TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

TTU IT College

Mark Selezenev 192388IAAM

IMPROVEMENT OF DATA PROCESSING IN ESTONIAN NATIONAL E-HEALTH INFORMATION SYSTEM ON THE BASIS OF LUNG CANCER ANALYSIS

Master's thesis

Supervisor: Priit Raspel

MSc

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

TTÜ IT Kolledž

Mark Selezenev 192388IAAM

ANDMETÖÖTLUSE PARANEMINE EESTI TERVISE INFOSÜSTEEMIS KOPSUVÄHIANALÜÜSI NÄITEL

Magistritöö

Juhendaja: Priit Raspel

MSc

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Mark Selezenev

27/03/2021

Abstract

The aim of the master's thesis is to examine the current state of data processing in the Estonian Health information system and provide possible solutions on how the situation can be improved. The work will examine the current state of data processing and the possibilities where the improvement for data processing can go in the near future on the basis of lung cancer analysis. Based on that and established international methods, the author will provide recommendations on how to proceed which is in line with the current technological and legal limitations that TEHIK and other health information system developers in Estonia need to abide by.

In the master thesis business and systems analysis will be conducted on the proposed solution that will come out of the recommendation. Business analysis will cover the AS-IS and TO-BE processes, technological and legal limitations, and technological viabilities of the project. Systems analysis will cover the system requirements, use cases, architectural location within the healthcare information system as a whole and a possible prototype.

The result of the work can be used as complementary information for the Ministry of Social Affairs in conducting the procurement of additional developments for Estonian health information system.

The master's thesis is written in English and contains 87 pages of text, 6 chapters, 36 figures.

Annotatsioon

Selle magistritöö eesmärgiks on uurida andmetöötluse praegust olukorda Eesti Tervise infosüsteemis ning anda võimalikud lahendused, kuidas olukorda parandada. Töö uurib andemtöötluse olukorra tänast olekut ja võimalusi, kuhu suunas võiks andmetöötluse paranemine tulevikus liikuda kopsuvähianalüüsi näitel. Toetudes sellele, lisaks väljakujunenud rahvusvahelistele meetoditele, annab autor soovitusi kuidas edasi liikuda, võttes arvesse tänapäevaseid tehnoloogilisi- ja seaduslikke piiranguid, mida TEHIK ning teised tervise informatsioonisüsteemide arendajad Eestis järgima peavad.

Magistritöö äri- ja süsteemianalüüs viiakse läbi eelnevatest analüüsidest tulenevatest lahendustest. Äritegevuse analüüs katab AS-IS ja TO-BE protsesse, tehnilistest- ja seadustest tulenevatest piirangutest ja projekti tehnoloogilisest rakendatavusest. Süsteemianalüüs katab süsteeminõuded, kasutusmallid, tervikliku arhitektuuri asukoha tervishoiu süsteemis ja prototüübi.

Töö lõpptulemust saab kasutada kui täiendavat informatsiooni Sotsiaalministeeriumi Eesti tervise infosüsteemi täiendavate arenduste hanke läbiviimiseks.

Magistritöö on kirjutatud inglise keeles ning sisaldab teksti 87 leheküljel, 6 peatükki ja 36 joonist.

List of abbreviations

TEHIK	Tervise ja Heaolu Infosüsteemide keskus or Center for Health and Welfare Information Systems.
IT	Information Technology
ID	Identity Document
MoSCoW	Must, Should, Could, Would method
FURPS	Functionality, Usability, Reliability, Performance, Supportability.
BPMN	Business Process Model and Notation
C4 Model	The C4 model is an "abstraction-first" approach to diagramming software architecture, based upon abstractions that reflect how software architects and developers think about and build software.
HIMSS	Healthcare Information and Management Systems Society, Inc. (HIMSS) is a global advisor and thought leader supporting the transformation of the health ecosystem through information and technology.
HAAM	Healthcare Analytics Adoption Model
EDW	Enterprise Data Warehouse
EMR	Electronic Medical Records
ICD	International Classification of Diseases
CIO	Chief Information Officer

KDD

Knowledge Discovery in Databases

X-Ray

An electromagnetic wave of high energy and very short wavelength, which is able to pass through many materials opaque to light.

CT scan

A computerized tomography (CT) scan combines a series of X-ray images taken from different angles around your body and uses computer processing to create cross-sectional images (slices) of the bones, blood vessels and soft tissues inside your body. CT scan images provide more-detailed information than plain X-rays do.

PET-CT scan

Is a nuclear medicine technique which combines, in a single gantry, a positron emission tomography (PET) scanner and an x-ray computed tomography (CT) scanner, to acquire sequential images from both devices in the same session, which are combined into a single superposed (co-registered) image.

DICOM

Digital Imaging and Communications in Medicine is the standard for the communication and management of medical imaging information and related data.

PSNR

Peak signal-to-noise ratio is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed as a logarithmic quantity using the decibel scale.

MSE

Mean squared deviation of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

JPEG

JPEG stands for “Joint Photographic Experts Group”. It's a standard image format for containing lossy and compressed image data. Despite the huge reduction in file size JPEG images maintain reasonable image quality.

CSV

CSV stands for Comma Separated Values. A CSV file is a plain text file that stores tables and spreadsheet information.

PNG

PNG is a popular bitmap image format on the Internet. It is short for “Portable Graphics Format”. This format was created as an alternative of Graphics Interchange Format (GIF).

SOAP

(formerly an acronym for Simple Object Access Protocol) is a messaging protocol specification for exchanging structured information in the implementation of web services in computer networks.

XML

Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.

HL8

The HL8 Clinical Document Architecture (CDA) is an XML-based markup standard intended to specify the encoding, structure, and semantics of clinical documents for exchange.

SQL

Stands for Structured Query Language. SQL is used to communicate with a database. According to ANSI (American National Standards Institute), it is the standard language for relational database management systems.

Table of Content

Table of Figures

Introduction

The main goal of the master’s thesis is to analyse the current data processing done within the Estonian health information system and find ways on how the data processing can be improved. The outcome of this work should be able to provide answers to questions such as what the current state is, where can the systems move on from the current state and how this change can be implemented while also providing requirements and a prototype for such a system. The work can be divided in to 3 distinct parts.

The first part of the work will discuss the problem at hand, why it is important, what are the limitations and the general way the author will proceed until the goals are reached and the questions are answered. Methods and tools used can also be considered a piece of the first part of the work.

The second part of the work will discuss the theoretical part of where Estonian systems are right now and what can be done to improve them. This part will also cover the limiting factors for the choices on how the systems can be improved.

The third and the final parts will try to answer the question of how this change can be implemented on a concrete example of lung cancer analysis. These parts will consist of the business and systems analysis. The main outcomes of these parts will be answers to important technical questions, process improvements, requirements, architectural models, and a prototype.

1 Description of the problem and formulation of the assignment

This chapter of the master thesis is designed to give a general overview of the situation and an introduction to the problem. The 2nd half of the chapter will focus on the limitations and how the author envisions reaching a solution.

1.1 General Overview

Currently the sector of government provided tools for healthcare services in Estonia can be divided into 3 main groups of their users: people living in Estonia, doctors or healthcare professionals and large institutions like private and public hospitals or clinics. The most commonly used tools in this area are the two main portals for general healthcare in the country: patient portal and doctor portal. Patient portal is mostly used by the people living in Estonia to register for appointments and to check or modify their own medical documents. Doctor portal is mostly used by the doctors and other healthcare professionals along with private enterprises to keep track and register all the documents in the system so they will be available to other healthcare providers. Both of the portals are connected together using the national health information system along with supportive technologies and system that facilitates the storage and exchange of the information between the different portals and services both inside and outside the country. In total there are more than 1000 different institutions and databases that send information to the national healthcare system with 400 of them doing it on everyday basis.

The number of connected systems and databases to the Estonian health information system seems staggering. This allows the system to access a very large amounts of data stored either in the central data warehouses or outside the system.

If we look at this from a historical standpoint, then the goals of the current system have been to move the different functions in the healthcare industry from the paper format into the digital format which made management and work quicker and streamlined. It allowed to decrease the amount of manpower that is needed for the management and storage of those documents and to move the expenditure in to the healthcare itself instead of management or to decrease the spending depending on the situation.

But at the same time this created a large amount of health data that is not possible to go through for any single individual. So majority of the data is not being used for anything productive and only a small portion is being manually handled by healthcare professionals. Or as the E-Healthcare strategy for 2020 puts it:

The data volume used in the process of treatment and diagnosis is increasing, and in many cases systematic handling thereof without the aggregation and preliminary analysis of data is virtually impossible for health care professionals. Given the constantly increasing amount of health data and various treatment suggestions, also people are in a situation where they need reliable support for decision-making.[1]

1.2 Description of the problem

The current developments for our healthcare systems in Estonia in the regard to the above-mentioned systems can be described as maintenance of technology and adding of additional functionality, which is based on state policy in the IT[2] sector and the general increasing needs of the healthcare industry in general.

Although the current systems definitely were successful in the previously mentioned goals in paragraph 1.1, the technology has already moved forward compared to the times when those goals were formulated. One of the main developments in technology compared to the 1990s and 2000s is that general technology and ability of the IT industry in the sphere of processing data has improved significantly. Other private and public enterprises in Estonia and beyond are already actively employing and developing these technologies. The Estonian state has also acknowledged the need to implement such systems inside our own state services for decision making and increase in service quality:

The state has a lot of data through information systems and services, but we do not use it enough to make better policy decisions and provide better services. In the coming years, the capacity to use public sector analytics solutions will be significantly increased in order to use the data effectively for its own benefit.[3]

Health information system is a very reactive system right now. It is mostly used by all 3 categories of users when it is necessary and they do not provide any active steps to prevent diseases or to actively assist doctors or other health professionals with diagnosing and treatment. The implementation of modern data processing technologies can assist in all of those categories.

Current absence of such technologies in our systems means that our solutions will start to lag behind the alternatives as the time goes by but an inclusion of them will allow us to further increase the quality of our services in the healthcare sector and also improve upon the goals that were previously set. The 2020 E-health strategy has the following to say about decision-supporting solutions:

The goal of the measure is to improve the decision-making quality in clinical practice and health promotion and increase the efficiency of the decision-making process, actively applying the decision-supporting applications based on digital data processing in the eHealth information systems.

Decision-supporters are IT applications helping doctors, other health care professionals and individuals make clinical or health decisions by connecting the automatically aggregated health information about a person with evidence-based knowledge. The algorithms of decision-supporters consist of evidence-based knowledge that is compared to the health information of individuals and that provide different decision-making options based on that. The development of decision-supporters must ensure their correspondence to the principle of evidence-based medicine, including the treatment instructions, and the possibility of flexible consideration of the results of research constantly updated over time.[4]

1.3 Limitations

There are some limitations that need to be taken in to account while analysing and creating changes for the above-mentioned systems and portals. The author foresees and would like to generalise those limitations in to 3 categories:

1) Technological limitations of the current infrastructure – According to the Estonian state policy we should use an existing technology, that has already been developed.[5] In regard to this work this means that we should follow the

general ideas and technology behind the current infrastructure layout: ID cards, client-server architecture, X-Road, microservices and databases in between.

2) Limitation connected to data requirements for the technology – this is a two-way problem. From one side, there are limitation of what the data can be used for. Although there might be a large amount of it, it still needs to be pre-processed to become usable for analysis. From the other side, the locality of this data is important. Sometimes it is not possible to use the data due to different factors like health, eating habits, difference in height etc.

3) Limitations connected to data requirements from a legal standpoint – while the data might be there, there are legal guidelines on how you need to store the data coming from EU or Estonian laws and IT development goals on top of requirements and limitations on how you can use the data already stored.[6]

1.4 Goal and stages to be completed in the master's thesis

In this master's thesis the author will be discussing on how we can improve the current systems in regard to data processing to improve the quality of the service and decision making on the basis of the illness analysis and our existing infrastructure.

Stages to be completed in the master's thesis:

1) Conduct interviews with TEHIK current project managers, architects, and analysts to check if something similar is planned or if something similar was discussed before and analyse how ready we are for the change and where to start.

2) Check if there are any existing solutions under development or already developed in other states in and outside of Estonia.

3) More in-depth research on the limitations, especially ones coming from the first and third categories.

4) Check what data we currently have in the system and are able to use. If we are not able to use it then how the problem of permissions can be solved first

because without the data, there cannot be any solution due to the data requirements on the algorithms or neural networks.

5) Once the current state of the systems and the data situations are clarified then the author will need to make a choice on what illness to use as the basis for this work. Another question that will need to be answered related to the illness is how viable are the different data processing methods, algorithms, or neutral networks for analysing this disease.

6) Outlining the AS-IS processes of the current systems and the architecture.

7) Prioritization and analysis of the requirements and other information gathered during points 1,2,3 and 4.

8) Proposition of solutions to the problems coming out of the requirements and study of alternatives.

9) Outlining the To-BE processes in the future system and the architecture.

10) Explaining the use cases, architectural questions, and the position of the new solution in the existing healthcare IT ecosystem.

11) Prototyping the views and the solution.

12) Retrospective. Creation of the framework.

The main goal of the master's thesis is to provide a prototype solution along with requirements to such a system and a framework on how this change can be implemented in the national E-health information system by overcoming the limitations and problems to facilitate the goals outlined by the Estonian state that were discussed in chapters 1.1 and 1.2 by the author. All of this will be done on the basis of illness analysis and existing infrastructure.

2 Methods and tools

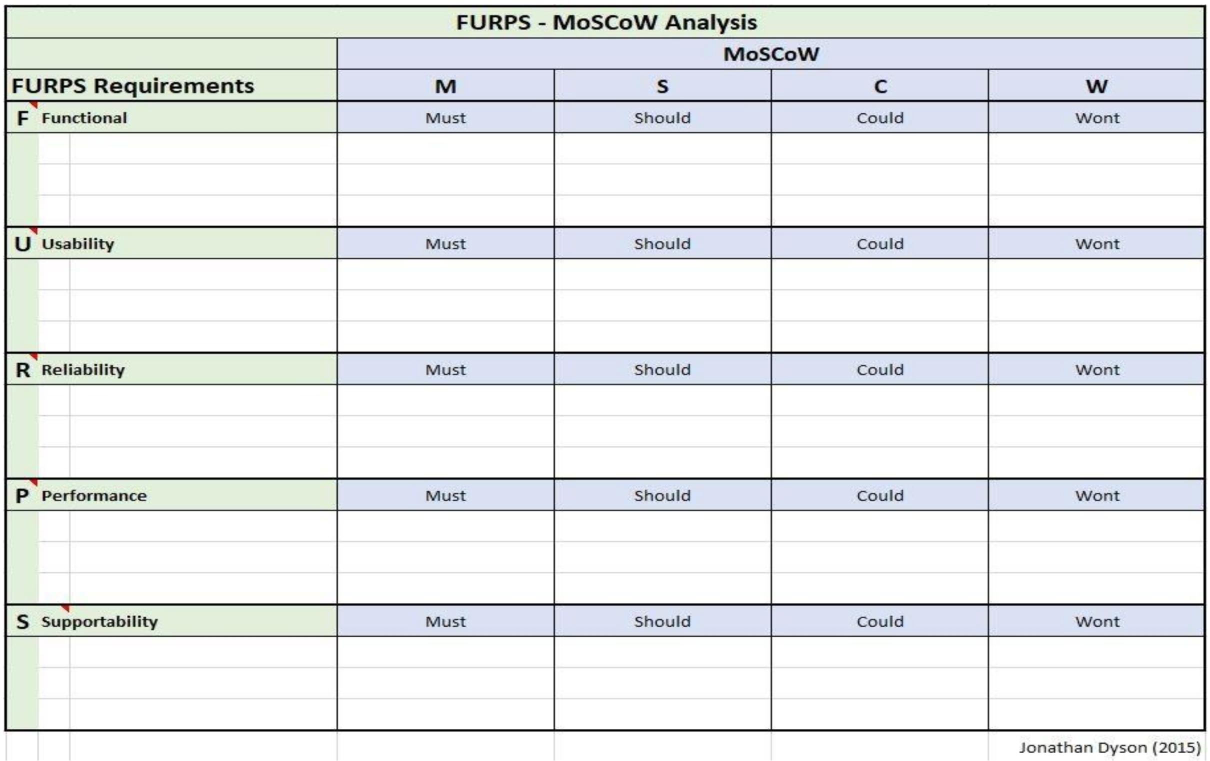
This chapter of the master thesis is designed to give an overview of the methods that will be used for analysis or to display the information as diagrams or in other formats. This chapter will also cover the specific tools the author will use to create diagrams, pictures or do analysis in general. Finally the chapter will discuss the interviewing process. This chapter will not include very specific methods that are the foundations of the 3rd chapter and require more in-depth explanations and have thesis related questions coming out of them.

2.1 Overview of the methods

One of the most important tasks that needs to be done, is to gather the requirements and prioritise them. Prioritization is important because it lets the project run smoothly and understand where to prioritise the resources. Some would say that is "more important than anything else". Without a clear prioritization, there is very few leeway what road a project can take and becomes very fragile and unrobust. Also in such cases, some requirements that might be considered "Must have" might get dropped in the process that will lead to undesirable outcomes.

So as a prioritization method the author would like to use a combined FURPS + MoSCoW method which should allow us both to structure the requirements and sort them out according to priority. The main idea behind it is to structure the requirements along the lines of both FURPS and MoSCoW.

Based on the Jonathan Dyson approach to the model, it looks something like this:



FURPS - MoSCoW Analysis				
	MoSCoW			
FURPS Requirements	M	S	C	W
F Functional	Must	Should	Could	Wont
U Usability	Must	Should	Could	Wont
R Reliability	Must	Should	Could	Wont
P Performance	Must	Should	Could	Wont
S Supportability	Must	Should	Could	Wont

Jonathan Dyson (2015)

Figure 1. FURPS – MoSCoW Analysis (Author: Jonathan Dyson, Source: https://www.linkedin.com/pulse/conjoining-furps-moscow-analyse-prioritise-jonathan-dyson)

The authors reasoning for the choice of MoSCow and FURPS is because the author believes that it will allow for a more coherent structure which goes under specific competences. This means that the author will be able to consult the proper professionals usually responsible for those categories more easily.

For process description the author decided to use BPMN 2.0. It is the global leader for these particular tasks and is easily understood by both IT and management side of the project.

Reasoning for the choice is because it is the most commonly used framework in the IT analysis field. According to the 2014 statistics it is used by more than 60% out of all Projects.[7]

For architectural description, the author will use the C4 Model.

