorts of data. Considering how much data there can be when it has all been collected, shifting through the data and making sure that all of it is correct, is not really fun, though sometimes necessary. Therefore checking the data when it is being entered should help out later when data integrity is verified. In the example system, the data is entered by hand through a form, which can be tedious and long forms can increase the risk of something being entered incorrectly, e.g. 0,01 instead of 0,001.

Document structure
Baseline information
Existing medication and possible laboratory results
Pre-operation information
Operation details
Post-operation results
Medication after operation
Follow-up information

Figure 5.3: Example of an eCRF document structure

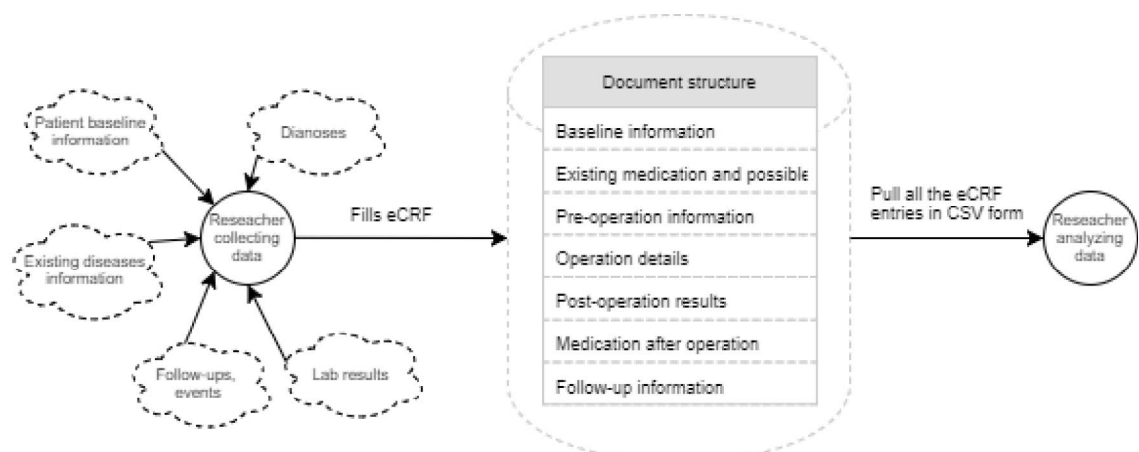


Figure 5.4: Full data flow in the case system

Figure 5.4 shows the whole simplified data flow of the case system. The data is first

collected from various sources mentioned before and saved in a specific format which the form is used for. Later the data is exported for analytic purposes.

### **Data collection in the case system**

As the data is collected possibly from many different systems and sources, the researcher usually opens up the form that the eCRF system produces and starts entering the information one source at a time. The data is usually fetched with the patients social security number (SSN) or other identifying information from the source systems, or from paper forms. The SSN is not entered to the eCRF however, as the eCRF generates an individual patient number for the patient, so that they can later be identified by the researchers. The researchers would keep a patient identification log that they can use to know which patient number in the study data corresponds to which patient, so that they can later update their information after events such as follow-ups. This log is kept separate from the research database.

Some problems that can be present involve entering this data and moving from source to source. Some information or fields can be left empty by accident or if the information is missing and can later be assumed that the data did not exist in the first place, even if the researcher only missed it. More problems come from the data format. Information such as numbers, dates and laboratory results can be in different formats in different sources. While making the eCRF it needs to be determined in what format this type of data is expected. When designing the form and the questions, it is better to leave as little as possible for the researcher to enter manually, but some information such as laboratory results have to be entered as text.

The form could also include calculations or scores for different risk assessment scores, such as the score for atrial fibrillation stroke risk (CHA<sub>2</sub>DS<sub>2</sub>-VASc score) where patients are given a score on how likely it is they might suffer a stroke. This particular score is based on variables such as the patient's age, gender and some existing diseases [16].

These calculators generally monitor the form while it is being filled and update the score accordingly. For instance if there was indication of a previous stroke in a yes/no question, it would add one point to the score. The scores are usually calculated dynamically, updating the score in real time, but in the past they have also been calculated when the data is exported.

## 5.2 Data extraction

Once all the data is collected it is exported from the eCRF system and analyzed to answer the research question. The data can also be exported during the collection phase to look at how the collection is going, how the data is formed and maybe to get preliminary results for the research.

In the case system the data is pulled from the database and exported into a Comma Separated Values (CSV) file. It is basically a text file where every line represents one patient and the line contains all the values in the database for that patient, separated by a comma or other specified character. In this case a semicolon is used as some of the text fields in the records could include a comma. Optionally, in the case system, the first line contains headers that specify which what values are in question in any given field. Since CSV files are fairly common, most spreadsheet software can import CSV files to a spreadsheet and most detect automatically which delimiters are used to separate the fields.