The C4 model is an "abstraction-first" approach to diagramming software architecture, based upon abstractions that reflect how software architects and developers think about and build software. The small set of abstractions and diagram types makes the C4 model easy to learn and use.[8]

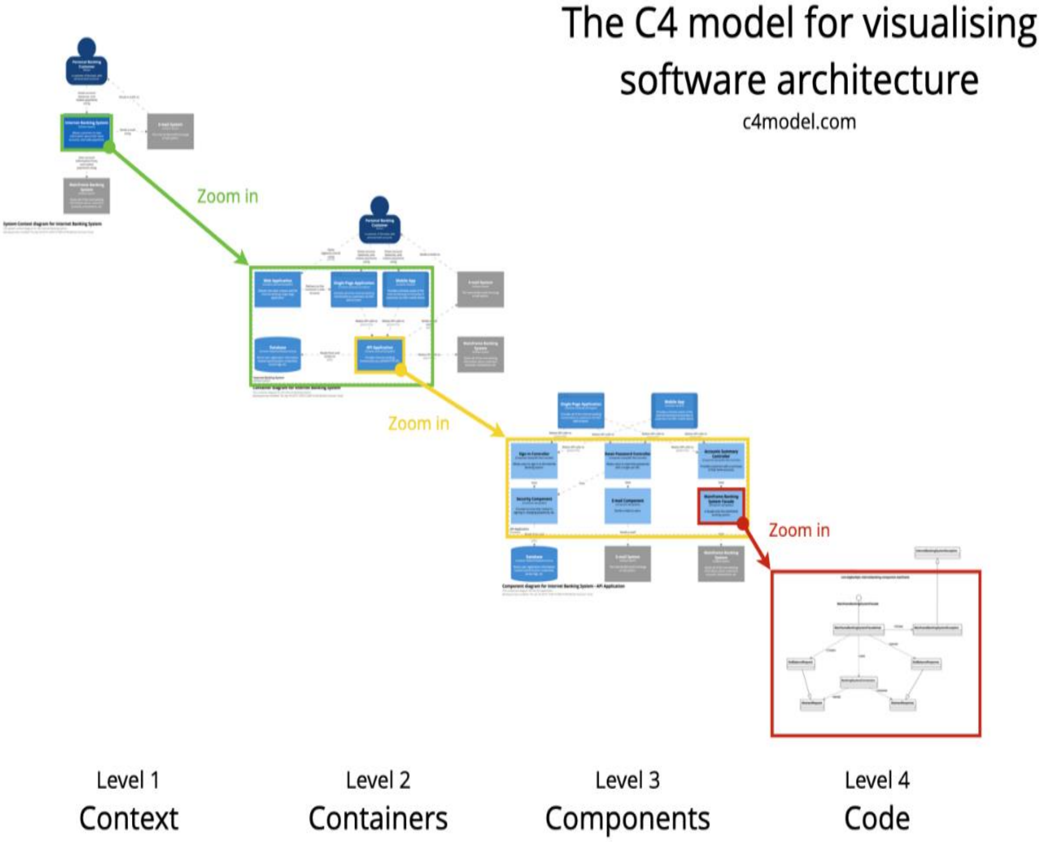


Figure 2. C4 Model for visualising software architecture (Author: Simon Brown, Source: https://c4model.com/)

Reasoning for this choice is quite similar to the choice of FURPS and MoSCoW. It allows a better communication with proper stakeholders and subdivides the architecture to different detail level according to responsibility.

For general better understanding of system interactions with the people and systems, Use Case diagram will be used. This will allow to visualise the basic flows better.

A use case diagram is usually simple. It does not show the detail of the use cases. It only summarizes some of the relationships between use cases, actors, and systems. It does not show the order in which steps are performed to achieve the goals of each use case.[9]

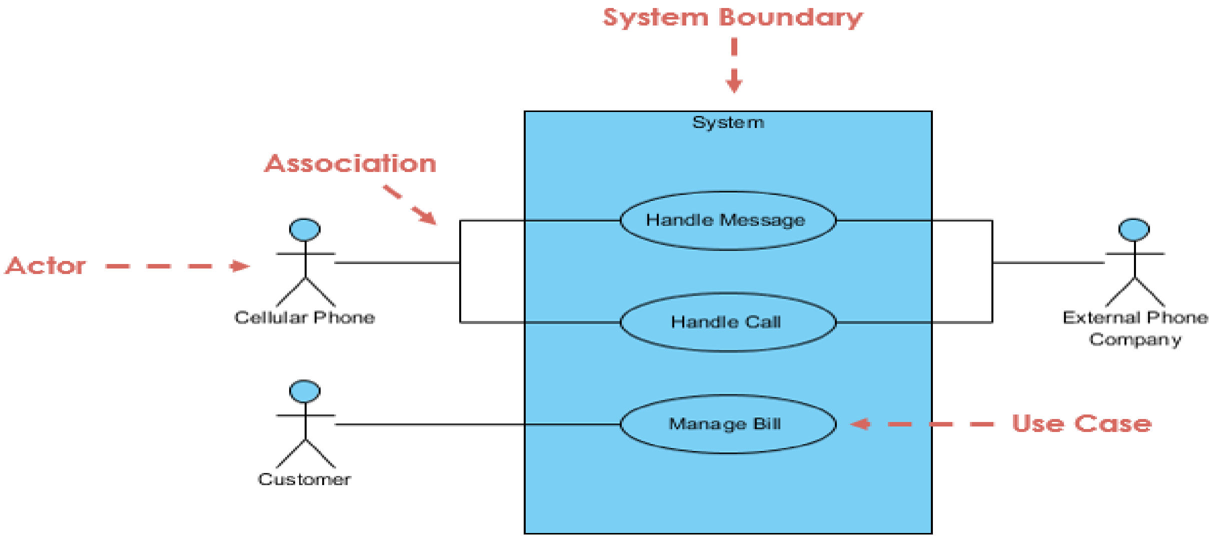


Figure 3. Use case diagram (Author: Visual paradigm, Source: https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-use-case-diagram/)

Reasoning: The main goal to use the Use Case diagram is to describe the behaviours within the system. The alternative to Use Cases is User Stories which are more descriptive of needs. The author believes FURPS and MoSCoW will be enough in that regard.

2.2 Overview of the tools

The tools that will be used are the following:

1) Star UML – Offline tool that has no limitations but has smaller variety of models to choose from.

2) Visual paradigm – Online modelling tool that is very easy to use but has limitations when using the free version.

3) Structurizr – Structurizr lets you create multiple diagrams from a single model and is used for c4 modelling.

4) Snipping tool – Tool within Windows 10 that allows to create custom pictures from any image displayed on the screen.

5) Adobe Photoshop – Used to create prototype pictures by editing existing images or creating new ones.

2.3 Overview of the interview process

Interviews will be conducted in a semi-structured personal format. The author will pre-prepare some questions to ask and will give the respondents the ability to talk freely on the subject. The interviews will be used to gather the requirements. The interviews will be conducted both with specialists in their own fields and also regular doctors and people who have used the healthcare information systems before. The question will be structured as open questions so the interviewee will have full freedom at answering them.

3 Description of E-health system readiness and the field

This chapter of the master's thesis is designed to analyse the current data processing situation in the Estonian Healthcare information systems. The first parts will cover the method used for this and the questions that arise from that method. The later parts of this chapter will also try to explain "How" the change can be conducted that comes out of the Healthcare Analytics Adoption Model.

3.1 Current state of Estonian systems

First of all, we will need to understand where we are currently located and how can we move forward theoretically. Instead of inventing the bicycle in this regard, the author wants to use an existing framework for such change: HIMSS created Healthcare Analytics Adoption Model(Onwards HAAM) in 2002 and the latest update has been in 2019.

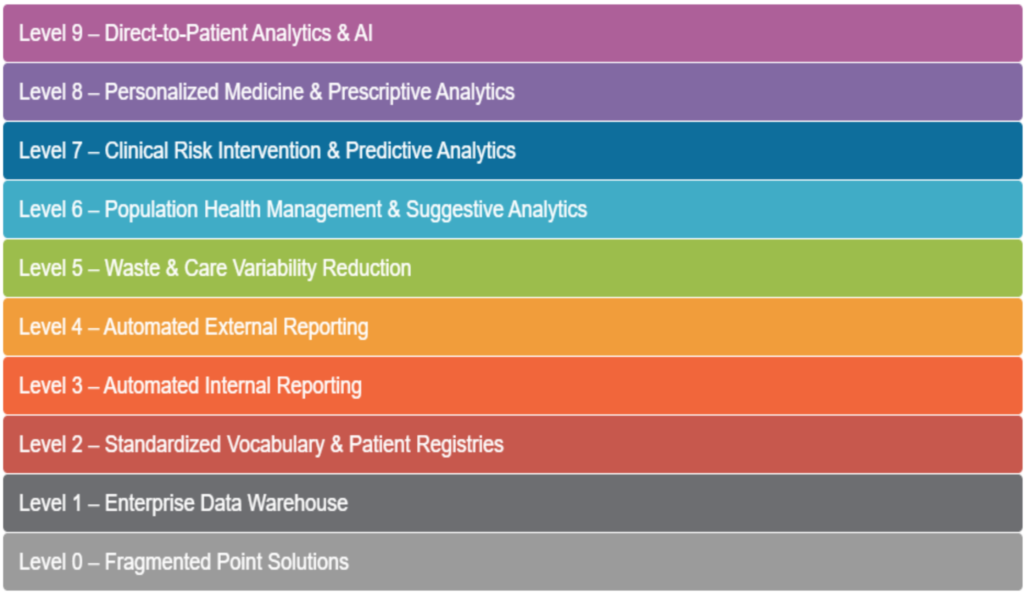
The HAAM provides three major benefits to health systems looking to grow in analytics maturity:[10]

1) A framework for evaluating the industry's adoption of analytics.

2) A roadmap to measure progress toward analytic adoption.

3) A framework for evaluating vendor products.

When health systems use the HAAM to its full potential, and follow it closely step by step, they will fully understand and leverage the capabilities of their analytics and achieve the ultimate goal that has been so hard to achieve for many organisations – to improve the quality of care while lowering costs and enhancing clinician and patient satisfaction.[11]



Level 9 – Direct-to-Patient Analytics & AI

Level 8 – Personalized Medicine & Prescriptive Analytics

Level 7 – Clinical Risk Intervention & Predictive Analytics

Level 6 – Population Health Management & Suggestive Analytics

Level 5 – Waste & Care Variability Reduction

Level 4 – Automated External Reporting

Level 3 – Automated Internal Reporting

Level 2 – Standardized Vocabulary & Patient Registries

Level 1 – Enterprise Data Warehouse

Level 0 – Fragmented Point Solutions

Figure 4 Healthcare analytics adoption model (Source: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/)

Each level in this model is defined and can be matched to our E-health information system. This model will need to be adapted a bit to fit the requirements of the state and the government- based healthcare system.

Fragmented Point Solutions or level 0 is defined the following:[12]

1) Vendor-based and internally developed applications are used to address specific analytic needs as they arise.

2) The fragmented point solutions are neither co-located in a data warehouse nor otherwise architecturally integrated with one another.

3) Overlapping data content leads to multiple versions of analytic truth.

4) Reports are labour intensive and inconsistent.

5) Data governance is non-existent.

This level 0 has definitely been surpassed. Our systems currently address the problems of overlapping data content and the different components are connected together by either the E-health information system or the central healthcare system along with the X-Road system.

Enterprise Data Warehouse or level 1 is defined the following: [13]

1) At a minimum, the following data are co-located in a single data warehouse, locally or hosted: HIMSS EMR Stage 3 data, Revenue Cycle, Financial, Costing, Supply Chain, and Patient Experience.

2) Searchable metadata repository is available across the enterprise.

3) Data content includes insurance claims, if possible.

4) Data warehouse is updated within one month of source system changes.

5) Data governance is forming around the data quality of source systems.

6) The EDW reports organizationally to the CIO.

Level 1 has definitely been surpassed also. Although this level talks about the implementation of a classical Enterprise Data Warehouse, this has to be adjusted a bit if we want to talk about our systems. Our systems are based on the Data embassy concept that the Estonian E-Health systems use. The basic idea behind it is that we do not run a large very centralised database on the territory of Estonia but that our data systems are decentralised and can be split-up into different components that can run both locally and outside the country. In regard to data, this means that our databases are usually separated from each other but the data can easily be accessed through X-Road within each component when needed to compile documents, reports and perform analytics.

Standardized Vocabulary & Patient Registries or level 2:[14]

1) Master vocabulary and reference data identified and standardized across disparate source system content in the data warehouse.

2) Naming, definition, and data types are consistent with local standards.

3) Patient registries are defined solely on ICD billing data.

4) Data governance forms around the definition and evolution of patient registries and master data management.

Level 2 has also been surpassed or achieved. Currently TEHIK has a special department of data management specialists who are responsible for having proper structure for the databases and also to have standardized vocabulary in all of our products and databases. This is an ongoing process to keep things like LOINC codes and other standards up-to-date and updated in our systems. Data types are usually managed within the databases and different systems by the same teams or if needed, through admins. The governance exists and has been evolving regarding patient registries and patient data separately due to state requirements.

Automated Internal Reporting or level 3: [15]

1) Analytic motive is focused on consistent, efficient production of reports supporting basic management and operation of the healthcare organization.

2) Key performance indicators are easily accessible from the executive level to the front-line manager.

3) Corporate and business unit data analysts meet regularly to collaborate and steer the EDW.

4) Data governance expands to raise the data literacy of the organization and develop a data acquisition strategy for Levels 4 and above.

5) Efficient, consistent production and agility

Level 3 also exists to a large degree in Estonian systems and processes. Regular reports are produced based on performance indicators and they are easily accessible to specific teams and people if necessary. Our data management specialists/analysts definitely work closely together with other departments to steer the development of the EDW and our systems.

Automated External Reporting or level 4: [16]

1) Analytic motive is focused on consistent, efficient production of reports required for regulatory and accreditation requirements, payer incentives and specialty society databases.

2) Adherence to industry-standard vocabularies is required.

3) Clinical text data content is available for simple key word searches.

4) Centralized data governance exists for review and approval of externally released data.

Level 4 also exists to a large degree in Estonian systems and processes. The situation is made a bit more straightforward because the E-health services are managed by the state while the state itself is the regulator also. The other regulator that Estonian E-health services need to answer to is the European Union regulators and they also receive reports from our systems to be in sync and to follow the standards of the European Union. Industry-standard vocabularies are in place as already discussed and clinical text data content is available for search and access through either the doctor, patient portals or the new up-and-coming data observer system. Our data releases are reviewed and approved for release if the information is being published to institution or people outside our organisation or outside the internal ministry.

Clinical Effectiveness & Accountable Care or level 5:[17]

1) Analytic motive is focused on measuring adherence to clinical best practices, minimizing waste, and reducing variability.

2) Data governance expands to support care management teams that are focused on improving the health of patient populations.

3) Population-based analytics are used to suggest improvements to individual patient care.

4) Permanent multidisciplinary teams are in-place that continuously monitor opportunities to improve quality, and reduce risk and cost, across acute care processes, chronic diseases, patient safety scenarios, and internal workflows.

5) Precision of registries is improved by including data from lab, pharmacy, and clinical observations in the definition of the patient cohorts.

6) EDW content is organized into evidence-based, standardized data marts that combine clinical and cost data associated with patient registries.

7) Data content expands to include insurance claims (if not already included) and HIE data feeds.

8) On average, the EDW is updated within one week of source system changes.

This is the first step that in the author opinion has not been achieved. While certain characteristics of this level are present, the main idea behind this phase has not been achieved. While Estonian systems definitely allow for quick access of data to healthcare specialists in the country and our databases and registries include data from labs, pharmacies, clinical observations and so on, TEHIK does not use that data to improve the quality of the services in the classical sense. Currently the system that Estonian healthcare uses as EDW is structured along the idea of evidence-based, standardized data marts but TEHIK does not actually use this data to improve the quality of the processes in the healthcare sector. So while the existence of such a system definitely helps the doctors and other specialists with their work, it is not used to measure the best practices, minimize waste, or reduce variability. In addition, this step might be further problematic because the monetary aspect of our healthcare is not perfectly accounted in our systems.

Population Health Management & Suggestive Analytics or level 6:[18]

1) The "accountable care organization" shares in the financial risk and reward that is tied to clinical outcomes.

2) At least 50% of acute care cases are managed under bundled payments.

3) Analytics are available at the point of care to support the Triple Aim of maximizing the quality of individual patient care, population management, and the economics of care.

4) Data content expands to include bedside devices, home monitoring data, external pharmacy data, and detailed activity-based costing.

5) Data governance plays a major role in the accuracy of metrics supporting quality-based compensation plans for clinicians and executives.

6) On average, the EDW is updated within one day of source system changes.

7) The EDW reports organizationally to a C-level executive who is accountable for balancing cost of care and quality of care.

Level 6 cannot be achieved right now because the level 5 main ideas have not been implemented. If the cost-effective analysis is not done in the 5th level then the further inclusion of additional data streams from bedside devices, home monitoring data and so on is irrelevant. The organisation or system cannot share in the financial risk and reward if the cost-efficiency analysis has not been done.

3.2 What are the possible ways forward

After having discussed where Estonian it systems are right now, the author wants to know what the possibilities are according to this framework to go forward. Also the author wants to say that being at stage 4.5 or around stage 5 is not that bad because according to the author of the framework, the situation in 2016 was that the leading healthcare providers in the US were around stage 6 or 6.5 and only some elements of stages 7, 8 and 9 were implemented here and there but a full maturity of those stages was not achieved.[19]

The basic ideas on moving forward can be understood by understanding the framework a bit better. Here is a good picture where the framework is divided into blocks. This picture is used to display the tools that the creator of the framework can suggest in America but this can serve our purposes also:

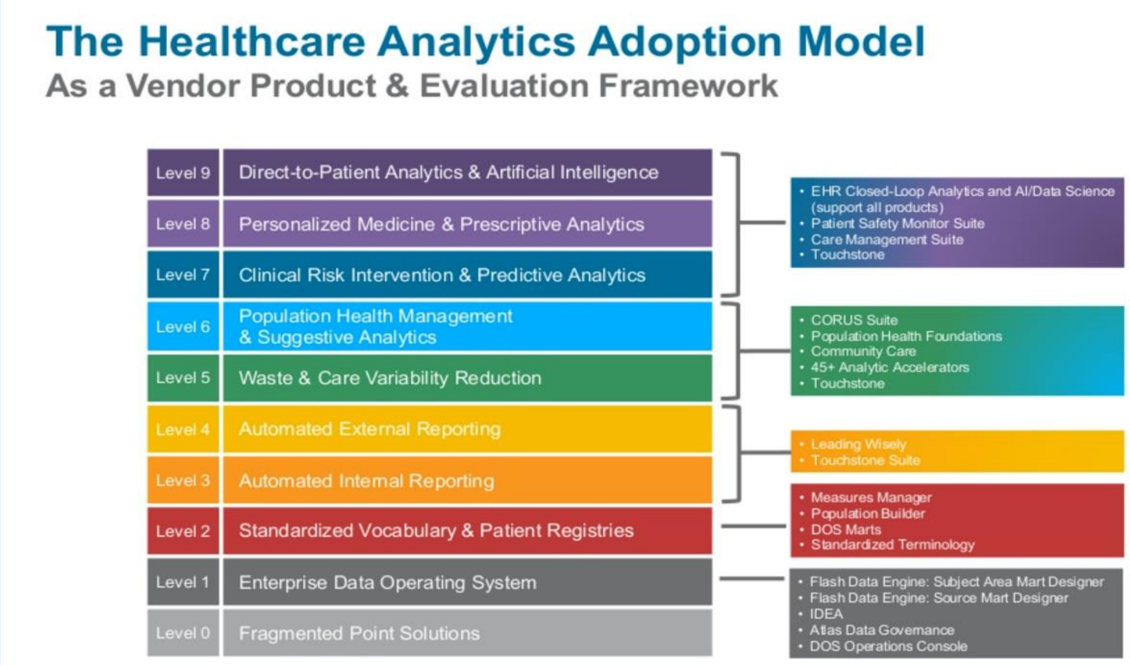

The Healthcare Analytics Adoption Model
As a Vendor Product & Evaluation Framework
Level 9
Direct-to-Patient Analytics & Artificial Intelligence
Level 8
Personalized Medicine & Prescriptive Analytics
Level 7
Clinical Risk Intervention & Predictive Analytics
Level 6
Population Health Management & Suggestive Analytics
Level 5
Waste & Care Variability Reduction
Level 4
Automated External Reporting
Level 3
Automated Internal Reporting
Level 2
Standardized Vocabulary & Patient Registries
Level 1
Enterprise Data Operating System
Level 0
Fragmented Point Solutions
• EHR Closed-Loop Analytics and AI/Data Science (support all products)
• Patient Safety Monitor Suite
• Care Management Suite
• Touchstone
• CORUS Suite
• Population Health Foundations
• Community Care
• 45+ Analytic Accelerators
• Touchstone
• Leading Wisely
• Touchstone Suite
• Measures Manager
• Population Builder
• DOS Marts
• Standardized Terminology
• Flash Data Engine: Subject Area Mart Designer
• Flash Data Engine: Source Mart Designer
• IDEA
• Atlas Data Governance
• DOS Operations Console

Figure 5 Healthcare Analytics Adoption Model and tools (Author: Dale Sanders PhD, Source: https://www.youtube.com/watch?v=EQ5c5xZfkUA&ab_channel=HealthCatalyst)

By looking at the picture, there are distinct 5 blocks that can be noticed. One block is for stage one or implementation of the enterprise data warehouse. 2nd block is for Standardized vocabulary and patient registries. 3rd block is for automated external and internal reporting for the purpose of management and accreditation. 4th block is for process/decision optimisation using statistical and cost data. 5th block involves personal care analytics improvement through predictive, prescriptive, and direct to patient analytics.