### **Data extraction in the case system**

The second problem is how the data is handled in the eCRF system and database, if at all, and when the data is exported so that it can be used as it is intended to be used for the given study. The data could also have its uses in other studies as well, when the initial study it was collected for has concluded.

The data is exported in CSV format and it is formed by iterating through the database's table of patients for the given study center/location, or all of them if needed. The CSV

format is used so that the exported file can easily be opened in spreadsheet programs as well as programs such as SPSS which is a statistical analysis software.

The form saves the data in literal form to the database and the data needs to be changed to a more usable form. How much of the data needs to be formatted depends on the study and preferences of the researchers. Usually at least yes/no answers are changed to 1/0 and dates are changed to DD.MM.YYYY format. Also empty or '0' answers are changed to dots (.) as that is least used in SPSS to signify empty data.

There have been some problems with the CSV format not working with all SPSS versions and it needs to be researched whether it would be possible for the eCRF to export straight to SPSS used file format.

Some other problems have recently been discovered with how the form saves empty data to the database and how it cannot be easily distinguished from data where the actual value is 0, when exported.

### **5.3 Data privacy according to GDPR**

The information entered into the case system is pseudonymized. There is never even any fields that the researcher could enter such information as social security numbers, names, addresses or any types of information that would immediately single out a patient. However it is a concern that some possibly identifiable user data is saved at some point in the system. Therefore it is reasonable to research what the new regulations specify about what kind of information can be collected. The goal is being able to collect all the relevant information that is needed for the study at hand, at the same time protecting the patients' and users' information and anonymity.

### **Data privacy in the case system**

Third problem comes with the new data protection regulations, the GDPR. The health records used in secondary purposes, at least from the perspective of this eCRF, are reasonably pseudonymized even before they are entered to the system as no directly identifying data is entered to the eCRF. However some information from the users, that is the researchers, is saved in the user database when the user is logged in. It should be explored what data is saved and whether something needs to be done about, in regards to the regulations.

In summary the problem here is to figure out how GDPR might impact the current system and find solutions for complying with the regulations.

## **5.4 Summary of problems and concerns**

This is a summary of the problems and concerns mentioned above. It is meant to make it easier to refer to these problems in the next chapter when we explore different solutions to these problems. Problems will be marked in order in which they naturally can manifest during data collection and usage of said data. Problems are marked P1, P2 and P3.

### **P1: Data collection forms**

The form needs to be more clear to the user about which fields in the eCRF have been filled and which have not, because especially in very long forms it could be easy to miss some fields. Some fields also require the value to be in a specific format, so it should be made clear to the user how the value is correctly formatted and entered.



**P2: Data exporting**

The data exported from the case eCRF system needs to be formatted in such a way that it can be further used in different software used to analyze the data for the studies. There have been some problems with some of the form values getting lost when exporting due to limitations in the back-end. If the current export format does not support all analyzing software it should be explored whether adding more formats would help with this.

**P3: Data privacy**

Find out what restrictions the new privacy regulations pose on the system and find out how the data collection should be changed especially concerning potentially identifying information. There have also been some talks of moving the servers to a location that is more under the hospital's control.

## **6 Case eCRF system**

In this chapter we will look at solutions to problems mentioned in Chapter 5. These problems deal with manually inputting patient record data into an eCRF, extracting all the relevant data from every patient from the eCRF and how the system can be made to follow the new privacy regulations. First we will explore solutions to the problems that exist when entering the health record information for secondary use to the eCRF system, and take two of the biggest problems with data entry currently and find solutions to those. The second part will focus on how the research data must be formatted or handled and exported so that it can be used in the healthcare studies in question, more specifically how exporting the data can be improved so that it will work with the data analyzing software. Lastly we will look at solutions to the possible problems that the GDPR might cause to this eCRF system and its usage.

### **6.1 P1 - Data collection forms**

The problem with data collecting with electronic forms can be mostly solved with a technique called form validation, where the information filled in the form is checked as it is entered and then validated that it matches the agreed format or style. The eCRF in question is browser based and these kinds of validations are usually done dynamically using JavaScript, which executes the scripts locally in the browser. Some of the validations can also be done when the form itself is submitted, but mostly it is more efficient to do it while the form is being filled. The goal with P1 is to check the data as it is being filled. More

specific goals are listed later. [17]

The eCRF system itself is quite old and mainly programmed using PHP: Hypertext Pre-processor (PHP) with a MySQL for the database solution. PHP is used for database connections and creating the forms from the data fetched from the database and submitting filled form data back to the database. Hyper Text Markup Language (HTML) is used to actually display the forms and other web page elements such as the menus. Cascading Style Sheets (CSS) are used for web page styles. Considering PHP for processing the form data one has to know that the PHP code is executed on the server side and therefore can only be used to evaluate the data when it has already been submitted to the server. Using only PHP, especially without any supporting libraries, it becomes impossible to produce modern dynamic and responsive web pages. The case eCRF was developed in 2009 and has had some updates done to it when needed. [18]