As the author sees it, one of the ways forward is to continue improving our systems and tools according to the framework itself. If our systems are around stage 4.5 then we can focus on achieving stage 5 and 6 for the purposes of process and decision optimisation. The main idea behind this is to analyse the historical data that we have regarding the healthcare choices that our institutions make and the cost-success of those decisions. For this to happen then we will need to pair our treatment data with our financial data regarding the price and the cost of those treatments along with the medication and materials used. The author of the framework, points to the minimal required data for this in stage 2:

At a minimum, the following data sources are co-located in a single local or hosted data warehouse: (1) HIMSS EMR Stage 3 clinical data, (2) financial data (particularly costing data), (3) materials and supplies data, and (4) patient experience data.[20]

While this way forward might look straightforward, it will also have its downsides. Since we are talking about global decisions on healthcare practices then this particular change requires not only statistical input from statistical analysis and raw data but also input from the side of healthcare specialists. Also this is not something that TEHIK alone can decide because the scope of decisions is outside just an e-health information system perspective. While it is possible for us to create tools to make those decisions easier, it will ultimately not be possible for us to force healthcare provides like hospitals or clinics to use specific treatments for x or y because they are more efficient statistically or outcome wise. These stages can be implemented by TEHIK by providing the tools for the healthcare providers to do those kinds of decisions or by the internal ministry giving TEHIK more competence/delegating competence in this regard.

Another way forward that the author sees from this framework is the possibility to go straight for 5th block of stages 7,8 and 9. While this might seem a bit counterintuitive since the framework has other steps before it for a good reason, implementation of stages 5 and 6 might be outside the competence of TEHIK as mentioned before. Also the majority of our systems are more focused on clinical data and outcomes instead of costs which would make the implementation of these steps for personal care a bit more straightforward in regard to the data. The main ideas behind this block is to introduce algorithms and advancements from machine learning and neural networks first for the purposes of diagnosing patients, forecasts for patient treatment and risk analysis and then go in the direction of decision making for healthcare specialists like prescriptions regarding medicine and treatment of disease ultimately leading to a system the end-goal of which is a direct-to-patient healthcare without any involvement of healthcare specialists and the system deciding almost everything from diagnosis to treatment and medicine. This might seem a bit futuristic but the way forward is clear and there are publicly available technologies that can be implemented for at least the diagnosis stage and with some help from healthcare specialists prescriptive stage might also be added based on the diagnosis. These kinds of solutions are also a bit more in the scope of TEHIK to create as an assistant or assistance for healthcare specialists.

The third way forward that the author sees comes from the framework and the experience of the frameworks creator that he has stated in his presentations. The creator of the framework mentioned that blocks 1, 2 and 3 are the pillars for the blocks 4 and 5. And while he has implementing block 4 in real life, it was somewhat hard and relatively unsuccessful because automatic reporting from block 3 was consuming a lot of resources and time for itself.[21]

In this sense we can put ourselves the goal to implement block 4 but instead of starting with block 4 we can do a full review of blocks and stages that come before. The main ideas behind this would be to do an overview of our reporting and data warehouse solutions that we already have or are currently under development. These solutions would need to be modified to include not only the clinical data but also the cost and price information attached to that clinical data and outcomes.

This way forward should be simpler compared to the other 2 because it would not require implementing new technologies or creating specific teams for statistical analysis comprising of data scientists and healthcare specialists. Nor does it need increasing competence of TEHIK to do these kinds of decisions. The largest downside to this change is the cost and the time required. Doing a full overview of our existing databases and doing a full refactor to fit the future needs is no small task considering how many systems are already connected to them.

3.3 Decision between the 3 ways forward

In the authors opinion, the decision between the 3 ways forward is straightforward. Since we are talking about TEHIK then there are certain legal boundaries that we need to operate in. These boundaries are set by the main statute for the Health and Welfare Information Systems Centre. The current statute states the following about the field of activity and tasks for the centre in the 2nd chapter:[22]

> 17. In its field of activity, the Centre shall perform the following tasks:
>
> 1) Organize and co-ordinate the development, management and implementation of strategies, development plans and budgets related to information and communication technology administrative and development processes.

2) Co-ordinate the digital development of the services in the area of the ministry, including the strategic planning of information systems and e-services, taking into account the objectives of the area of governance of the ministry and the possibilities of information technology.

3) Participates in the development of strategies and development and action plans of the area of governance of the ministry in view of the possibilities of e-services and the objectives of the information society.

11) Organize the development and implementation of the best methods of personal data protection in the information systems and databases of the area of governance of the ministry.

12) Implements the information security policy of the area of governance of the ministry within the limits of its field of activity.

15) Ensure central services for the provision of information and communication technology in the area of governance of the ministry: infrastructure, including data communication, data security, backup, system administration; software support for the effective implementation of information and communication technology management and development; systems integration, maintenance, and computer support.

17) Organize the development and management of data transmission formats, data control rules and classifications related to information systems.

21) Co-ordinates and co-operates with data providers and supervisory authorities in the field of data quality management and data acquisition.

26) Participate in the development of legislation concerning its field of activity, express an opinion on the draft submitted to the Ministry for approval and make proposals for the amendment of legislation.

27) Co-operates with the area of governance of the Ministry and other state agencies and international institutions for the development and implementation of e-solutions in the field of health, work, and social affairs, including participation in the work of relevant committees, councils and working groups.

29) Make proposals to the management of the ministry for the organization of the functions of the authorized processes of information systems and registers in the field of health, labour and social affairs and perform the functions of the authorized processes within the limits of its competence.

30) Perform other functions specified by legislation or assigned by the ministry.

The author has selected the clauses from that paragraph that are relevant for the purposes of the current work. As can be seen from those clauses, TEHIK has broad competence in a lot of relevant areas: data handling, data warehousing, data security, development of systems and even strategic decision making to some degree. Although TEHIK is involved in strategic decision making to some degree, it has no real competence in the questions of broad healthcare optimisation or/and cost optimisation for the healthcare sector. Theoretically this competence can be given to TEHIK either through special government panels or by other legal acts however it does not have it right now. Which means that currently the only possible choice that would follow the legal limits of TEHIK is the 2nd choice of going for 5th block and stages 7,8,9 of the model. Technically option 3 also falls in within the competence of TEHIK but this choice would be severely constrained by time, budget limitations and work required to finish it. A whole refactoring of our data warehousing solution will inevitably lead to a need to refactor the hundreds of systems already connected to it. So the most optimal choice right now is option 2.

3.4 How can we proceed onwards with those analytics

Since we know where we want to get now then the author needs to chart a path for that direction. In regard to proceeding onwards there is an existing widely accepted way of doing these kinds of analytics. It is called Knowledge Discovery in Databases or KDD for short. This process can be described with the following picture:

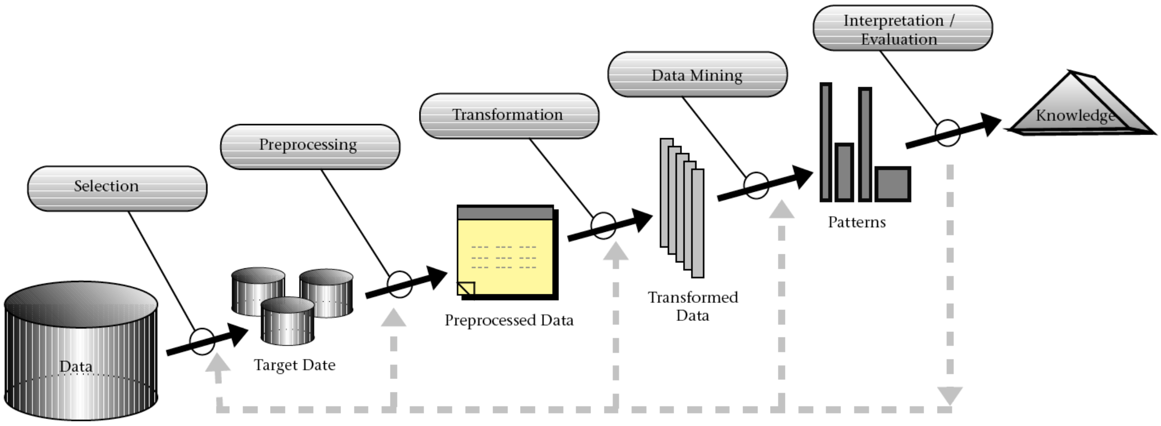


Figure 6 From data mining to knowledge discovery in databases. (Author: [Fayyad et al., 1996] U. Fayyad, G. P.-Shapiro, and P. Smyth. AI Magazine, 17(3):37-54, Fall 1996. Source: http://citeseer.ist.psu.edu/fayyad96from.html)

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:[23]

1) Developing an understanding of
    a. The application domain
    b. The relevant prior knowledge
    c. The goals of the end-user
2) Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.
3) Data cleaning and pre-processing.
    a. Removal of noise or outliers.
    b. Collecting necessary information to model or account for noise.
    c. Strategies for handling missing data fields.
    d. Accounting for time sequence information and known changes.

4) Data reduction and projection.

    a. Finding useful features to represent the data depending on the goal of the task.

    b. Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

5) Choosing the data mining task.

    a. Deciding whether the goal of the KDD process is classification, regression, clustering, etc.

6) Choosing the data mining algorithm(s).

    a. Selecting method(s) to be used for searching for patterns in the data.

    b. Deciding which models and parameters may be appropriate.

    c. Matching a particular data mining method with the overall criteria of the KDD process.

7) Data mining.

    a. Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

8) Interpreting mined patterns.

9) Consolidating discovered knowledge.

3.5 Formulating questions arising from KDD

The author believes that the prior knowledge has been established for part one. Also the author has spent quite a lot of time explaining the goals of the end user and narrowing them down. The current goal is to introduce algorithms and advancements from machine learning and neural networks first for the purposes of diagnosing patients, forecasts for patient treatment and risk analysis and then go in the direction of decision making for healthcare specialists like prescriptions regarding medicine and treatment of disease ultimately leading to a system the end-goal for which is a direct-to-patient healthcare without any involvement of healthcare specialists and the system deciding almost everything from diagnosis to treatment and medicine. The long-term objective should be narrowed to predictive analytics or diagnosing in this case since the other types or analytics will have to start from diagnosis.

The next question would be what disease or illness the work is going to be predicting and focusing on. The author believes that this is a very complicated question which requires professional medical expertise to answer or at the least, some kind of cost-effective analysis of the whole Estonian healthcare system to determine where assistance is required and where costs can be reduced which is way outside the scope of this work. The author believes that since the whole idea for the work is too show how this change can be implemented and the data processing can be introduced then it is of a lower importance which illness exactly will be predicted or diagnosed. Perhaps a separate department or group needs to be set up within the data management branch of TEHIK to answer these kinds of questions. For clarity of the work, the author wants to try to predict or diagnose lung cancer in patients. The author wants to do this choice because lung cancer related illnesses is one of the two leading causes of death in the world.

Then the next questions that we need to answer is what data we have in this regard and how can we get permissions to use them if we do not already have those permissions. To answer this question the author will need to consult medical professionals and TEHIK data management department.

After the questions regarding the data is clarified then the author needs to provide guidance on the algorithm or a machine learning method. The author will use pre-existing research on the subject to decide on the alternatives and their effectiveness.

When the questions about the data and the algorithms are answered then the author can tackle the questions of data pre-processing, reduction and interpretation of results depending on the choice of the algorithm.

And once all of the algorithm and data questions are answered then we need to position the algorithm and the movement of data somewhere in the system. In here, we will answer the questions of architecture and communication between different components within the Estonian national e-health information ecosystem.

4 Business analysis

This chapter of the master’s thesis is designed to answer process related questions along with other technical questions that might be related to the algorithm or data. The first part will focus on the AS-IS process along with the data that is needed for the algorithm and the algorithm itself. The second part will focus on data pre-processing, access, and the TO-BE processes.

4.1 AS-IS process for data processing and lung cancer analysis

The general way the data works right now is the following:

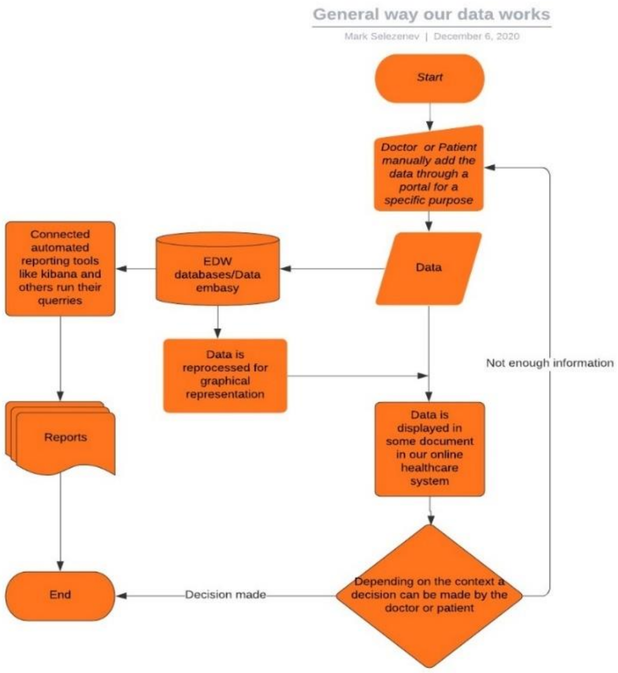

General way our data works
Mark Selezenev | December 6, 2020
Start
Doctor or Patient manually add the data through a portal for a specific purpose
Data
EDW databases/Data embasy
Connected automated reporting tools like kibana and others run their querries
Data is reprocessed for graphical representation
Not enough information
Data is displayed in some document in our online healthcare system
Reports
End
Decision made
Depending on the context a decision can be made by the doctor or patient

Figure 7 General way our data works (Source: Author created.)

The flow starts with the patient or doctor wanting to manually enter any data in the system for some purpose. After the data is entered and saved then it moves to the databases that work on data ambassy concept and can be comparable to a EDW. After that 2 situations can happen:

1) This data will be used by patient or doctors for decision making regarding treatment, medicine and so on. To do this decision the data will usually be re-processed to be represented in some document in a graphical manor.
2) The data will go into a report through an automated tool. This might be an internal or external report.

Adding of lung cancer diagnosis works in a similar way and the methods that are used by Estonian doctors for diagnosing are the same as worldwide. While there are tests that are used to exclude some diagnosis like blood tests they are still not the main focus of this chapter because we are more interested in diagnosing lung cancer instead of ruling out other non-lung cancer related diseases. The main methods and information used for lung cancer diagnosing are:

1) Chest X-ray.[24]

A chest X-ray is usually the 1st test used to diagnose lung cancer. Most lung tumours appear on X-rays as a white-grey mass. However, chest X-rays cannot give a definitive diagnosis because they often cannot distinguish between cancer and other conditions, such as a lung abscess (a collection of pus that forms in the lungs). If a chest X-ray suggests you may have lung cancer, you should be referred to a specialist in chest conditions. A specialist can arrange more tests to investigate whether you have lung cancer and, if you do, what type it is and how much it is spread.

2) CT scan.[25]

A CT scan is usually the next test you will have after a chest X-ray. A CT scan uses X-rays and a computer to create detailed images of the inside of your body. Before having a CT scan, you will be given an injection containing a special dye called a contrast medium, which helps to improve the quality of the images. The scan is painless and takes 10 to 30 minutes.

3) PET-CT scan.[26]

A PET-CT scan may be done if the results of a CT scan show you have cancer at an early stage.

The PET-CT scan (which stands for positron emission tomography-computerised tomography) can show where there are active cancer cells. This can help with diagnosis and choosing the best treatment. Before having a PET-CT scan, you will be injected with a slightly radioactive material. You will be asked to lie down on a table, which slides into the PET scanner. The scan is painless and takes 30 to 60 minutes.

4) Different versions of bronchoscopy and biopsy.[27]

If a CT scan shows there might be cancer in the central part of your chest, you may be offered a bronchoscopy. A bronchoscopy is a procedure that allows a doctor to see the inside of your airways and remove a small sample of cells (biopsy). During a bronchoscopy, a thin tube with a camera at the end, called a bronchoscope, is passed through your mouth or nose, down your throat and into your airways. The procedure may be uncomfortable, so you will be offered a sedative before it starts, to help you relax, and a local anaesthetic to make your throat numb. The procedure takes around 30 to 40 minutes. A newer procedure is called an endobronchial ultrasound scan (EBUS), which combines a bronchoscopy with an ultrasound scan. Like a bronchoscopy, an EBUS allows a doctor to see the inside of your airways. However, the ultrasound probe on the end of the camera also allows the doctor to locate the lymph nodes in the centre of the chest so they can take a biopsy from them. The procedure takes around 90 minutes.

The results are kept in the following formats:

1) CT scan – Picture

2) X-ray – Picture

3) PET-CT – Picture

4) Bronchoscopy and biopsy – mostly as information which is part of databases in words and numbers but in some cases like EBUS then pictures/videos can also be included.

Also while the data and information are in the formats above there is still a final outcome for the diagnosis which is usually in a pre-determined format of either being positive or negative. In case of a positive lung cancer diagnosis there are also descriptions and how far the cancer has progressed.

Once the tests are done and the pictures are taken then they are entered into the system and saved. Based on those pictures or procedures, the doctor compiles a medical document where a final diagnosis is included along with the explanation of how far the lung cancer has progressed if the outcome was positive. When this information is fully entered in the system then it can be used for reporting purposes or to display them on any of the portals.