Over time there has been some need to add more dynamic functionality to these forms. Functionality such as automatic calculations for different scores like ‘Major Bleeding Risk’ score (HAS-BLED) or ‘Bleeding Risk in Atrial Fibrillation’ score (OBRI). With these scores it was needed for those scores to update as the form was being filled. This sort of active updating the view of the web page would have been difficult with PHP and so JavaScript was used as it makes it easy to evaluate and update the form information on the clients end, that is the browser, and no data needs to be sent to the server until it is ready to be submitted. It would have been possible to calculate the score after the information was submitted and then display it to the researcher, but having it update the score immediately also decreases the chance of wrong data being submitted as the researcher can monitor the score to see whether it matches what can be expected from the data that has been entered on the form.

Many modern dynamic websites are implemented using JavaScript frameworks where JavaScript is the main language used for the whole programming stack, that is front-end, middleware and back-end, supported by extensive libraries. Somewhat recent tools

and frameworks have made JavaScript arguably the most used language for making web applications, and easy to deploy modules provide a lot of functionality.

As stated earlier PHP is unfit for modern dynamic pages and it is better used for back-end data manipulation such as database connections and data formatting. Since the PHP part of the system provides most of the functionality, to use a newer JavaScript framework would require rewriting the code rather than just implementing new features.

To add more features on top of our existing software we will use JavaScript, but we can add simple JavaScript script files that are loaded when needed to support the existing site. While some of the older scripts have been programmed using pure JavaScript, JavaScript itself is not that simple, or easy, to use so we will use jQuery as a supporting library. jQuery makes the script a lot easier and efficient to write and it also makes it easier to select different area and component ids and classes throughout our forms, and execute code and give different styles to different areas. With jQuery you can easily assign custom CSS style definitions to specific elements, and you can add different logics to these scripts on how and when these styles are applied.

We will break P1 into two distinct issues. First is **checking that all the fields are filled in the form**. Second is **numeric value format checking**. It will also be limited to these as these are the two biggest problems or nuisances in the form currently. Other reason is that the form is designed to be simple to fill and contains a lot of values that are entered simply by clicking. Some of the very few text entries that might exist in these forms usually contain additional information about various subjects, and would be hard to validate.

### **Form validation - Checking that fields are filled**

The first part of form validation problem we will look at and solve is signaling the person filling the form about empty fields. The goal is to make it difficult to miss empty fields, as when there can be over a hundred questions or fields in one form, it becomes easy to accidentally miss one or two. This should also help researchers who later come

back to the form to fill out missing information, and people in charge of validating that information. We will highlight unfilled fields with a noticeable but not distracting color. The main colors of the form are black text on white background and even minor color changes would be easily noticeable, as seen in Figure 6.1 which shows a small part of an eCRF where patient background information is asked. We will be using this part of the form to demonstrate most the changes that are made to the whole form regarding form validation.

<b>Patient number</b>	<input type="text"/>	<b>Test group</b>	<input type="text"/>
<b>Gender</b>	Female <input type="radio"/> Male <input type="radio"/>		
<b>Year of birth</b>	<input type="text"/>	<b>Month of birth</b>	<input type="text"/>
<b>Additional information</b>	<input type="text"/>		
<b>Status</b>	Diagnosed <input type="checkbox"/> Discharged <input type="checkbox"/>		

Figure 6.1: Part of the eCRF without modifications

Basically what we need to do for this is create a script that highlights the background or the input element with a different color if there is no value inputted or if no selection has been made.

Each of these fields in this form have their own element inside a HTML div tag, which in this case is used to group these elements and allocate some space on the page for them. The fields need to have a unique class identifier, which requires a small change to how the HTML tags for the fields are generated. We will assign a running number to all the class names to that they can be easily identified by jQuery scripts.

In jQuery there are wildcard selectors that can be used to match any div element and then fetch the specific class identifier so that we can apply changes so that specific element. We will want to have jQuery to monitor any changes that might happen in the form, but then apply the changes to the changed element.

```
1 $(':input')...
2 $(':radio')...
3 $(':checkbox')...
4 $('select')...
5 .on('input', function() { ... })
6 .on('change', function() { ... })
7 .parent()
8 .val()
9 .css( ... style definitions here ... )
```

Figure 6.2: jQuery selectors and functions

In Figure 6.2 we can see some basic jQuery selectors and functions which will be used later when applying changes discussed earlier. From first to last, with these selectors we can:

1. Select all the inputs in the form, such as text fields and text areas
2. Select all radio input fields
3. Select all checkbox fields
4. Select all dropdown input fields
5. Monitor when the input value is inserted, for example when typing information in a text field
6. Monitor when a value is changed inside a selector
7. Select the parent element / class of a selector
8. Get the value inside a selector, such as text in a text field

### 9. Change the styles of a class, such as background color

Alone these selectors and functions will not accomplish much, as they have to be combined to achieve monitoring the form and applying changes. Quite simply though, we can take the input selector, combine it with the ‘on input’ function, fetch the inputs parent container and add a CSS change to it. There needs to be some other logic as well, since just changing the background color to one color will not change it back if the text is later erased, but simple if-else statements are enough. The logic code will look mostly the same for every type of input field, though some small changes are needed.