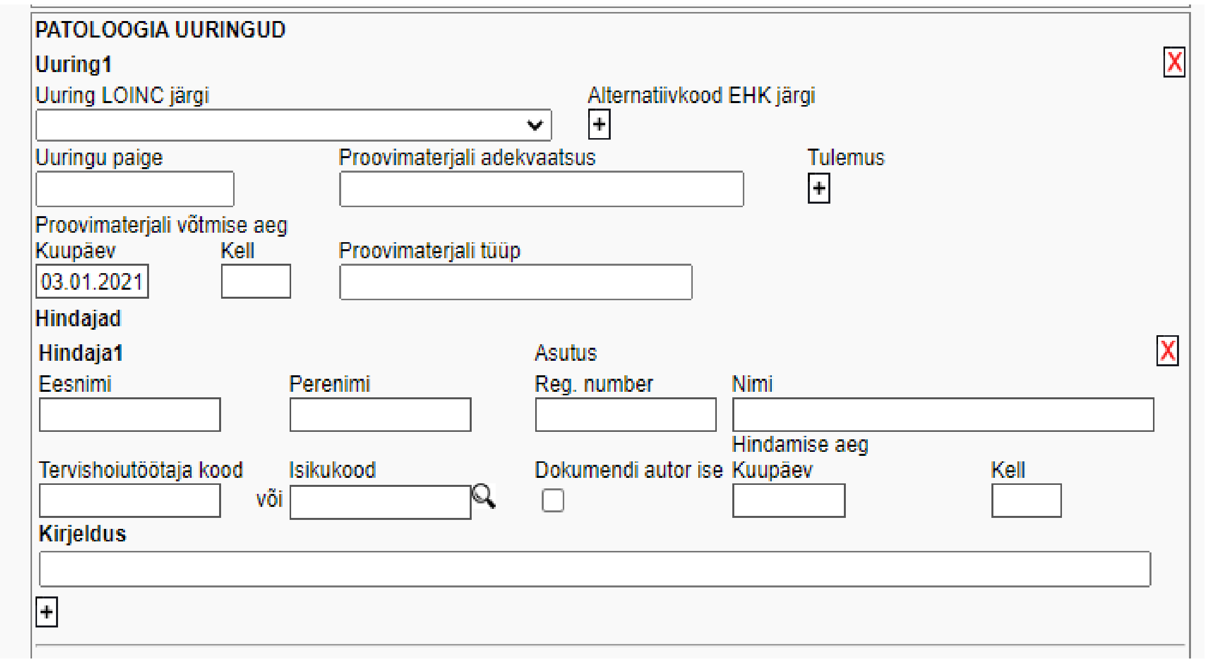
4.2 Data storage

As mentioned in the previous paragraph, the data that the author is interested in are images for CT, X-ray, PET-CT, EBUS and regular data in word format regarding biopsies and bronchoscopy.

After talking to Kerli Linna ,who is the Head of Data Management at TEHIK[28], the following information was disclosed. Currently the imaging data and regular data in word format are stored in two different places. The regular data is stored in central healthcare systems databases while pictures are managed by a separate entity which is called “Sihtasutus Eesti Tervishoiu Pildipank” or just “Pildipank” which is translated as bank for pictures from Estonian. Right now this organisation is under the Tartu University Clinic/Hospital. Currently the pictures are stored in DICOM format[29].

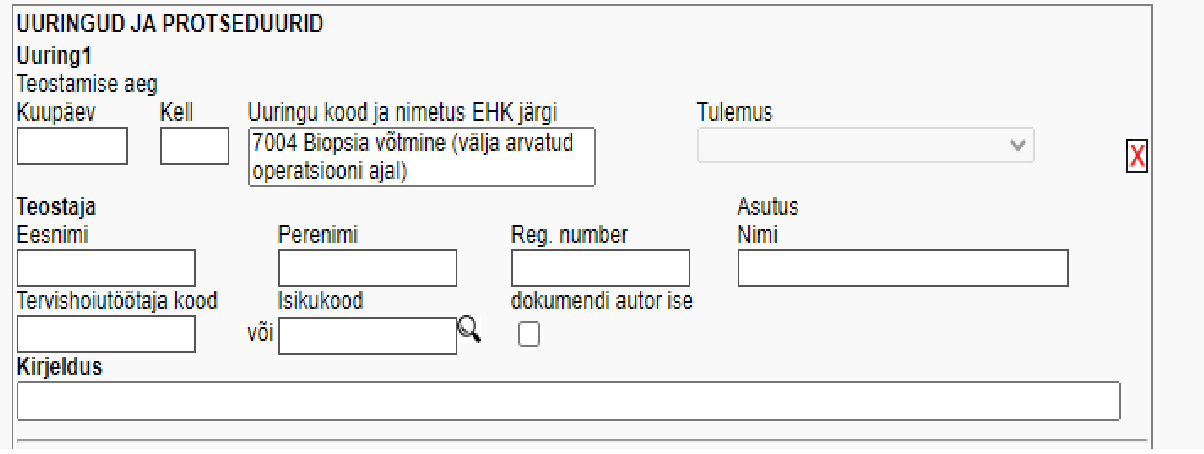
After the author contacted the Pildipank regarding the standards for the images, he was able to clarify that there are no Estonia specific DICOM standards and the system accepts any images that are allowed for the DICOM standard. This means that most of categories will contain images in different formats which will heavily depend on the equipment used to take those images. The general modality or dimensions can range from 64x64 to 4000x5000 and beyond depending on the equipment.

Regarding the biopsy the situation is a bit more simpler. While the general information about lung cancer can be and is coded under different parts of Estonian medical documents which includes pictures and text but the authors main aim is biopsy right now. It is coded under the “Patoloogia” and “Protseduurid ja Uuringud” which have the following data in them:



Kuupäev 03.01.2021

Figure 8 Pathologies (Source: Author created from Arstiportaal.ee)



Uuringu kood ja nimetus EHK järgi 7004 Biopsia võtmine (välja arvatud operatsiooni ajal)

Figure 9 Probes and Procedures (Source: Author created from Arstiportaal.ee)

The full latest XML 8.1 standards can be located at the standards page for Estonian healthcare system: http://pub.etervis.ee/standards2/Standards/8.1. It can be located in any xml if the document contains those blocks, for example in the file called “Vastus saatekirjale.xml”.

4.3 Methodologies used to predict lung cancer and their viability

The most common ways that are used today are based on CT images. The general process looks like this[30]:

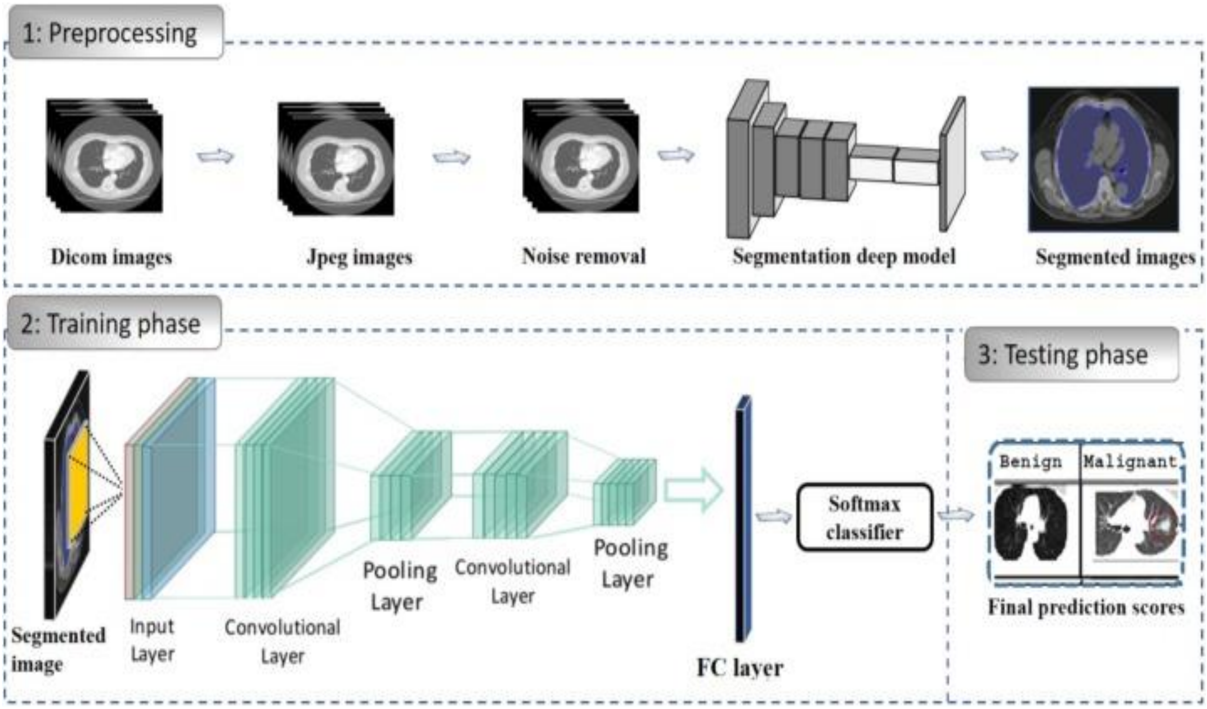


Figure 10 Machine supported framework for lung nodules prediction.(Source: https://www.sciencedirect.com/science/article/pii/S1876034120305633#bib0435)

The CT scan DICOM image is taken and then it is converted to a more understandable format for the machine-learning tools like .Jpeg. After the conversion is done then unnecessary elements are removed from the image. Once the elements are removed then the image can be used for training or prediction depending on the situation.

The alternative methods that are used usually involve machine learning algorithms like:

1) Decision Trees.

2) K-Nearest Neighbours.

3) Logistic regression.

4) Random Forest.

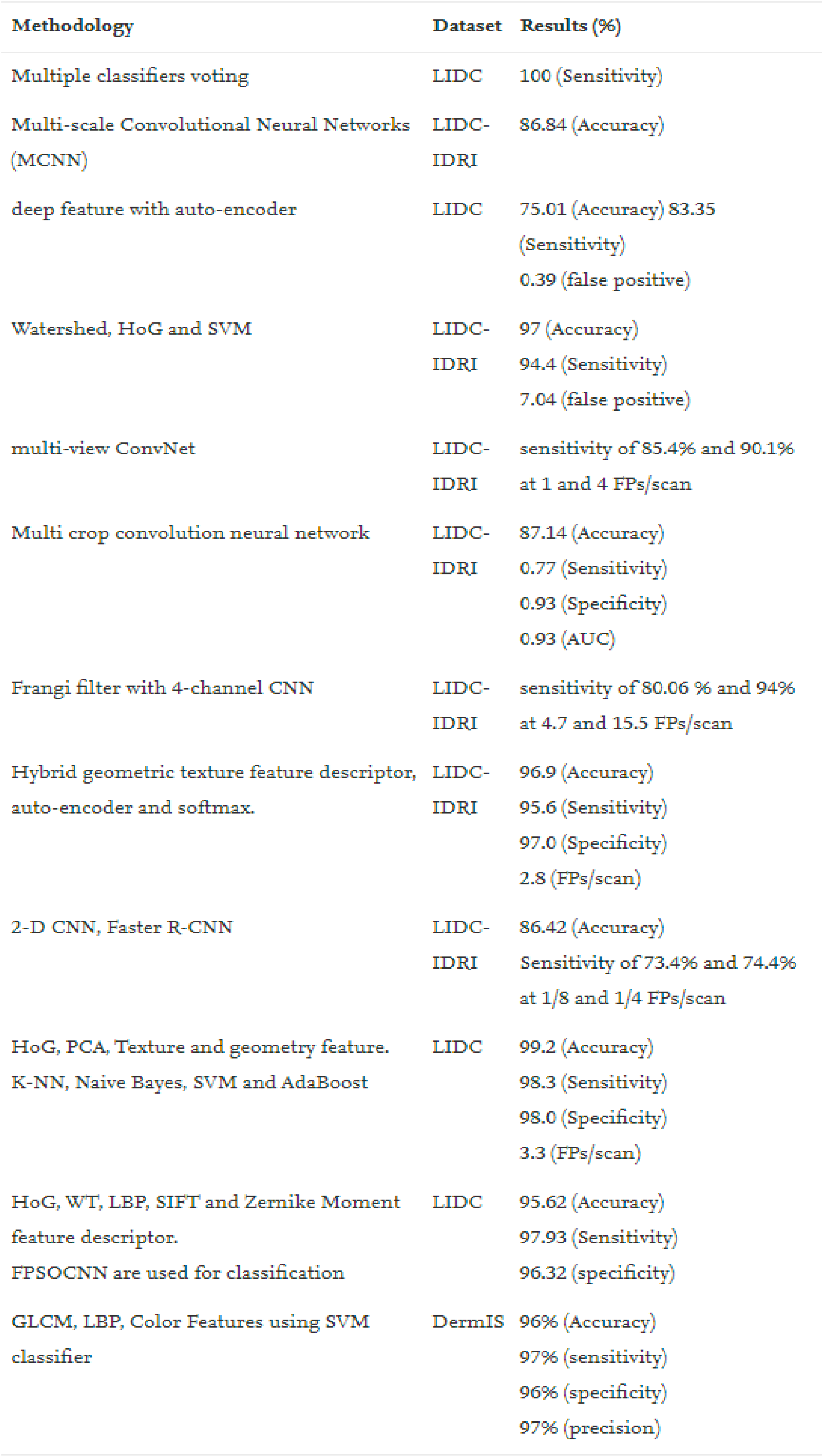
5) Support vector machines.

These methods involve gathering data and then formatting it for usage with those algorithms. Usually means transforming them into numerical formats. After that the data can be split between two datasets so the algorithms will be able to train on that data and then analyse the accuracy of those algorithms on the 2nd dataset. An example of such work can be found here.[31]

The difference between the two approaches is important for the purposes on this work. CT images are a standardized format that will contain the information which the Estonian Central E-health system already has. On the other hand, the machine learning algorithm approach is a lot less standardized in regard to the data and is usually done on public datasets that are available. Those datasets usually contain data that might not be stored in the Estonian healthcare databases at all or in a format that can be easily accessed. While it is possible to adapt the machine learning algorithms to use more or less data if needed, it is a field that will require further research and investment in general. Or alternatively, it might require redoing how and what data the databases store.

In the authors opinion, the best choice that we can do right now is use the pre-built image processing solutions and methods that are available and require adaptation to our systems with minimal research instead of using solutions like machine learning algorithms that might require significant amount of research into the data used and algorithms applied or to the structure of systems and databases.

Another question that the author needs to answer is how accurate different approaches are for CT image-based processing. There has been research on the subject and the following analysis has been published very recently:



Methodology	Dataset	Results (%)
Multiple classifiers voting	LIDC	100 (Sensitivity)
Multi-scale Convolutional Neural Networks (MCNN)	LIDC-IDRI	86.84 (Accuracy)
deep feature with auto-encoder	LIDC	75.01 (Accuracy) 83.35 (Sensitivity) 0.39 (false positive)
Watershed, HoG and SVM	LIDC-IDRI	97 (Accuracy) 94.4 (Sensitivity) 7.04 (false positive)
multi-view ConvNet	LIDC-IDRI	sensitivity of 85.4% and 90.1% at 1 and 4 FPs/scan
Multi crop convolution neural network	LIDC-IDRI	87.14 (Accuracy) 0.77 (Sensitivity) 0.93 (Specificity) 0.93 (AUC)
Frangi filter with 4-channel CNN	LIDC-IDRI	sensitivity of 80.06 % and 94% at 4.7 and 15.5 FPs/scan
Hybrid geometric texture feature descriptor, auto-encoder and softmax.	LIDC-IDRI	96.9 (Accuracy) 95.6 (Sensitivity) 97.0 (Specificity) 2.8 (FPs/scan)
2-D CNN, Faster R-CNN	LIDC-IDRI	86.42 (Accuracy) Sensitivity of 73.4% and 74.4% at 1/8 and 1/4 FPs/scan
HoG, PCA, Texture and geometry feature. K-NN, Naive Bayes, SVM and AdaBoost	LIDC	99.2 (Accuracy) 98.3 (Sensitivity) 98.0 (Specificity) 3.3 (FPs/scan)
HoG, WT, LBP, SIFT and Zernike Moment feature descriptor. FPSOCNN are used for classification	LIDC	95.62 (Accuracy) 97.93 (Sensitivity) 96.32 (specificity)
GLCM, LBP, Color Features using SVM classifier	DermIS	96% (Accuracy) 97% (sensitivity) 96% (specificity) 97% (precision)

Figure 11 Imaging methodologies for lung cancer (Source: https://www.sciencedirect.com/science/article/pii/S1876034120305633#bib0435)

So while the accuracy is not 100% and there are obviously false positives, still the author considers the different approaches acceptable for the intents and purposes of the new system.

4.4 Data pre-processing, reduction, and result interpretation

Since the situation is more or less understood then to be able to use any algorithms pre-processing of images needs to be completed. This would usually be done using the PyDicom library for Python which has been designed for different kind of manipulations of the DICOM format. For the intents and purposes of this work the most important functions of this library are:

1) Working with pixel data.[32]

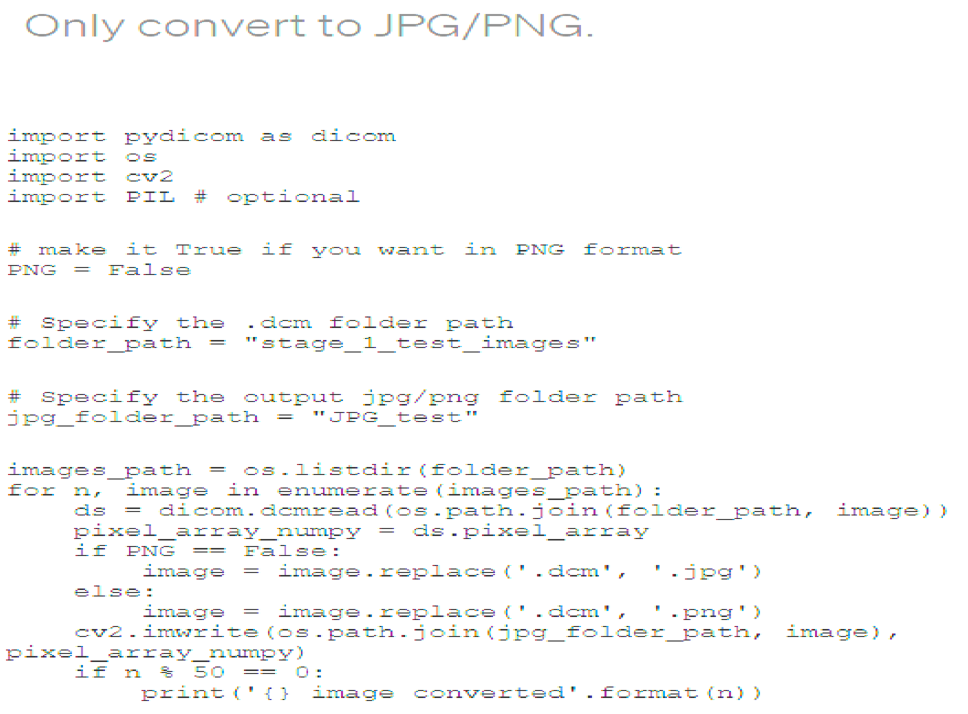
2) Writing and reading information from the DICOM file.[33]

What this means is that the library allows to get access to the information that is needed for the algorithms to work:

1) Images through Pixel manipulation.

2) Patient information that is also stored in the DICOM file.