Before we add these functions to the input fields, we need to add some code that is executed when the page is loaded, in the beginning of the script which changes the background colors of each field to the desired color. We will be using light but noticeable red as the ‘not filled’ color. While the background colors could also be defined using the CSS file, it will not work when opening the form later if some of the form has already been filled. This part of the script needs to go through every input field in the form and change the background color of those that contain no value.

In Figure 6.3, we can see the code that accomplishes changing background color for fields that contain no information. The first line specifies that the following code block is executed when the document is ready, meaning when the page has finished loading. Second line selects all the input fields in the form, fetches their parent elements and changes their background to the light red color. The third lines gets all the inputs fields again and iterates through them, executing the code in the next few lines. These lines search and go through those fields that have already been filled and change their parent elements background color back to white. Checkboxes and radio buttons need their own piece of code. The type of the field is determined by the ‘attr’ function which is used in an if-else clause to execute the right code for the right type of field.

On line 5 in Figure 6.3 the code for checkboxes fetches the parent element of the

```
1 $( document ).ready(function() {
2   $(':input').parent().css("background-color", "#f27d7d");
3   $(':input').each(function(){
4     if ($(this).attr('type') === 'checkbox') {
5       if ($(this).parent().find(":checked").length >= 1) {
6         $(this).parent().css("background-color", "white");
7       }
8     }
9     else if ($(this).attr('type') === 'radio') {
10      if (this.checked) {
11        $(this).parent().css("background-color", "white");
12      }
13    }
14    else {
15      if ($(this).val() && $(this).val() !== "-") {
16        $(this).parent().css("background-color", "white");
17      }
18    }
19  });
20 });
```

Figure 6.3: jQuery script which changes background color of select elements

checkbox and then searches for any checked fields within the element to see if any of the checkboxes are checked. This is because a user can check multiple checkboxes and it needs to be determined whether any of them are checked before we can reset the background color. For radio inputs it is easier as only one of the radio inputs can be checked in an element, so we do not have to search for other checked input within the element. For any other type of field it is enough to check if there is any value entered, or if the value is not '-' for dropdown selections which is used to mark not answered or empty in these forms.

Next is the actual form monitoring for field value changes. It will be somewhat similar to the earlier code, but the code parts are executed when the triggers are activated. For different types of fields, there are different triggers like 'on change' and 'on input' as seen in Figure 6.2.



```
1  $("input").keyup(function() {
2      if ($(this).val()) {
3          $(this).parent().css("background-color", "white");
4      } else {
5          $(this).parent().css("background-color", "#f27d7d");
6      }
7  });
8
9  $("radio").on('change', function() {
10     if (this.checked) {
11         $(this).parent().css("background-color", "white");
12     } else {
13         $(this).parent().css("background-color", "#f27d7d");
14     }
15 });
16
17 $("checkbox").on('change', function() {
18     if (this.checked || $(this).parent().find(":checked").length >= 1) {
19         $(this).parent().css("background-color", "white");
20     } else {
21         $(this).parent().css("background-color", "#f27d7d");
22     }
23 });
24
25 $("select").on('change', function() {
26     if ($(this).val() && $(this).val() !== "-") {
27         $(this).parent().css("background-color", "white");
28     } else {
29         $(this).parent().css("background-color", "#f27d7d");
30     }
31 });
```

Figure 6.4: Form monitoring and background color changing

First with this implementation we will define the selector that monitors the text inputs, and the code that it will execute. Note that there is some overlap with these code blocks in Figure 6.4, as well as with the code in Figure 6.3. The code could be made to be more concise but this figure shows more clearly how different input fields need some consideration on how the events and field values are handled.

Beginning on line 1 in Figure 6.4, we see the code that monitors normal text fields and executes the following code when key is released while typing. If a value exists, the background color of the parent element is changed. On line 9 there is a monitor for value changes in radio fields, if checked, the background color is changed. Line 17 has the monitoring for checkboxes, if any of the checkboxes are checked within the parent element, the background is changed. Last the on line 25 is the code for dropdown menus,

or selections. The background color is changed, if the selection has a value and that value is not '-'.

Patient number

Test group

Gender Female  Male

Year of birth

Month of birth

Additional information

Status Diagnosed  Discharged

Figure 6.5: The end result in the form

In Figure 6.5 we can see the end result of the scripts applied on the same part of the form as in Figure 6.1. This helps in distinguishing fields that have not been filled from those that have, especially in forms that might have hundreds of fields to fill. The goal mentioned earlier would be accomplished as these unfilled fields would be difficult to miss.

### **Form validation - Numeric value format checking**

This is another part of the script that adds an option for form fields to monitor what the users enters as value. The goal here is to add a parameter, which contains a pattern, to a field and the value entered to that field must match the pattern. If the value entered does not correspond to the format that is required of the value, it will display a warning message. This is only used for text input fields as for other fields the values have already been defined in the form.

The form loads the specific information about every field from the database. There it is defined what kind of fields there are in the form, in what order they are displayed and

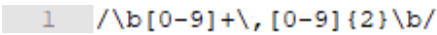
what options they might have. It would be good for the form to also get the formatting instructions from there, but there might be some challenges with getting these instructions from the database through PHP to work with jQuery. The most robust solution might also be for the instructions to be defined with regular expressions.

Regular expressions in this context are sort of instructions on how information should be formatted, or how they are formatted so that specific pieces of that information can be easily parsed from larger segments. The expression usually consists of starting and ending characters and in the middle what characters, numbers and how many and in what order they are matched against the larger piece of data. When there is a match, the expression can be used to return the matches and within them different groups of data segments. It is useful for example parsing pieces of data that contain a lot of unwanted characters.

We can use these expressions for example to instruct the form on the proper format of the value inputted in the field. On those fields that the regular expression is defined in the database, the form would match the inputted value with the expression to see that they are a match. In the case there is not a match, when the expression expects a certain pattern in the value, the form would display a warning message and the correct pattern.

Strictly in the context of this eCRF we want to monitor some numeric values that are entered. Some values such as laboratory results might require the value to be entered with say the precision of two decimals.

The numeric value fields in the form look just like the text value fields in Figure 6.1, with the occasional addition of the unit such as mg/l added next to the text box. We will start this by defining the regular expression needed to match a numeric value with two decimals.



```
1 /\b[0-9]+\., [0-9]{2}\b/
```

Figure 6.6: Regular expression to match inputted value

In Figure 6.6 we see the expression that will accomplish what we need for now, as it

was tested using several different variations of values using online tools [19].

To break it up, first there is the beginning and ending symbols (*/*). Next `\b` strings define the strict area to which the expression is matched. The string cannot be longer in any direction than specified by the expression, or it will not be a match. The rest of the expression has the actual logic for matching the value. `[0-9]` means matching any number from 0-9 with the `(+)` symbol meaning one or more. Comma symbol after that is escaped with `\` and is always expected when entering the value. Last `[0-9]{2}` means matching again numbers 0-9 and two of them are always expected.

Next to implement the expression to the form we need to add information of the expression to the database, where it is needed. A new column needs to be added to the table which stores the fields displayed in the form. We will call this column 'pattern'. After that we enter the expression for those fields that require validation and leave it empty for most other fields which do not need the expression.

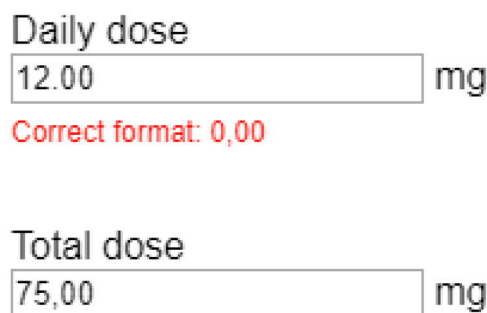
To the PHP code that generates the form field by field, we will add a piece of code that checks if the field currently being generated contains anything in the pattern part. To the fields that contain this, we will add an extra element which will display the warning message under the value field, when needed. To accomplish all this we will make the PHP code add a function call to these input field which triggers when a key is pressed in the field.

```
1 <input type="text" name="d_dose" onkeyup="matchPat('d_dose', $pattern)" />
2
3 function matchPat(inputName, pattern) {
4     var input = $('input[name='+inputName+']');
5     if (!input.val().match(pattern)) {
6         input.parent().find('.warning').show();
7     } else {
8         input.parent().find('.warning').hide();
9     }
10 }
```

Figure 6.7: HTML input tag definition and javaScript funtion

In Figure 6.7 on line 1 we can see the HTML input tag which contains the function call

for matching the value in the field. To the JavaScript/jQuery function which we trigger as we type in the field, we will send the name of the input field and the pattern we want to match. In the JavaScript function on line 4 in Figure 6.7, we find the input field that corresponds to the name we passed to the function. After that on line 5 we determine if there is a value entered to the field and whether that value matches the regular expression. If it does not then we display the warning message, on line 6 where we first find the warning element from the input's parent element and then show it. If the value matches the expression then we hide the warning message again.



Daily dose  
12.00 mg  
Correct format: 0,00

Total dose  
75,00 mg

Figure 6.8: Input fields with expression matching and warning message

Figure 6.8 shows the finished pattern checking in the form, with two input fields that have the expression matching enabled. The first field has the numeric value entered incorrectly and shows the warning message displaying the proper format. The second field has the number entered correctly, therefore there is no message. The goal was to be able to enable input value format checking and that has been achieved with these additions.