When the data is accessed and extracted then it can be saved almost in any format that is needed. The most usual formats to save that data is JPEG or PNG for images and CSV files for regular information. Here is an example on how it is done for conversion:

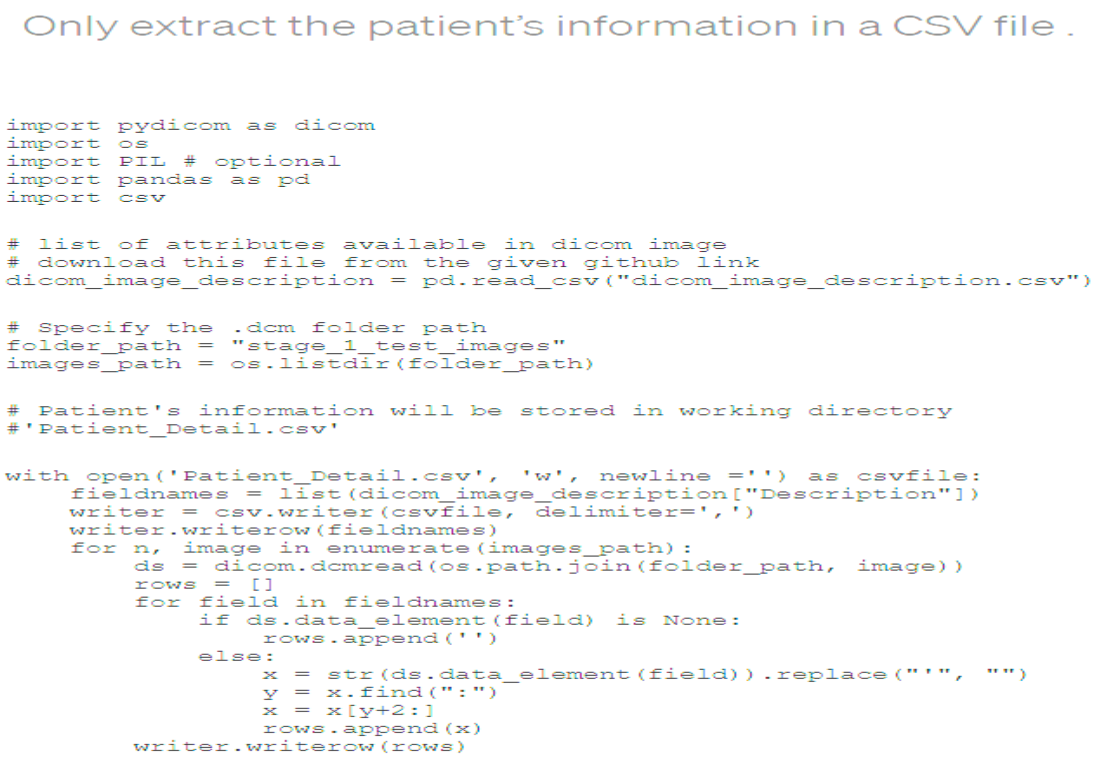


Only convert to JPG/PNG.

```
import pydicom as dicom
import os
import cv2
import PIL # optional

# make it True if you want in PNG format
PNG = False

# Specify the .dcm folder path
folder_path = "stage_1_test_images"

# Specify the output jpg/png folder path
jpg_folder_path = "JPG_test"

images_path = os.listdir(folder_path)
for n, image in enumerate(images_path):
    ds = dicom.dcmread(os.path.join(folder_path, image))
    pixel_array_numpy = ds.pixel_array
    if PNG == False:
        image = image.replace('.dcm', '.jpg')
    else:
        image = image.replace('.dcm', '.png')
    cv2.imwrite(os.path.join(jpg_folder_path, image),
pixel_array_numpy)
    if n % 50 == 0:
        print('{} image converted'.format(n))
```

Figure 12 Conversion of DICOM to PNG or Jpeg (Source: https://medium.com/@vivek8981/dicom-to-jpg-and-extract-all-patients-information-using-python-5e6dd1f1a07d)

And example for storing data in csv format:



Only extract the patient's information in a CSV file .

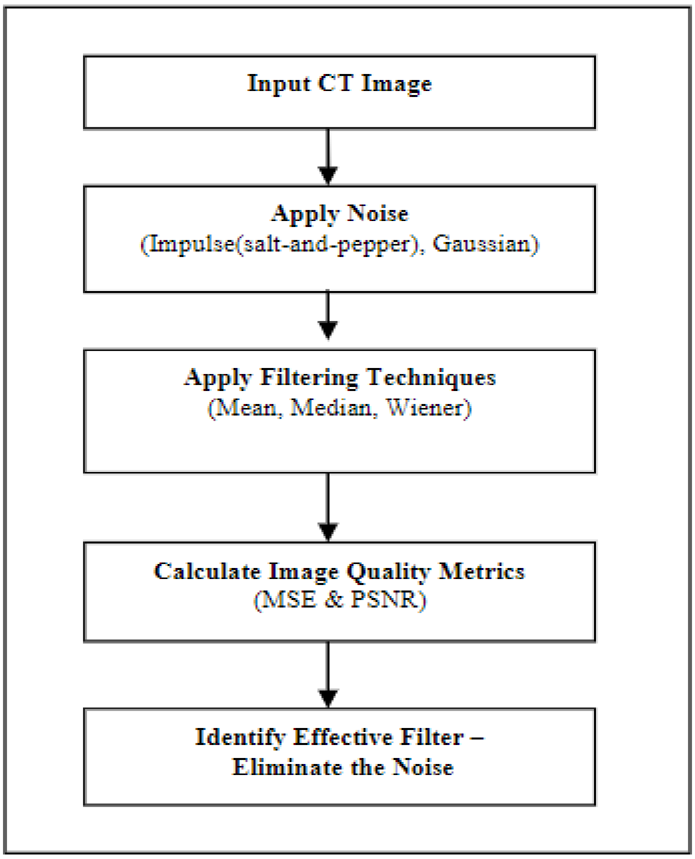
```
import pydicom as dicom
import os
import PIL # optional
import pandas as pd
import csv

# list of attributes available in dicom image
# download this file from the given github link
dicom_image_description = pd.read_csv("dicom_image_description.csv")

# Specify the .dcm folder path
folder_path = "stage_1_test_images"
images_path = os.listdir(folder_path)

# Patient's information will be stored in working directory
#'Patient_Detail.csv'

with open('Patient_Detail.csv', 'w', newline ='') as csvfile:
    fieldnames = list(dicom_image_description["Description"])
    writer = csv.writer(csvfile, delimiter=',')
    writer.writerow(fieldnames)
    for n, image in enumerate(images_path):
        ds = dicom.dcmread(os.path.join(folder_path, image))
        rows = []
        for field in fieldnames:
            if ds.data_element(field) is None:
                rows.append('')
            else:
                x = str(ds.data_element(field)).replace("'", "")
                y = x.find(":")
                x = x[y+2:]
                rows.append(x)
        writer.writerow(rows)
```

Figure 13 Extracting of patient data to CSV (Source: https://medium.com/@vivek8981/dicom-to-jpg-and-extract-all-patients-information-using-python-5e6dd1f1a07d)

The next question then is what kind of denoising techniques can be used. This is ultimately a very hard question in the authors opinion because it requires experimentation and goes a bit outside the scope of this work.

For the intents and purposes of this work, the author was able to locate a research on the subject which was conducted this year. The name of the research is “Elimination of Noise in CT Images of Lung Cancer using Image pre-processing Filtering Techniques”. [34]It was published in the International Journal of Advanced Science and Technology Volume 29, No. 4s, pp. 1823-1832.

The research was conducted in the following manner:



Input CT Image
Apply Noise
(Impulse(salt-and-pepper), Gaussian)
Apply Filtering Techniques
(Mean, Median, Wiener)
Calculate Image Quality Metrics
(MSE & PSNR)
Identify Effective Filter –
Eliminate the Noise

Figure 14 Research process(Source: http://sersc.org/journals/index.php/IJAST/article/view/6991/4163)

In the first step, the images were collected and then converted into grey level images of 8 bit so it was possible to apply the pre-processing techniques. The next step was used to identify and measure the noise present in the image. After the noises were identified then different filtering techniques were used to restore the images in spite of noises. After the filters were applied then image quality metrics were checked using Mean-Squared Error and Peak-To-Signal Noise Ratio and based on that a conclusion was produced regarding different methods.[35]

The results of this research are as follows:

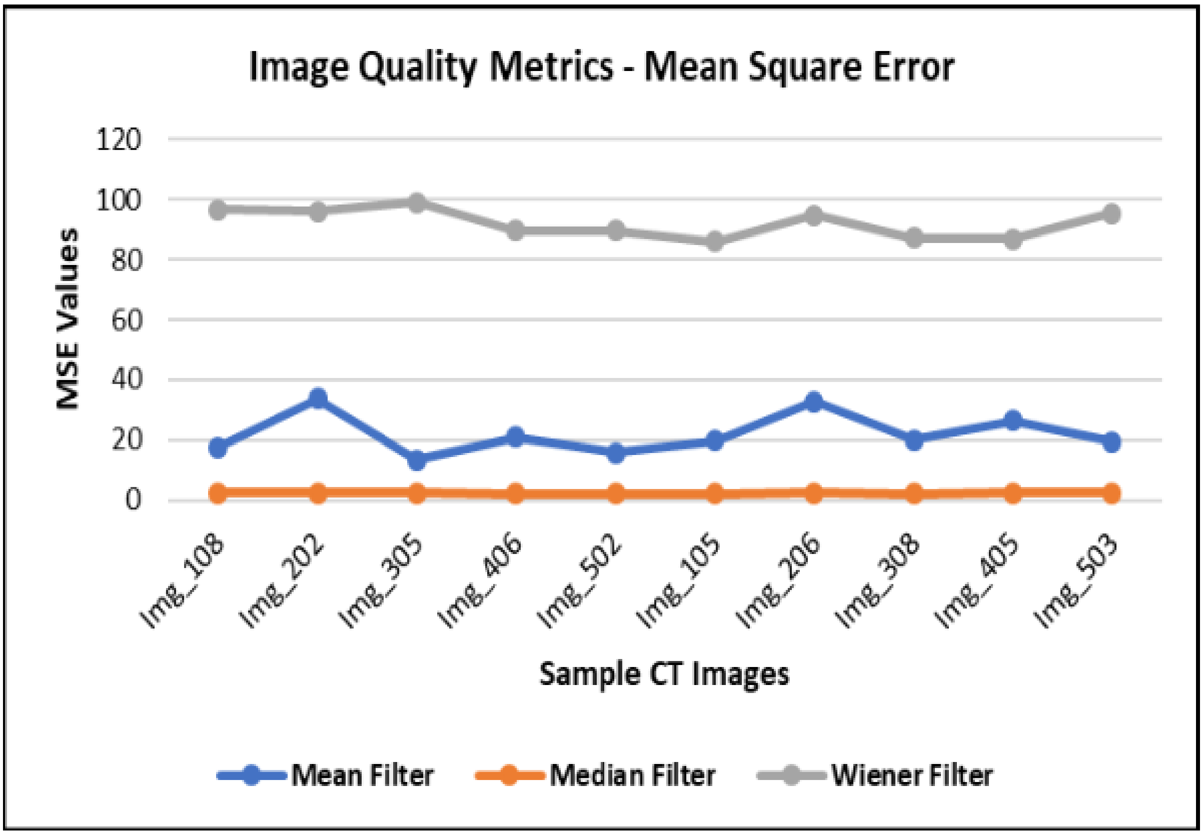


Figure 15 Image Quality Metrics - Mean Square Error(Source: Page 1830, http://sersc.org/journals/index.php/IJAST/article/view/6991/4163)

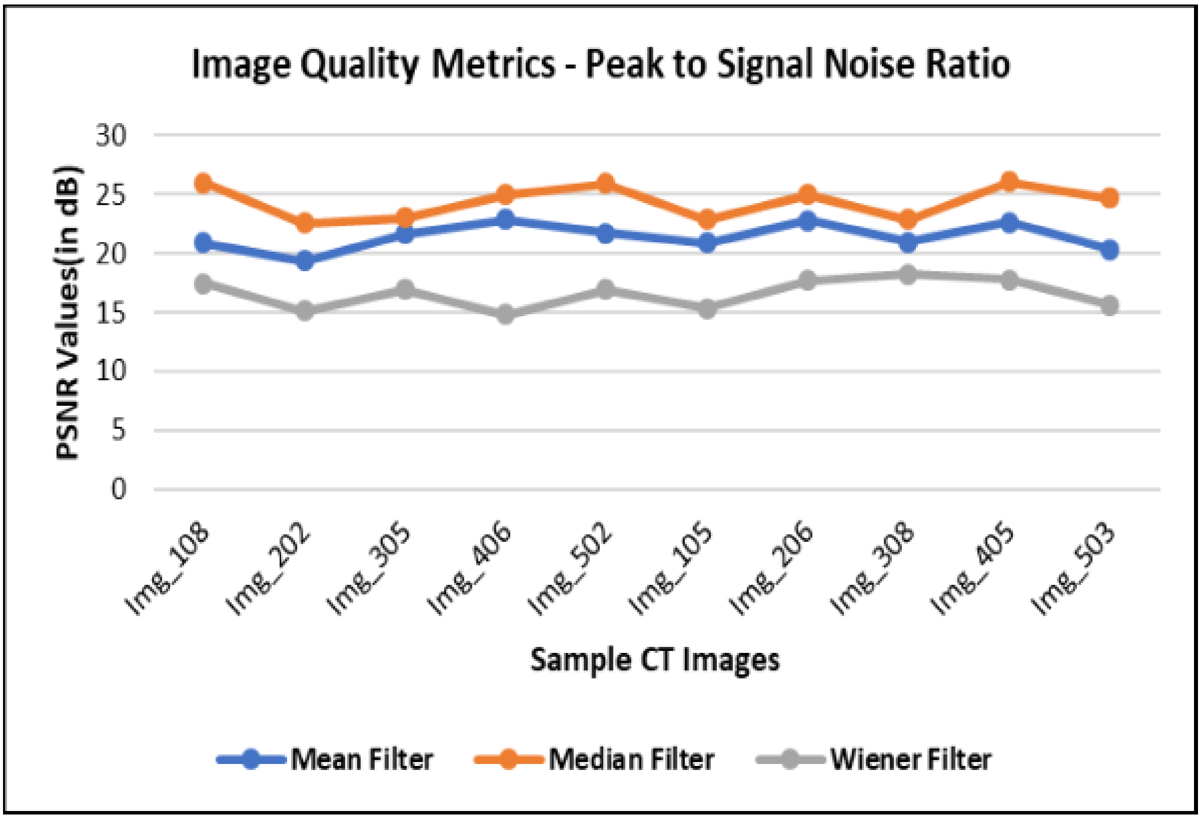


Figure 16 Image Quality Metrics - Peak to Signal Noise Ratio(Source: Page 1830, http://sersc.org/journals/index.php/IJAST/article/view/6991/4163)

This paper implemented various filters such as Mean, Median and Wiener on CT images for lung cancer. The performance of these filter was analysed based on the image quality metrics such as Peak-to-Signal Noise Ratio (PSNR) and Mean-Squared Error (MSE). Based on the results obtained for image quality metrics, it is noted that the Median filter is effective in compare to other filters in eliminating the noise by having high PSNR and low MSE values. This work was implemented using MATLAB 14a.[36]

The interpretation for the results through image recognition usually produces two results:

1) Benign or tumour not found - A growth that is not cancer. It does not invade nearby tissue or spread to other parts of the body.[37]

2) Malignant tumour - A tumour that invades surrounding tissues, is usually capable of producing metastases, may recur after attempted removal, and is likely to cause death unless adequately treated.[38]

4.5 TO-BE process

The new process will need to be divided in to two different categories:

1) The category that is responsible for the usage of the trained algorithm to produce results or suggestions for the users.

2) The category that is responsible for the training of the algorithm and replacing the previous version in the system.

The training process begins with gathering of the data in raw format that TEHIK might have in the systems. This step is not automated because the algorithms might use different sources or different data for the purposes of training and this is something that the data scientist needs to decide when it will be needed to optimise the algorithm. After the data is gathered then the data needs to be formatted into the appropriate format and any inconsistencies need to be removed to make the results more accurate. These 2 steps need to be automated because later on this code can be re-used for the purposes of producing results or suggestions for the users. After that the data scientist will need to decide if the amount of data is sufficient or is it needed to increase the amount of data using artificial means of data augmentation. This step can be manual. After the data scientist decides that the amount is sufficient then the algorithm or the neural network can be trained and after replaced in the system.

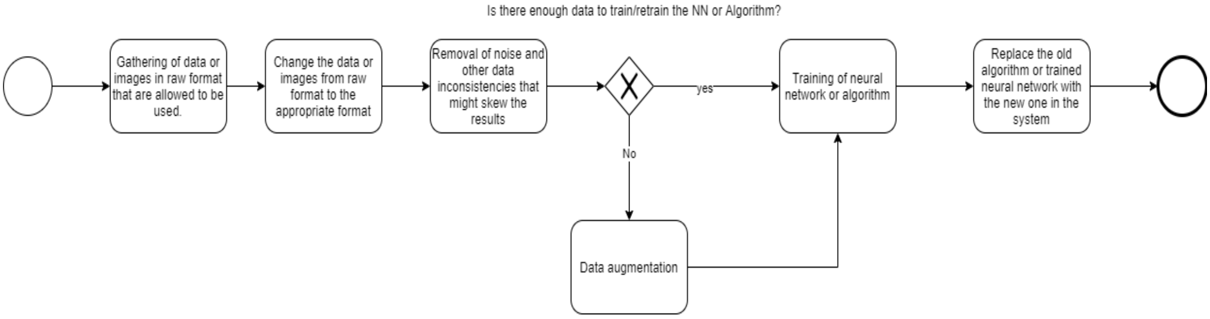

Is there enough data to train/retrain the NN or Algorithm?
Gathering of data or images in raw format that are allowed to be used.
Change the data or images from raw format to the appropriate format
Removal of noise and other data inconsistencies that might skew the results
yes
No
Training of neural network or algorithm
Replace the old algorithm or trained neural network with the new one in the system
Data augmentation

Figure 17 Training process (Source: Author created)

The process for lung cancer analysis begins with the user wanting to get analysis, diagnosis, or assistance for lung cancer. This process can be inbuilt into any of the systems that might require this both for patients and also doctors. Then the user will go to the portal that he wants to use. This can be anything from patient portal to doctors portal or any other system that we use in Estonia. After the user opens the portal then he needs to sign in with the usual process of selecting a method, entering his credentials and then the portal authenticating the user.

After that is done then the user will need to fill in a document for the request which should contain 3 main parts:

1) A request for the analysis using the algorithm.

2) Partial or all data that is required for the algorithm from the user.

3) Permissions to use the data from the side of the owner of the data.

After the document is filled in and sent then the data part of the document will need to go into the databases depending on what kind of data it is through the validator that will check for the integrity of that data, the formats, or any other requirements the databases might have for them. The requests and the permission will be delivered to the central healthcare system.

Depending on the request, the system will decide what trained algorithm or neural network will be used to make the prediction. After the algorithm is decided then the system will request the data from the databases and check if the permissions were granted to use this data and if all of the required data is present. If it is not present or the permissions have not been granted then a message will be sent to the user through the appropriate portal and their email to provide these permissions and data. If the permissions are present along with the data then the system will use the code from the training process to reformat the images or data into the proper formatting and produce an analysis or decision. This decision will be stored in the central healthcare system databases and will be sent to the user through the portal.

After the user receives this information through the portal then the process ends. The process can also end if the user refuses to provide permissions to use the data or the data itself.