## 6.2 P2 - Data exporting

Data exporting happens when the researcher uses the option in the form to export data, when it is ready to be analyzed. The only option to export the data using the eCRF currently is exporting it as a CSV file. There are other similar options that might work better in some cases such as tab-delimited or tab-separated value file.

There is a software used to make accessing the database easier and using this one can also export the database in Structured Query Language (SQL) format. Though this function is currently reserved only for administration uses and most likely would not be practical for research use, as the data should be formatted for analyzing purposes when exported, at least to a degree. However the SQL exporting is used for backing up the data, which does not require any kind of formatting.

The problem with data exporting is that it has been reported that CSV importing does not always work properly with SPSS, the program this data is usually analyzed with. Requiring the researcher to perform some amount of manual formatting to the data so that it works properly. The goal here is to determine why the exported data might not work in SPSS and whether the tab-delimited data might work better.

To begin analyzing the data, as it can be currently exported, we use the eCRF to fetch the data. Exporting the data with the header option enabled, where the export function adds headers with field names on the first line of the CSV file. This basically helps with determining what all the different values mean, and it should not cause problems when opening the CSV file with SPSS. The SPSS version we will be using is version 20.

Opening the CSV file in SPSS is done like opening any other file. After opening the example CSV file it would seem that it does load properly. Further analyzing the data would require more knowledge on how the program works and on data analyzing, but it seems that the headers and rows of data are loaded as they should be. That is so that the header row is at the top of the screen with all the different columns divided based on the

field name, and under the header are rows of values corresponding to the above header.

While we were not able to reproduce the problem with the tools we have available, it would still be helpful to add more formats in which the data can be exported from the eCRF. Adding a feature to export the data in tab-delimited format should be somewhat simple as it should be accomplished just by duplicating and modifying the PHP file which is used to export the data. After copying the file, we will change the settings used to define which characters are used to format the essentially text file. Actually the CSV file the eCRF exports is not separated by commas, but semi-colons and new lines are defined by new-line characters (`\n`). Making the export use tabs instead of semi-colons is simply just changing the semicolon to tab (`\t`) in the code.

```
1 $csv_terminated = "\n";  
2 $csv_separator = ";";  
3 $csv_enclosed = '';
```

Figure 6.9: CSV delimited definitions in the eCRF

In Figure 6.9 we can see the CSV definitions in the code. They are defined like this so that they are easy to change if needed. On lines 1-2 we can see the characters that separate the fields and lines from each other. On line 3 we can see that values or singular fields are enclosed with quotes so that for free text fields there is no confusion on where one field begins and one starts.

For the tab-delimited option we will add a new link to the main menu of the eCRF which will be used to load the exported file. The link will load the copied and modified CSV exporter. With this feature added to the main menu, we will load the tab-delimited file and load it in SPSS the same way we loaded the CSV file. Opening the new file format however seems to produce results no different than loading the old CSV format, which would suggest that changing the file format or the data this way does not solve the problem. Essentially this file format could be considered very close to the CSV format.

As the data would seem to load fine on SPSS version 20, it would suggest that the inability to load the data would have happened on another version of the software, or caused by some kind of other error. Adding more file formats to be able to choose from might help in some cases, but in this case adding the tab-delimited option seems unlikely to help. This is because the files are still quite similar in how they are constructed, and CSV format being the more popular or supported one of these. SPSS does also support importing spreadsheet files such as Microsoft Excel files XLS or XLSX. The CSV file could first be imported into a spreadsheet program, saved as a spreadsheet and then opened with SPSS.

The problem mentioned in the beginning of this subsection have actually not been heard of after it was mentioned the first time, and not during when the solution was being investigated. We could conclude from this that the problem does not critically hinder the analyzing of the eCRF data for now. We could also expect to hear about this problem from more researchers if it was priority concern.

In the future it could be explored whether the form and data export could be made to support the actual SPSS '.sav' data format. Currently it would involve some bigger changes to the eCRF and the system itself, such as version upgrades which are planned to be made when the eCRF is moved to a new server and platform later. This was actually looked into around two years ago, but back then the options for including this feature were even fewer than today.



### **6.3 P3 - Data privacy under GDPR**

This eCRF system acts as a processor for the patient data, and the databases and other sources where the data is collected from are the actual registry controllers. These controllers include sources such as patient data registry and laboratory results database, where the data is linked to actual patients. The permissions to collect this data in the eCRF system is handled before the actual collection starts.

#### **Patient information**

As the data processor, case system's primary function is to collect, store and combine patient data used for healthcare study purposes. This data when collected is already pseudonymized by assigning the patient a unique identification number related only to the system, and it is assigned when the collection starts for every patient. In a case of a registry study, which the eCRF is mainly used for, a special permission is applied for and which is granted by relevant authorities. No actually identifiable data, such as names or social security numbers, is collected. The identification number is used to relate the patient to the data, but the means to do that are kept separate from the system.

While the data collected has been pseudonymized, there have been extra steps taken to further limit what data is collected and to protect the patient's privacy. Some potentially troublesome information have been limited, such as birthdays limited to the year and data collection locations hidden from the system.

At the moment there is no more that needs to be done to the system when it comes to patient data, as the system only processes pseudonymized patient data.

### **Researcher information**

When it comes to user accounts, the case system only saves the IP address from the users, that is the researchers or other data entry personnel. Most of the actual information that is dealt with comes from patient registers.

The users log in to the system with a username and password combination that have been generated for them by the system administrator, and generated so that no identifying data is used. The system creates a session for every user that is logged in, so that the researchers do not have to authenticate themselves every time they load any of the pages. The session and its validity is checked every time any of the pages in the eCRF are loaded. The session is saved in the database as well as a session cookie in the user's browser and when loading the page it is compared against the corresponding entry on the database.

For security reasons however the application saves the IP address of the user, so that even if someone would manage to copy the session from the user's browser, it would be limited where they could use it. The session cookie is also set to expire after a given time to further improve security.

It should be explored whether the user should be informed when their IP address is saved, and about storing the session cookies on to the user's browser. The IP address could be defined as sort of private information which could be used to identify a user, but in most cases it is not something that can identify a specific user. There would have to be a disclaimer telling the user that their IP address is saved every time they log in, which they can either accept or cease using the form.

According to the GDPR the IP address is considered user's identifiable private data if a specific person can be identified based on the IP address. As no other information is stored about the user and the user account could be used by more than one person or researcher, usually from a public place, it should be quite unlikely to be able to identify a person just based on that. Adding to this the article 6 of the GDPR mentions also that private information on the user can be temporarily saved without consent, if the data is

used to ensure the integrity and confidentiality of the system [11]. In the case system this is exactly why the IP address is saved. To better ensure the confidentiality of the data and not allow access to those who are not supposed to have access. The user data also does not form its own registry, but is part of the registry with the patient data.

When the research data is eventually exported from the system, when the collection is complete, the user data is never exported along with the patient data. To further increase the security of the data, data like the IP address could also be encrypted when it is saved in the database.

To summarize, there is no need to change the system when it comes to user information as only the IP address is collected and that is temporarily used to improve security in the system.

In the future there are plans to move the eCRF's data and other functions to a service provider which oversees other hospital systems and data pools. This should ensure better security for the data and clearer picture of how the security is implemented in the system and who has access to it in the end. The researchers are also encouraged to include their initials when submitting patient data and it could be explored whether this is necessary.

# 7 Conclusions

We looked at few different problems in an existing eCRF and came up with solutions for improving data gathering and looked at how the new privacy regulations affect the form and the system it is in. While most the the problems were solved, the one with data extraction was left unsolved but improved.

## 7.1 P1 - Data collection

Highlighting unanswered questions in the eCRF can make filling and updating patient information clearer, especially in longer and more complicated forms. Looking at Figure 6.5, we can see how the highlighting makes the unanswered fields stand out from the rest. Considering these forms can have hundreds of questions just for the base information, adding to that even more questions that are answered during follow-ups, we can conclude that it could be easy to miss some of the questions and the highlighting helps in that regard. Especially when returning later to fill the information, or if multiple people fill the information for the same patient. The feature was also somewhat easy to include in the program as the tools needed were already used to some degree.

With multiple different numeric values that might be required to be entered on the form, there needs to be at least some information on how those values are expected to be entered. Displaying the warning text, when the value is not what or how it is required to be, should at least help with formatting the data when there are many different people

entering information and when the data can also differ from patient to patient. This kind of data validation could potentially also be done when sending information to the server, so that information in wrong formats are not saved in the database. However the questions in the forms are mainly selections and typing values is not as common. As it is in the case system now, it is more of a suggestion or outline on how the data should be entered, but it has not been a very big problem in the past either. Usually the data is also checked and verified by the researchers before analyzing it.

Form validation as a whole improves how data is entered and prevents unwanted or wrongly formatted data to be entered to the database. On contrast form validation can also help to gather especially wanted information, or at least give more attention to it.

## **7.2 P2 - Data exporting**

For the data extraction part we can not really conclude that we came up with a specific answer to the problem of SPSS not being able to parse the research data from the CSV file. We added the support for extracting the data in tab-delimited format, but after further research it would appear to function almost in the same way as the CSV format file.

It could be that the problem exists in an older version of SPSS which in a way could solve itself eventually, or it could be that the problem was somekind of a temporary problem with the data extraction. As there has been no further reports of this error existing, it should not affect a wide range of users.

In the future it could be explored whether the support for extracting the data in native SPSS format could be included.