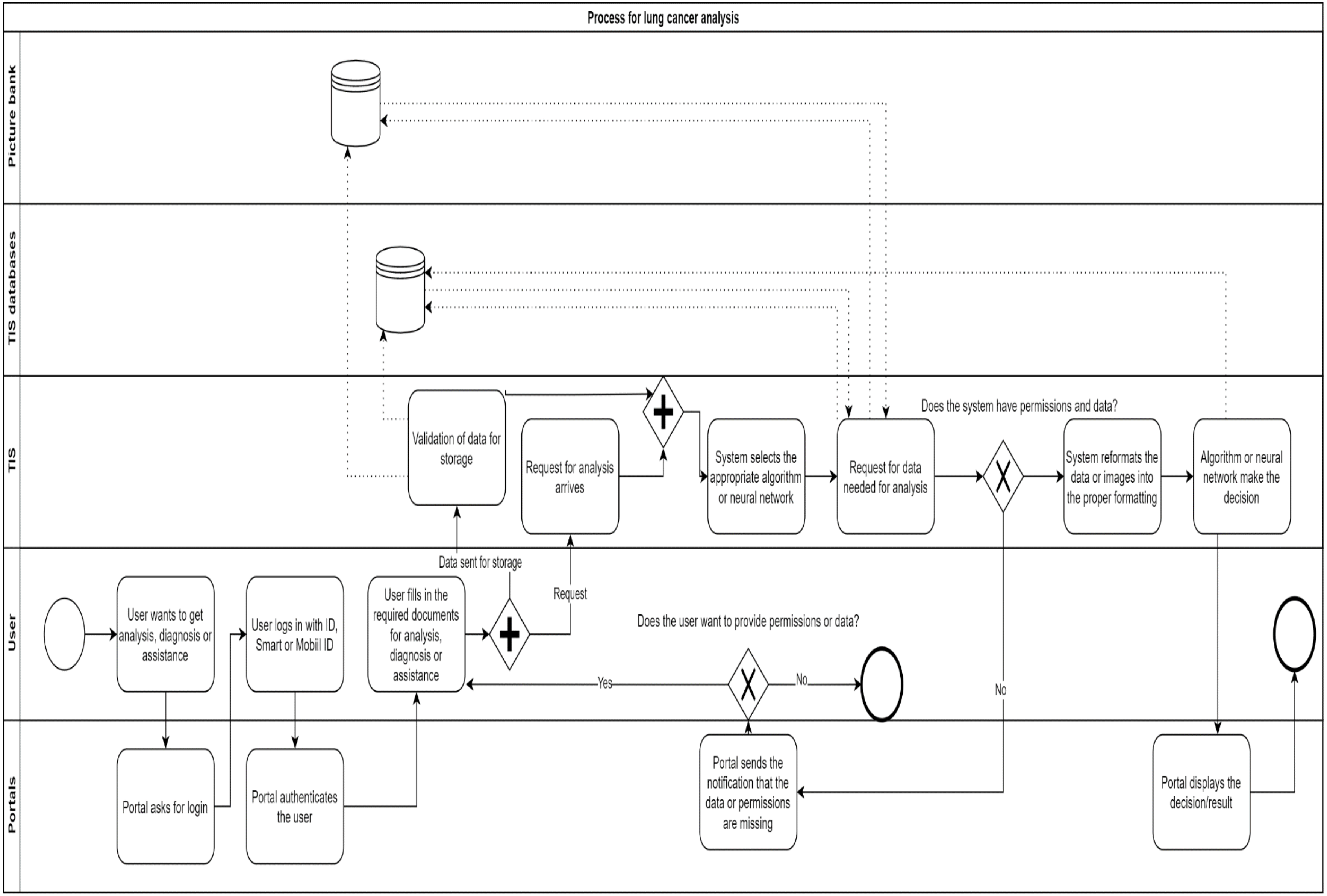

Process for lung cancer analysis
Picture bank
TIS databases
TIS
User
Portals
Validation of data for storage
Request for analysis arrives
System selects the appropriate algorithm or neural network
Request for data needed for analysis
Does the system have permissions and data?
System reformats the data or images into the proper formatting
Algorithm or neural network make the decision
Data sent for storage
Request
User wants to get analysis, diagnosis or assistance
User logs in with ID, Smart or Mobiil ID
User fills in the required documents for analysis, diagnosis or assistance
Does the user want to provide permissions or data?
Yes
No
No
Portal asks for login
Portal authenticates the user
Portal sends the notification that the data or permissions are missing
Portal displays the decision/result

Figure 18 Process for lung cancer analysis (Source: Author created)

4.6 Data extraction and training

Before the architectural and systems aspects can be discussed, the author needs to mention the process of data extraction and training of the algorithm or the NN. This process is conducted by data scientists to determine which algorithm or NN might be more efficient for the needs and purposes of the system. The problem of training can be divided in to 3 different parts:

1) Legal problems.

2) Technical problems.

3) Problems with the data.

Regarding the legal problems the author consulted with TEHIK jurist and the head of the data management department. (Laine Mokrik and Priit Raspel) Currently the legal environment is more or less clear but it is definitely not very comfortable for the usage of machine learning or neural networks. If a person wants to get some data from the Estonian state then he needs to go in front of the Research ethics committee and then explain aspects like why this data is needed and how this data will be stored. More information about this can be found in the Personal Data Protection Act Implementation Act paragraph 59 subparagraph 4.[39] Then the committee will need to make a decision and if they allow for it then the data can be provided. The problem with this legal regulation is that this system is not very flexible for situation of re-training of the algorithm. So every time new data arrives, the data scientists will need to go in front of the committee every time before access can be granted to that data for re-training purposes which is usually the general practice for such systems.

There is also another legal problem when it comes to the main process itself. The algorithm or NN can be used in two different ways. One is that the algorithm can be applied to all of the existing data as soon as the algorithm is trained and the other one being that the algorithm can be applied to data when the people actually request this to be done. While the first method is actually more efficient and will produce a lot more positive results, it is illegal according to Estonian laws. So only the 2nd method can be used which is described in the BPMN process diagram.

The extraction of the data is possible once the permissions are granted by the Research Ethics committee from a technical standpoint. The data would be extracted and then made anonymous. While there are limitations on the author on what he can discuss here, the process of extraction can still be described in a general manor. Currently there might be 3 places where the data can be extracted from. One would be done via SQL if we talk about the general health information system database. If information needs to be extracted from the Picture Bank then the information can be queried via the DICOM query/Retrieve services. Finally there is the biggest database where all the documents are kept and this database can be queried via the XQuery language that allows for manipulations with XML files.

Problems with the data can be different. The main problem might be related to the locality of the data to be used for training. It can be solved via sorting but this would require a deeper understanding on how the data is kept for Estonian citizens and citizens of other countries. Currently the databases should have documents regarding citizens of other countries which should be possible to exclude using different querying parameters that SQL, DICOM or XQuery use. This particular decision should be under the discretion of the data scientist dealing with the algorithm or the NN. The situation might be quite similar in cases when the data types will be missing or have a mismatch between each other. This can be fixed by the data scientist by conversion or exclusion. Those corrections preferably need to be done on the extraction phase but it might not always be possible since different actors manage different databases. In case of the document and SQL databases it can be done by TEHIK and should be easier but in case of the Picture bank it is more complicated due to it being under the Tartu clinic.

5 Systems and architectural analysis.

This chapter of the master's thesis is designed to answer questions regarding IT systems architecture, requirements and other IT analytical questions ending with a small prototype. The first parts will focus on the position of the algorithm within the system and requirements while the later parts of the chapter will explain the use cases and the position within the whole Estonian healthcare ecosystem itself ending with a prototype.

5.1 Where can the algorithm go architecturally

After having discussed the processes, the author thinks that to move forward with architectural and system requirements, it is needed to discuss the general layout of a lot of Estonian healthcare systems and how they could or should be connected to the process.

There are 4 categories that the author wants to bring up:

1) Databases where the data can come from or go to.

2) Users who use the system.

3) The health information system itself.

4) Portals and other programs that can either send or receive data for the purposes of the process.

All those categories also need to be subdivided and the author wants to start with the health information system. This particular system consists of many different components or microservices:

1) The central system that is under the Social ministry is the main part of the health information system. It is responsible for moving of the data from place to place within the system, storing them in the databases, performing some actions if needed and for enforcing the standards and requirements in some cases. In a sense, it is the backbone of the whole system without which it would not function.

2) Validator – Another system that is connected to the central system. The basic flow of the data goes through the validator into the central system. The main idea behind the validator is to give access to the data to the appropriate people and to validate the data itself for integrity, correctness, standards and so on depending on the requirements of the social ministry, system, or the law. The difference between the central system errors and validator errors are that the central system errors are usually server errors while the validator errors are errors produced by specific rule sets or requirements.

3) The security server – this particular system is responsible for safeguarding the system from unauthorized access and other forms of security threats. The information flows through the security server and validator into the central system. Most of the people know this security server as X-tee. So most of the external communication is done by the security server.

4) Other systems that are not used for the purposes of the process like data balancer or data balancer API etc.

The principle behind most of the governmental systems used for different kind of governmental services provided by the Estonian government is called data embassy. This means that the databases and information is very decentralised and there is no one large database that holds all of the information. For the intents and purposes of the process, these are the important databases:

1) Healthcare information system databases that are stored by TEHIK and other government services.

2) Picture bank which is managed by the Tartu Clinic right now.

3) Other external databases where the data can come from if requested which is not limited only to databases located in Estonia but also the ones located in other countries that the eco system has access to via the X-tee services.

Then there are the users who can be:

1) Regular people

2) Doctors and other medical professionals.

3) Private enterprises who develop or use healthcare related software.

And finally there are portals and the software, which are numerous and even more of them will be added in the future as time goes on. Some of the more notable ones are:

1) Digilugu.ee

2) Arstiportaal.ee

3) Family doctor software

4) Mobile and PC versions of e-first aid.

5) Data viewer (Currently in development)

All these systems have to function together with the algorithm to make sense and be useful for the greater benefit of the whole healthcare system of Estonia. How they currently function together is that the user starts his actions through one of the portals or any of the software. The user needs to usually authenticate using either the ID card or any other option available and then he initiates an action be it accessing of some data or performing some action. This sends a SOAP request from the portal or the software to the central system through the security server and the validator. The system finds the data or performs the changes that are needed and then returns the information through the security server and in SOAP format to the portal so the portal can display either that the action has been performed or show the data.

Under these conditions, the algorithm can exist as an external system in a form of a software or webservice or as an internal microservice within the health information system itself. There are benefits to both.

If the algorithm is a microservice of the health information system then the benefits are the following:

1) Handling of the data is done within government servers which is more secure.

2) This will create infrastructure that the private enterprises can re-use for the purposes of training or accessing the data if they want to conduct similar research either for machine learning research in healthcare or pure medical research.

3) This will create infrastructure and experience for TEHIK and other government institutions in regard to managing of processing power for training of the algorithm or neural networks.

4) This will create infrastructure where the algorithms can be stored and function.

If the algorithm is an external system then the benefits are the following:

1) The time to conduct improvement or change will be faster because the external system will be less complicated in general and will not have to follow strict guidelines on when and how can the changes be done.

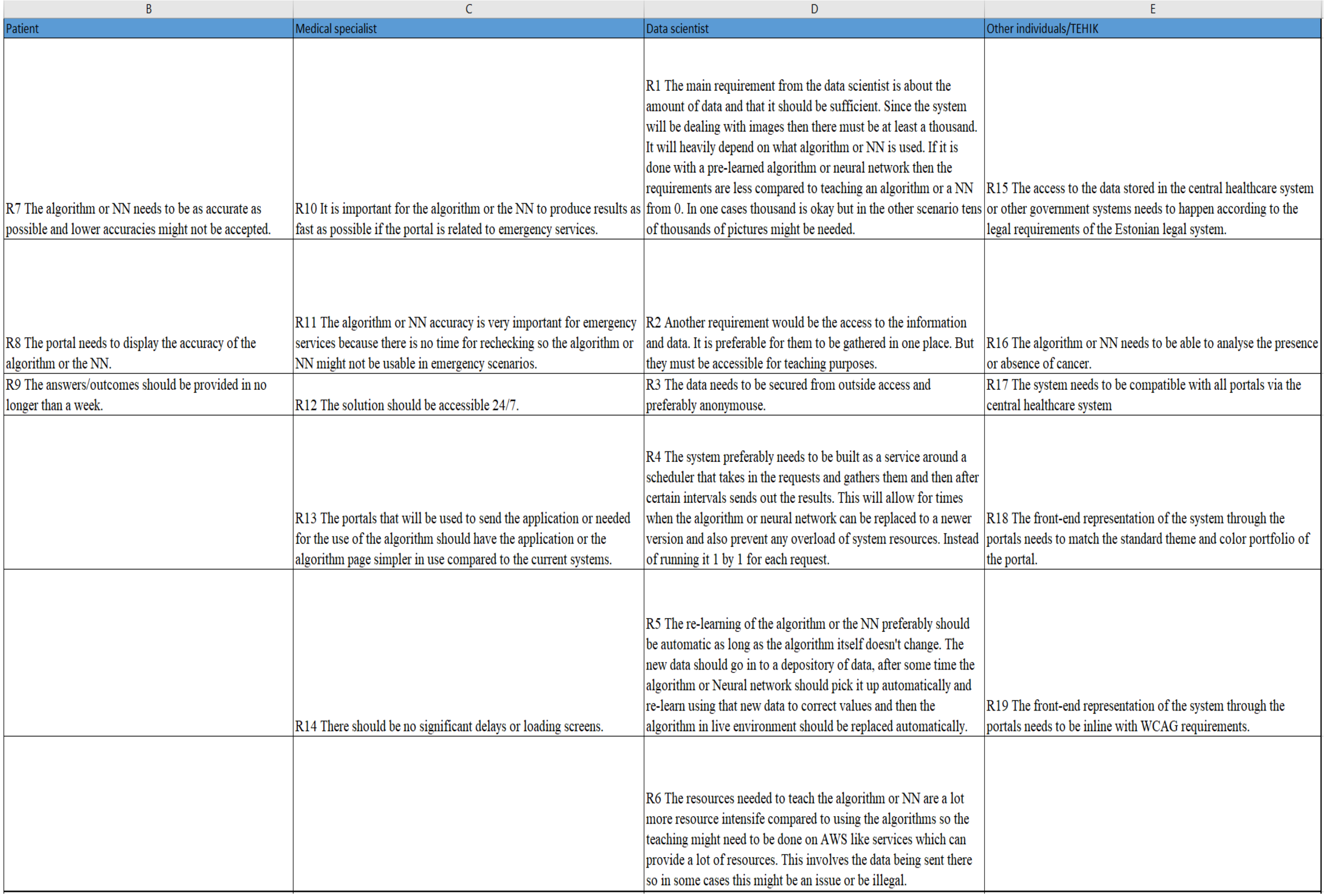
2) It is much easier to give access to the data and invite other developers to participate in the development of an external system.

In the authors opinion, the best approach here is to develop at least the starting algorithm as an internal component or microservice of the health information system because it will allow the government institutions and TEHIK to create the infrastructure that solves some of the needs of the private enterprises in this regard and which should allow them to conduct private development of such algorithms or neural networks in the future.

5.2 Full requirements

Here are the requirements that were gathered during the talks and interviews with different stakeholders from their ideas, opinions, wishes and desires that might be important for this project. The stakeholders are patients/regular people, doctors, nurses, data scientists and other individuals from TEHIK.

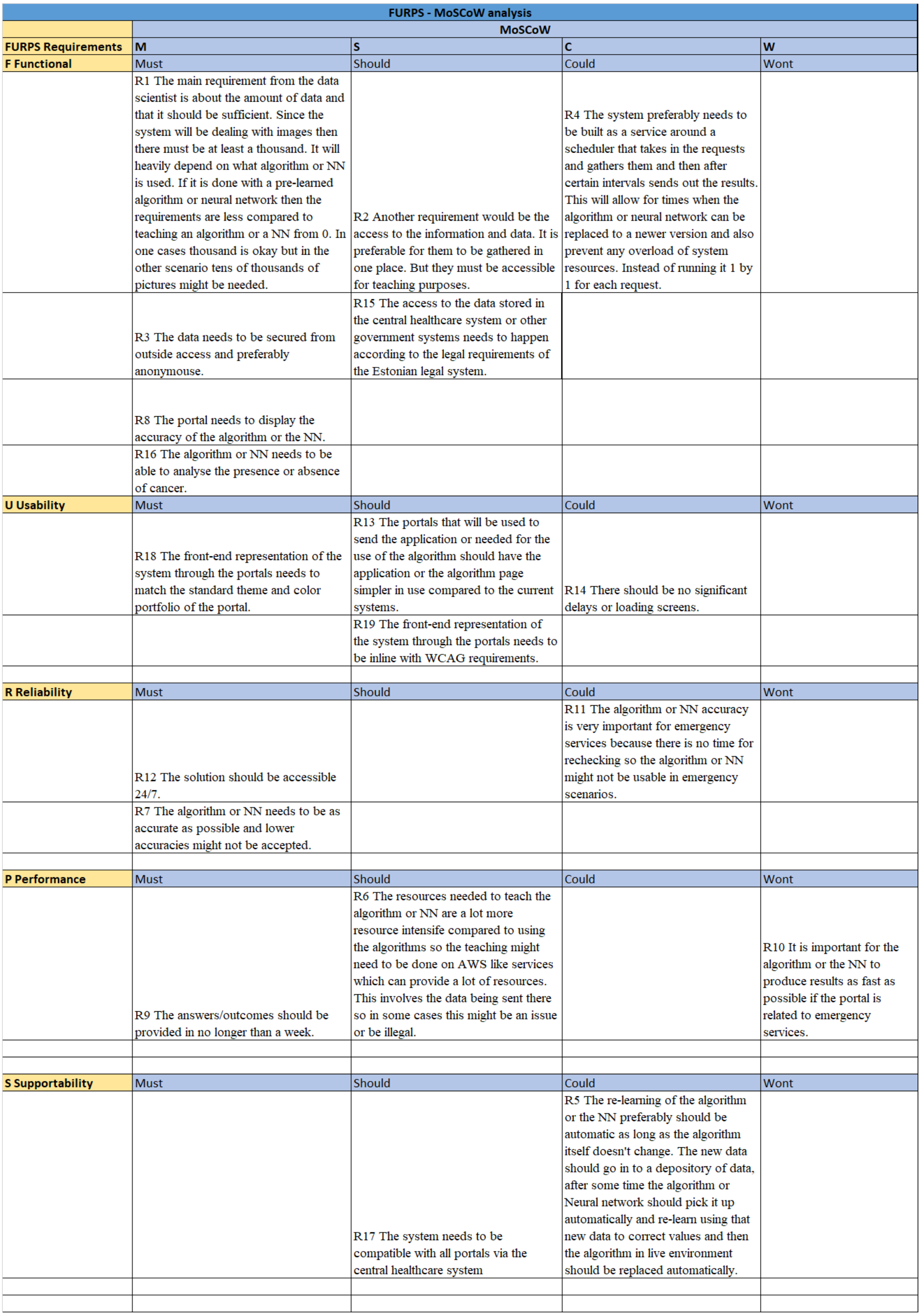
As the interviews have shown, medical specialists and patients/average people have no strong opinion regarding such a system besides it being a good idea and some requirements for performance and usability while both of those categories do not seem to care about their data being used for teaching purposes and improvement of the algorithms or NNs. The author fully acknowledges that such an outcome is not representative of the whole society in regard to the usage of the data. Data scientists on the other hand were more specific about requirements for the data instead of functionality or usability requirements for the system.