### 7.3 P3 - GDPR

The eCRF system was designed from the start to only collect relevant data for healthcare studies. It was however good to review exactly what kind of data is collected.

As a data processor the eCRF processes the data collected by actual registry controllers. The data is pseudonymized when entered into the system. Because of this, there is no need to change anything to how patient data is collected, for now.

There was nothing to be changed in the form when it comes to the users of the system. As there is no personal user account and the only data the form collects on its users is their IP address. The eCRF is also private in a way that only those who need have access to it and the user accounts generated are anonymous. The form also saves a cookie that is used to save the session for the user so that they do not have to login every time they load a page. Collecting the IP address from the user for security reasons is permitted by GDPR and it was also good to see that other types of logging that happens in the server is allowed.

In the future there are plans to implement a real user account manager and login page and that will most likely need more research on GDPR as well, but that will come after system is moved to a new location.

## References

- [1] C. Pagliari, D. Detmer, P. Singleton, "Potential of electronic personal health records", *BMJ: British Medical Journal*, vol. 335, pp. 330-333, Aug 2007
- [2] P. Tang, J. Ash, D. Bates, J. Overhage, D. Sands, "Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption", *Journal of the American Medical Informatics Association*, vol. 13, pp. 121–126, Mar 2006
- [3] S. Vainiomäki, H. Hyppönen, J. Kaipio, J. Reponen, J. Vänskä, T. Lääveri, "Potilastietojärjestelmät tuotemerkeittään arvioituna vuonna 2014", *Suomen Lääkärilehti*, vol. 69, pp. 3361-3371, 2014
- [4] Efficca, Pegasos ja Uranus jäävät historiaan, 2014, Accessed: 2018, October, [Online], Available: <https://www.medi uutiset.fi/uutisarkisto/effica-pegasos-ja-uranus-jaavat-historiaan-6082659>
- [5] H. Hyppönen, J. Hyry, K. Valta, S. Ahlgren, "Sosiaali- ja terveydenhuollon sähköinen asiointi - Kansalaisten kokemukset ja tarpeet", Terveyden ja hyvinvoinnin laitos, 2014
- [6] R. Dolin, L. Alschuler, "Approaching semantic interoperability in Health Level Seven", *Journal of the American Medical Informatics Association*, vol. 18, pp. 99-103, Nov 2010

- [7] P. Coorevits, M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, G. De Moor, D. Kalra, "Electronic health records: new opportunities for clinical research", *Journal for Internal Medicine*, vol. 275, pp. 547-560, Dec 2013
- [8] General Data Protection Regulation, Accessed: 2018, October, [Online], Available: <https://gdpr-info.eu/>
- [9] W. G. Voss, "First the GDPR, Now the Proposed ePrivacy Regulation", *Journal of Internet Law*, vol. 21, pp 3-11, Jul 2017
- [10] General Data Protection Regulation Key Issues, Accessed: 2018, October, [Online], Available: <https://gdpr-info.eu/issues/>
- [11] Art. 6 GDPR - Lawfulness of processing, Accessed: 2018, November, [Online], Available: <https://gdpr-info.eu/art-6-gdpr/>
- [12] M. Hintze, G. LaFever, "Meeting Upcoming GDPR Requirements While Maximizing the Full Value of Data Analytics", 2017 March, Accessed 2018, October, [Online], Available: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2927540](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2927540)
- [13] Art. 24 GDPR - Responsibility of the controller, Accessed: 2018, November, [Online], Available: <https://gdpr-info.eu/art-24-gdpr/>
- [14] Art. 28 GDPR - Processor, Accessed: 2018, November, [Online], Available: <https://gdpr-info.eu/art-28-gdpr/>
- [15] Art. 51 GDPR - Supervisory authority, Accessed: 2018, November, [Online], Available: <https://gdpr-info.eu/art-51-gdpr/>
- [16] H. van den Ham, O. Klungel, D. Singer, H. Leufkens, T. van Staa, "Comparative Performance of ATRIA, CHADS2, and CHA2DS2-VASc Risk Scores Predicting



Stroke in Patients With Atrial Fibrillation: Results From a National Primary Care Database”, *Journal of the American College of Cardiology*, vol. 66, pp. 1851-1859, Oct 2015

- [17] Form Data Validation, Accessed: 2018, October, [Online], Available: [https://developer.mozilla.org/en-US/docs/Learn/HTML/Forms/Form\\_validation](https://developer.mozilla.org/en-US/docs/Learn/HTML/Forms/Form_validation)
- [18] E. Aaltonen, ”Sähköinen tutkimustietojen kerääminen ja käsittely : eCRF-sovellus potilastutkimukseen”, 2010, Accessed: 2018, October, [Online], Available: <https://www.theseus.fi/handle/10024/16785>
- [19] RegExr: Learn, Build, & Test RegEx, Accessed: 2018, October, [Online], Available: <https://regexr.com/>