B	C	D	E
Patient	Medical specialist	Data scientist	Other individuals/TEHIK
R7 The algorithm or NN needs to be as accurate as possible and lower accuracies might not be accepted.	R10 It is important for the algorithm or the NN to produce results as fast as possible if the portal is related to emergency services.	R1 The main requirement from the data scientist is about the amount of data and that it should be sufficient. Since the system will be dealing with images then there must be at least a thousand. It will heavily depend on what algorithm or NN is used. If it is done with a pre-learned algorithm or neural network then the requirements are less compared to teaching an algorithm or a NN from 0. In one cases thousand is okay but in the other scenario tens of thousands of pictures might be needed.	R15 The access to the data stored in the central healthcare system or other government systems needs to happen according to the legal requirements of the Estonian legal system.
R8 The portal needs to display the accuracy of the algorithm or the NN.	R11 The algorithm or NN accuracy is very important for emergency services because there is no time for rechecking so the algorithm or NN might not be usable in emergency scenarios.	R2 Another requirement would be the access to the information and data. It is preferable for them to be gathered in one place. But they must be accessible for teaching purposes.	R16 The algorithm or NN needs to be able to analyse the presence or absence of cancer.
R9 The answers/outcomes should be provided in no longer than a week.	R12 The solution should be accessible 24/7.	R3 The data needs to be secured from outside access and preferably anonymouse.	R17 The system needs to be compatible with all portals via the central healthcare system
	R13 The portals that will be used to send the application or needed for the use of the algorithm should have the application or the algorithm page simpler in use compared to the current systems.	R4 The system preferably needs to be built as a service around a scheduler that takes in the requests and gathers them and then after certain intervals sends out the results. This will allow for times when the algorithm or neural network can be replaced to a newer version and also prevent any overload of system resources. Instead of running it 1 by 1 for each request.	R18 The front-end representation of the system through the portals needs to match the standard theme and color portfolio of the portal.
	R14 There should be no significant delays or loading screens.	R5 The re-learning of the algorithm or the NN preferably should be automatic as long as the algorithm itself doesn't change. The new data should go in to a depository of data, after some time the algorithm or Neural network should pick it up automatically and re-learn using that new data to correct values and then the algorithm in live environment should be replaced automatically.	R19 The front-end representation of the system through the portals needs to be inline with WCAG requirements.
		R6 The resources needed to teach the algorithm or NN are a lot more resource intensife compared to using the algorithms so the teaching might need to be done on AWS like services which can provide a lot of resources. This involves the data being sent there so in some cases this might be an issue or be illegal.	

Figure 19 Requirements (Source: Author created)

The prioritization of those requirements along the lines of FURPS + MoSCoW looks like this:



FURPS - MoSCoW analysis				
	MoSCoW			
FURPS Requirements	M	S	C	W
F Functional	Must	Should	Could	Wont
	R1 The main requirement from the data scientist is about the amount of data and that it should be sufficient. Since the system will be dealing with images then there must be at least a thousand. It will heavily depend on what algorithm or NN is used. If it is done with a pre-learned algorithm or neural network then the requirements are less compared to teaching an algorithm or a NN from 0. In one cases thousand is okay but in the other scenario tens of thousands of pictures might be needed.	R2 Another requirement would be the access to the information and data. It is preferable for them to be gathered in one place. But they must be accessible for teaching purposes.	R4 The system preferably needs to be built as a service around a scheduler that takes in the requests and gathers them and then after certain intervals sends out the results. This will allow for times when the algorithm or neural network can be replaced to a newer version and also prevent any overload of system resources. Instead of running it 1 by 1 for each request.	
	R3 The data needs to be secured from outside access and preferably anonymouse.	R15 The access to the data stored in the central healthcare system or other government systems needs to happen according to the legal requirements of the Estonian legal system.		
	R8 The portal needs to display the accuracy of the algorithm or the NN.			
	R16 The algorithm or NN needs to be able to analyse the presence or absence of cancer.			
U Usability	Must	Should	Could	Wont
	R18 The front-end representation of the system through the portals needs to match the standard theme and color portfolio of the portal.	R13 The portals that will be used to send the application or needed for the use of the algorithm should have the application or the algorithm page simpler in use compared to the current systems.	R14 There should be no significant delays or loading screens.	
		R19 The front-end representation of the system through the portals needs to be inline with WCAG requirements.		
R Reliability	Must	Should	Could	Wont
	R12 The solution should be accessible 24/7.		R11 The algorithm or NN accuracy is very important for emergency services because there is no time for rechecking so the algorithm or NN might not be usable in emergency scenarios.	
	R7 The algorithm or NN needs to be as accurate as possible and lower accuracies might not be accepted.			
P Performance	Must	Should	Could	Wont
	R9 The answers/outcomes should be provided in no longer than a week.	R6 The resources needed to teach the algorithm or NN are a lot more resource intensife compared to using the algorithms so the teaching might need to be done on AWS like services which can provide a lot of resources. This involves the data being sent there so in some cases this might be an issue or be illegal.		R10 It is important for the algorithm or the NN to produce results as fast as possible if the portal is related to emergency services.
S Supportability	Must	Should	Could	Wont
		R17 The system needs to be compatible with all portals via the central healthcare system	R5 The re-learning of the algorithm or the NN preferably should be automatic as long as the algorithm itself doesn't change. The new data should go in to a depository of data, after some time the algorithm or Neural network should pick it up automatically and re-learn using that new data to correct values and then the algorithm in live environment should be replaced automatically.	

Figure 20 FURPS - MoSCoW analysis (Source: Author created)

The main reason behind adding R10 in to the “Won’t” category is because the initial system cannot be planned for all possible portals. If the emergency assistance portals require an answer that is both fast and extremely accurate these particular requirements can be implemented at a later point or might not be feasible at all. For example, the accuracy might not reach the point where emergency services will consider it adequate.

All other requirement prioritization is straightforward in the opinion of the author with the exception of WCAG requirements(R19). These requirements come from European Union legislature which make those WCAG requirements obligatory for all government websites all across the European Union. Majority of the government websites in the European Union are not yet WCAG compatible. This requirement also heavily depends on the portal itself so if the portal is not compatible with WCAG right now then fulfilling this requirement might not be possible if we are planning to use the portal along with the algorithm.

5.3 Use Cases

After the examination of the process and the functionality, the author was able to create the following use cases:

1) UC1 Sign in.

2) UC2 Filling in of the application to process the data.

3) UC3 Providing permissions to use the data.

4) UC4 Providing the data.

5) UC5 Processing the data.

6) UC6 View the result.

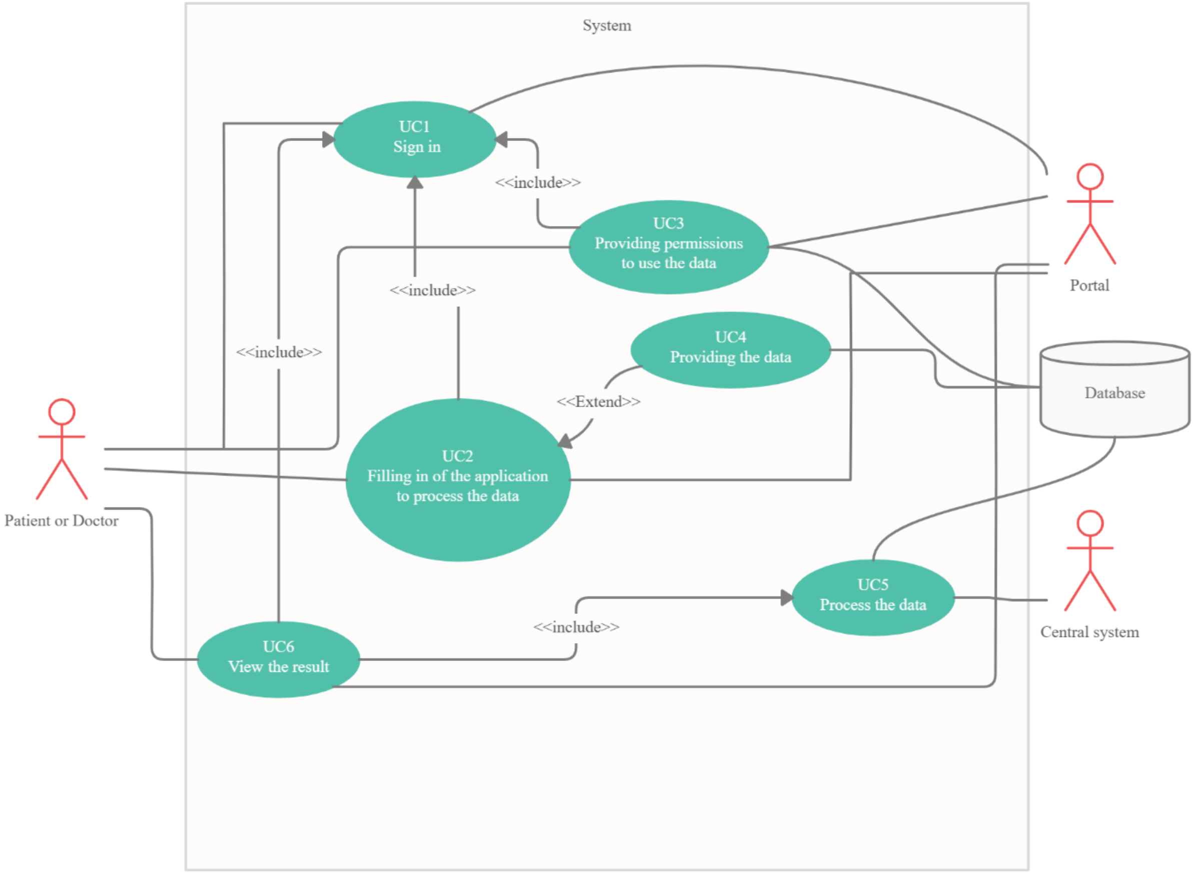


Figure 21 Use Case diagram (Source: Author created)

Name	UC1 Sign in
Description	The user wants to sign in into the system.
Actors	Patient or Doctor, Portal.
Pre-conditions	The user is not signed in.
Post-conditions	The user is authenticated by the system and the system allows access to the user to perform specific actions that are not available without authentication.
Main flow	1) User opens the website. 2) User clicks on sign in. 3) User chooses to sign in with ID card method. 4) User clicks sign in. 5) System asks to enter the Pin code and the user enters it. 6) User clicks "OK".

Alternative flow	AC1: 1) User opens the website. 2) User clicks on sign in. 3) User chooses to sign in with Mobile ID method. 4) User enters his mobile phone and ID code number. 5) User clicks to sign in. 6) User receives a code on his mobile phone and then enters it on the website. 7) User clicks "OK" AC2: 1) User opens the website. 2) User clicks on sign in. 3) User chooses to sign in with SMART ID method 4) User enters his ID code number. 5) User clicks to sign in. 6) User receives a code on his mobile phone and then enters it on the website. 7) User clicks "OK" AC3: 1) User opens the website. 2) User clicks on sign in. 3) User chooses any sign in method. 4) User enters the required details 5) User clicks to sign in. 6) The system informs the user that the sign in has failed. 7) User clicks "OK" 8) The user is returned to the sign in screen

Figure 22 UC1 Sign in. (Source: Author created)

Name	UC2 Filling in the application to process the data
Description	The user wants to fill in the application to allow his data to be processed.
Actors	Patient or Doctor, Portal.
Pre-conditions	Actor is signed in.
Post-conditions	The application is filled in and the system responds that it has been accepted and sent.

Main flow	1) User is located on the main page. 2) User selects data processing. 3) User selects the box to fill in the application. 4) User fills in the information, provides permissions or attaches the needed information. 5) User confirms the application with a digital signature.
Alternative flow	AC1: 1) User is located on the main page. 2) User selects data processing. 3) User selects the box to fill in the application. 4) User fills in the information, provides permissions or attaches the needed information. 5) User confirms the application with a digital signature. 6) User receives an error message with a description of what exactly went wrong.

Figure 23 UC2 Filling in the application to process the data (Source: Author created)

Name	UC3 Providing permissions to use the data
Description	The user needs to provide permissions to the system to process and use the data to produce an outcome. This is not a subcase of filling in the application because it can happen either when filling in the application or if the system requires these permissions and asks for them without an application.
Actors	Doctor or Patient, Database, Portal.
Pre-conditions	The user is signed in.
Post-conditions	The permissions to use the data are granted and saved in the database.
Main flow	1) The user receives a message that he needs to provide permissions to process his data. 2) The user clicks on the message or goes to "Managing accesses to health data". 3) User clicks on edit in the "Access to data for machine learning purposes" box. 4) User clicks on sign and enters his digital signature credentials. 5) User clicks "ok". 6) The system says that the information has been updated.

Alternative flow	AC1: 1) User starts filling in the application to process the data. 2) User agrees by ticking on a checkbox for his data to be used for machine learning purposes. 3) User digitally signs the application.

Figure 24 UC3 Providing permissions to use the data (Source: Author created)

Name	UC4 Providing the data
Description	The user needs to provide the data to be processed if it is missing in the system. This is a subcase of filling in the application.
Actors	Doctor or Patient, Database.
Pre-conditions	The user is signed in
Post-conditions	Data is provided and stored in the database.
Main flow	1) User starts filling in the application to process the data. 2) User provides the requested information if the information is not present in the database. 3) User digitally signs the application.
Alternative flow	None

Figure 25 UC4 Providing permissions to use the data (Source: Author created)

Name	UC5 Process the data

Description	The algorithm or NN processes the data to produce a result.
Actors	Central system, Database.
Pre-conditions	Data is present in the database.
Post-conditions	The results are stored in the database.
Main flow	1) The data is extracted from the database. 2) The data is pre-processed and fed into the algorithm or NN. 3) The algorithm or NN produce a result. 4) The result is sent to the database.
Alternative flow	None

Figure 26 UC5 Process the data (Source: Author created)

Name	UC6 View the result.
Description	The user wants to view the results that the algorithm or NN came to.
Actors	Doctor or Patient, Portal.
Pre-conditions	The data has been processed; the user is signed in.
Post-conditions	The result is visible and the user can view it.
Main flow	1) The user is located on the main page. 2) The user clicks on examination results. 3) The user clicks on the result. 4) The results are displayed in the standard analysis format.
Alternative flow	None

Figure 27 UC6 View the result (Source: Author created)

5.4 C4 Model

In here the author wanted to talk about the general architecture of the new system while dividing it between context view, container view, component view and then perhaps even the code. Currently there are legal limitations of the information being secret and protected by the state due to security concerns so there is limited amount of information that the author can provide on the subject. Having discussed the situation with the administrators and architects, most of the information beyond the context diagram is secret so the author will have to describe the situation with words instead of drawing a full diagram.

The general look for all of the systems connected to the IT healthcare ecosystem is:

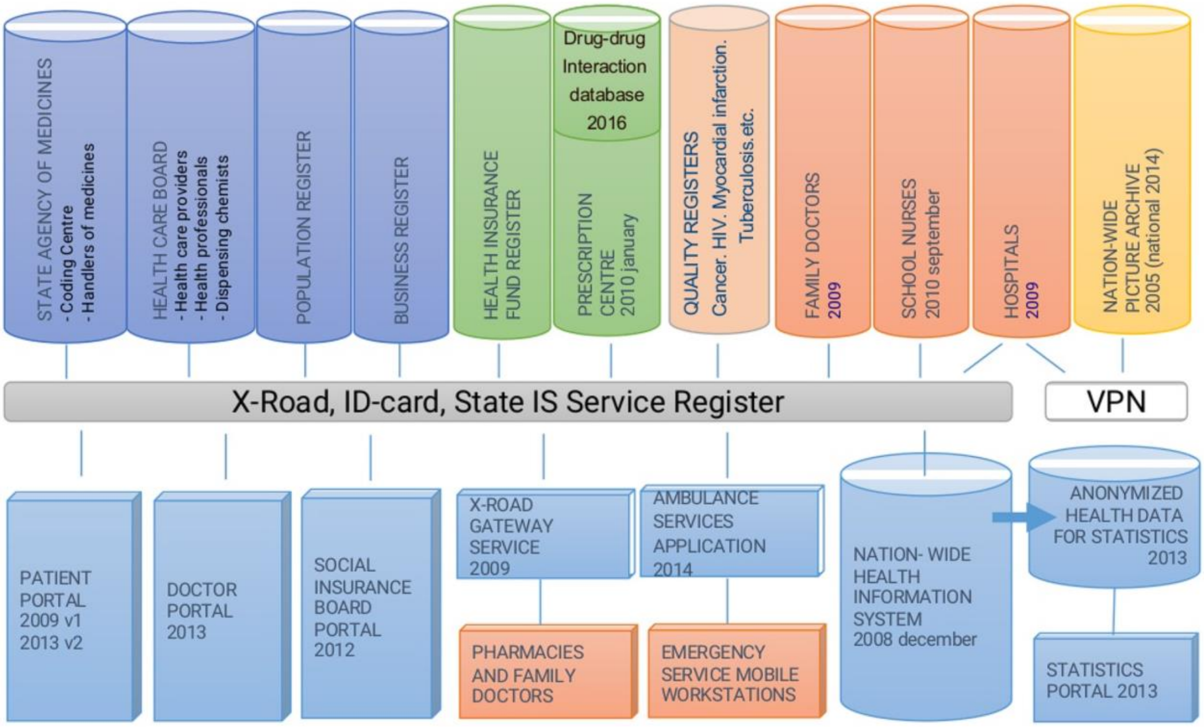

STATE AGENCY OF MEDICINES
- Coding Centre
- Handlers of medicines
HEALTH CARE BOARD
- Health care providers
- Health professionals
- Dispensing chemists
POPULATION REGISTER
BUSINESS REGISTER
HEALTH INSURANCE FUND REGISTER
Drug-drug Interaction database 2016
PRESCRIPTION CENTRE 2010 january
QUALITY REGISTERS
Cancer. HIV. Myocardial infarction. Tuberculosis.etc.
FAMILY DOCTORS 2009
SCHOOL NURSES 2010 september
HOSPITALS 2009
NATION-WIDE PICTURE ARCHIVE 2005 (national 2014)
X-Road, ID-card, State IS Service Register
VPN
PATIENT PORTAL 2009 v1 2013 v2
DOCTOR PORTAL 2013
SOCIAL INSURANCE BOARD PORTAL 2012
X-ROAD GATEWAY SERVICE 2009
AMBULANCE SERVICES APPLICATION 2014
PHARMACIES AND FAMILY DOCTORS
EMERGENCY SERVICE MOBILE WORKSTATIONS
NATION- WIDE HEALTH INFORMATION SYSTEM 2008 december
ANONYMIZED HEALTH DATA FOR STATISTICS 2013
STATISTICS PORTAL 2013

Figure 28 Estonian healthcare IT ecosystem (Source: Artur Novek, https://www.slideshare.net/igorbossenko/overview-of-estonian-health-information-system)

Most of the systems are connected together through the x-road system. This system has been mentioned before as the security server. X-road consists of 2 components, one is the central X-road system that is used for secure communications and the other is a gateway that is usually a part of every other system which is used to connect to the central X-road system.

The only system that is not connected to the X-road central system is the national picture archive which is the old name for the Picture bank which the author was discussing before. Right now this means that the Picture bank is not directly accessible to the other systems besides the hospitals using a VPN. So this situation will need to be resolved also by either connecting the picture bank to the x-road or establishing a separate VPN connection between the bank and the national health information system.

The context view for the proposed system would look like the following:

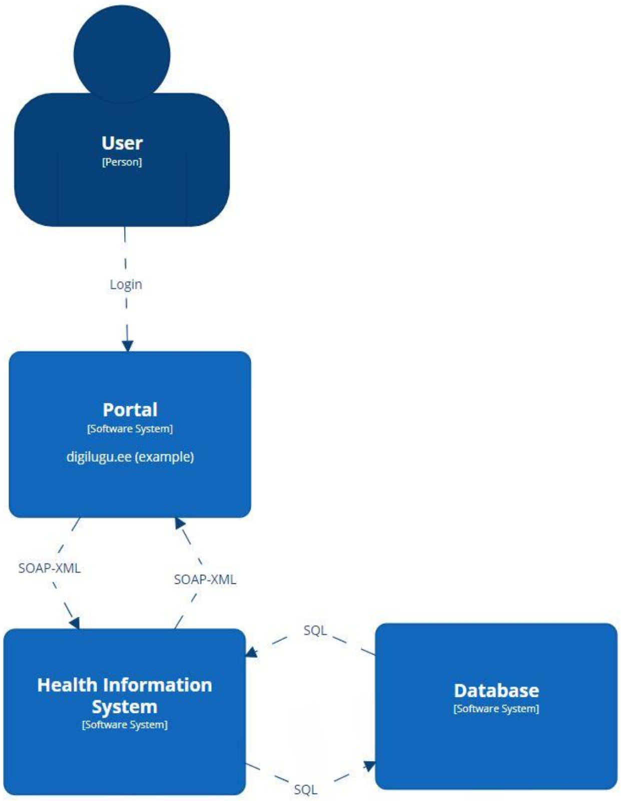


Figure 29 Context view (Source: Author created)

From the standpoint of the user he will be login into a portal of any sort. This would happen using the current methods that are used: ID card, Mobile ID, or Smart ID. In some cases logins through the bank are also allowed but they are not always implemented. The information then would move through the X-road system to the Health Information System after the permissions are granted or the application is filled.

This information would use the standard Soap protocol with XML HL8 format. After that the health information system can decide where it wants to store the data depending on the situation. Usually it would be the systems database which is accessed using SQL.

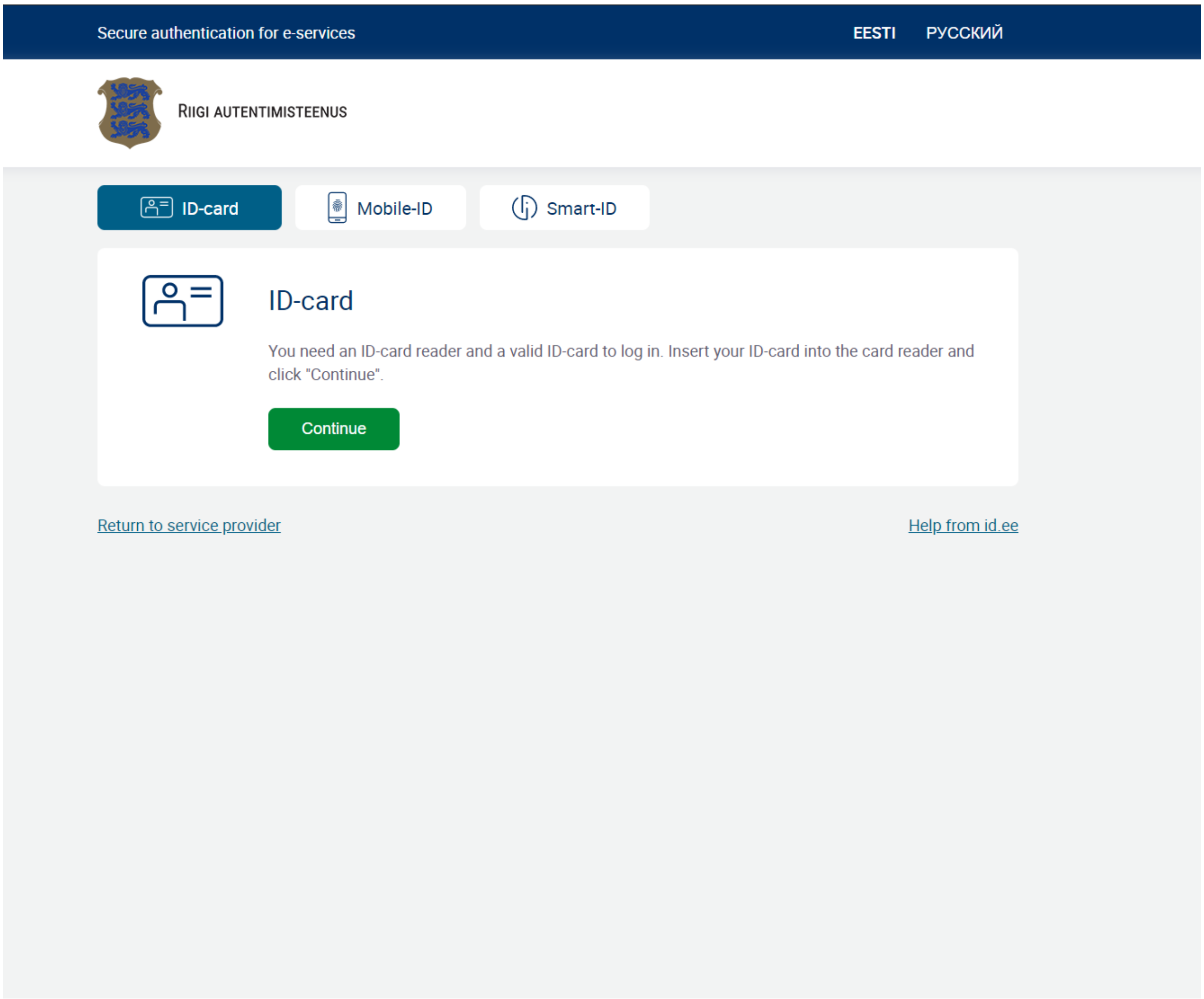
The algorithm would be a container or a packet within the health information system. It should be able to communicate with the databases by itself using the same mechanisms as the health information system. It should also be able to access the Picture bank using the x-road system.

The components of the system would be the algorithm itself, information pre-processor that should be able to take the data from the Picture Bank and pre-process it to the format that the algorithm can understand. As mentioned before, this would convert DICOM files and the information in them by taking out the image and then making it in to a standardized .JPEG file or any other format that might be preferable to the algorithm. Once the algorithm is done with its calculation and prediction then the result can be stored in the database using SQL.

5.5 Prototype

This prototype will use digilugu.ee as the foundation for the portal. Currently the interface of the patient portal can be changed a bit to suit the needs but the general systems can remain the same.

Login can be done via the state service that is developed by SK ID Solution called TARA:



Secure authentication for e-services

EESTI РУССКИЙ

RIIGI AUTENTIMISTEENUS

ID-card Mobile-ID Smart-ID

ID-card

You need an ID-card reader and a valid ID-card to log in. Insert your ID-card into the card reader and click "Continue".

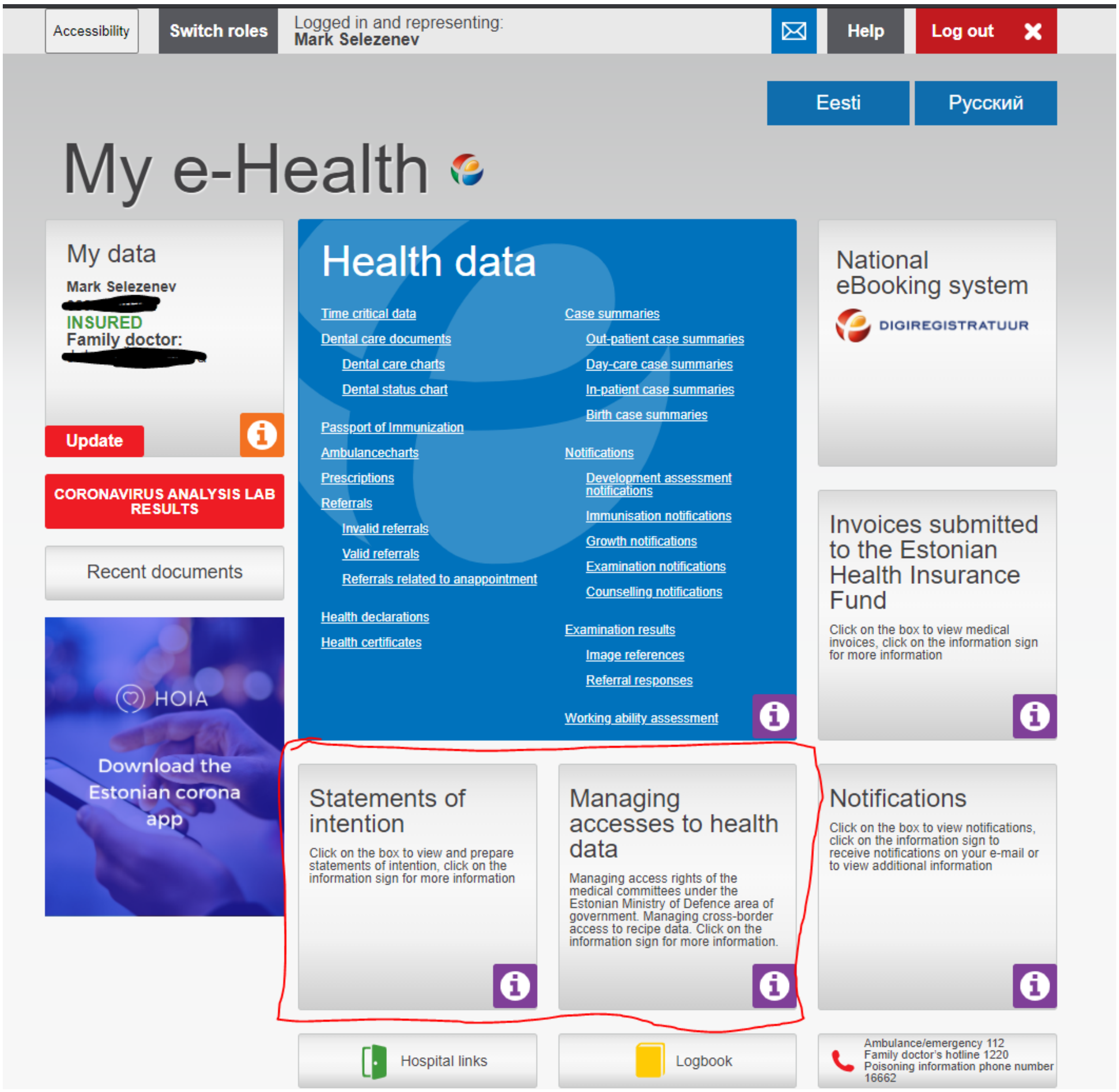
Continue

Return to service provider

Help from id.ee

Figure 30 Sign in page (Source: Digilugu.ee)

After the user is logged in then he will be redirected to the main screen that already contains the subcategories that are needed.



Logged in and representing:
Mark Selezenev

My data

Mark Selezenev
[redacted]
INSURED
Family doctor:
[redacted]

Invoices submitted to the Estonian Health Insurance Fund

Click on the box to view medical invoices, click on the information sign for more information

Statements of intention

Click on the box to view and prepare statements of intention, click on the information sign for more information

Managing accesses to health data

Managing access rights of the medical committees under the Estonian Ministry of Defence area of government. Managing cross-border access to recipe data. Click on the information sign for more information.

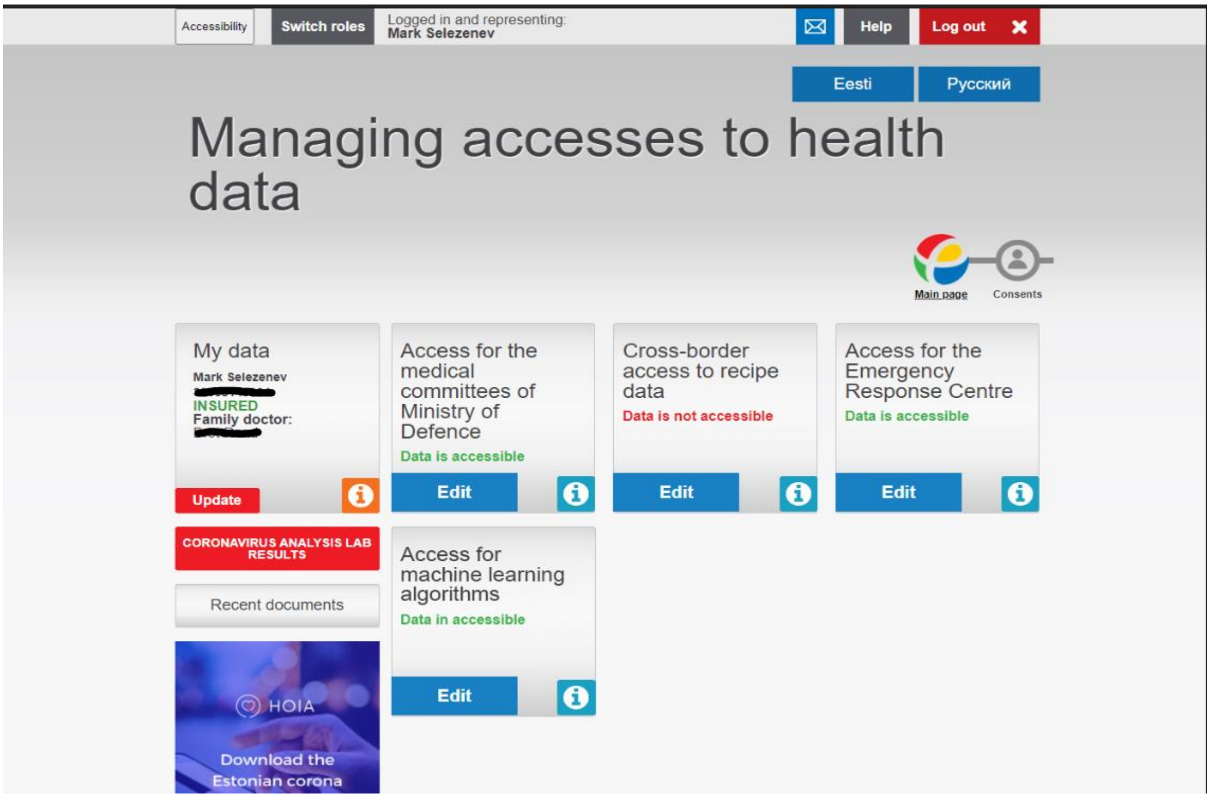
Notifications

Click on the box to view notifications, click on the information sign to receive notifications on your e-mail or to view additional information

Ambulance/emergency 112
Family doctor's hotline 1220
Poisoning information phone number 16662

Figure 31 Main page (Source: digilugu.ee)

Managing accesses to health data can be used to manage the permissions of the systems and in this case it can be used to enable or disable access for the algorithms or NN to the users personal data. Statement of intention can be used to start the process of filling in the application. The modified screens can look like this:



Logged in and representing: Mark Selezenev

Managing accesses to health data

My data
Mark Selezenev
[redacted]
INSURED
Family doctor:
[redacted]

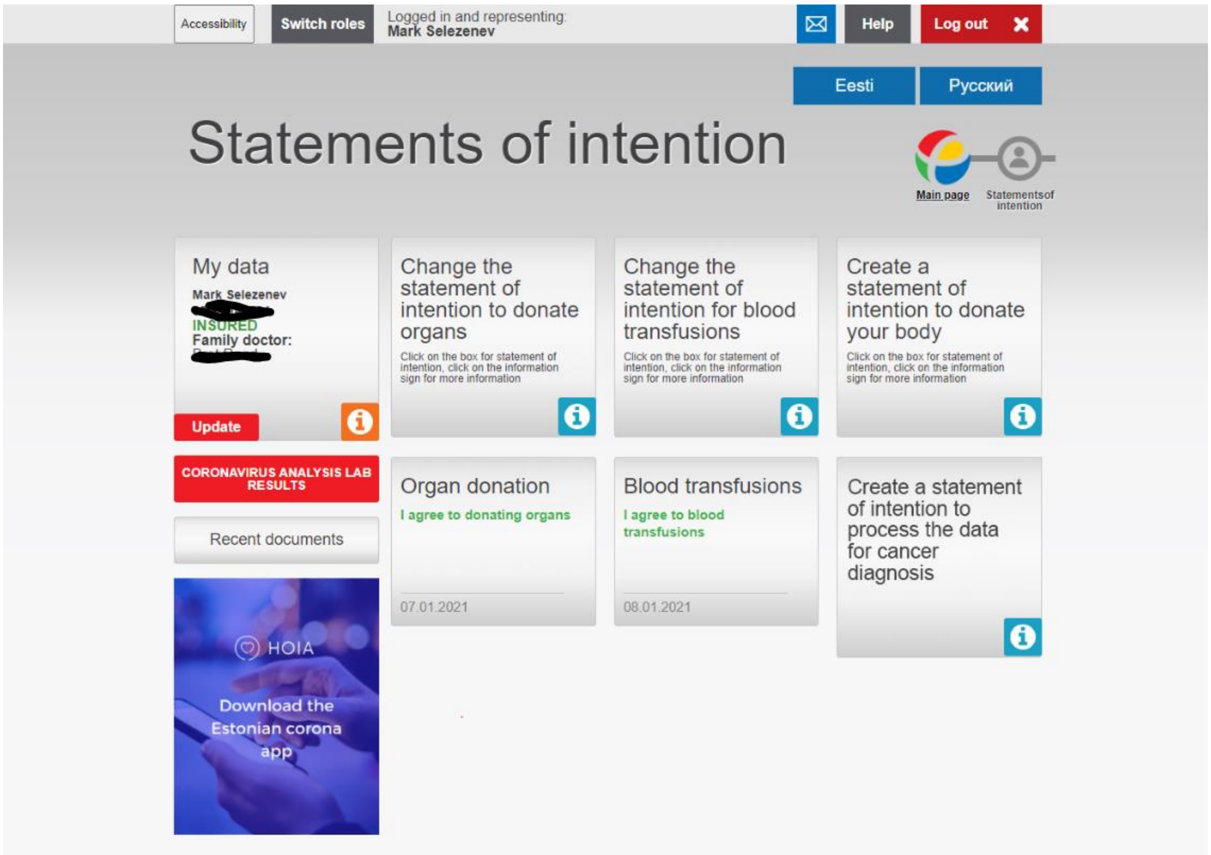
Access for the medical committees of Ministry of Defence
Data is accessible

Cross-border access to recipe data
Data is not accessible

Access for the Emergency Response Centre
Data is accessible

Access for machine learning algorithms
Data in accessible

Figure 32 Managing accesses to health data (Source: Author created)



Logged in and representing: Mark Selezenev

Statements of intention

My data
Mark Selezenev
[redacted]
INSURED
Family doctor:
[redacted]

Change the statement of intention to donate organs
Click on the box for statement of intention, click on the information sign for more information

Change the statement of intention for blood transfusions
Click on the box for statement of intention, click on the information sign for more information

Create a statement of intention to donate your body
Click on the box for statement of intention, click on the information sign for more information

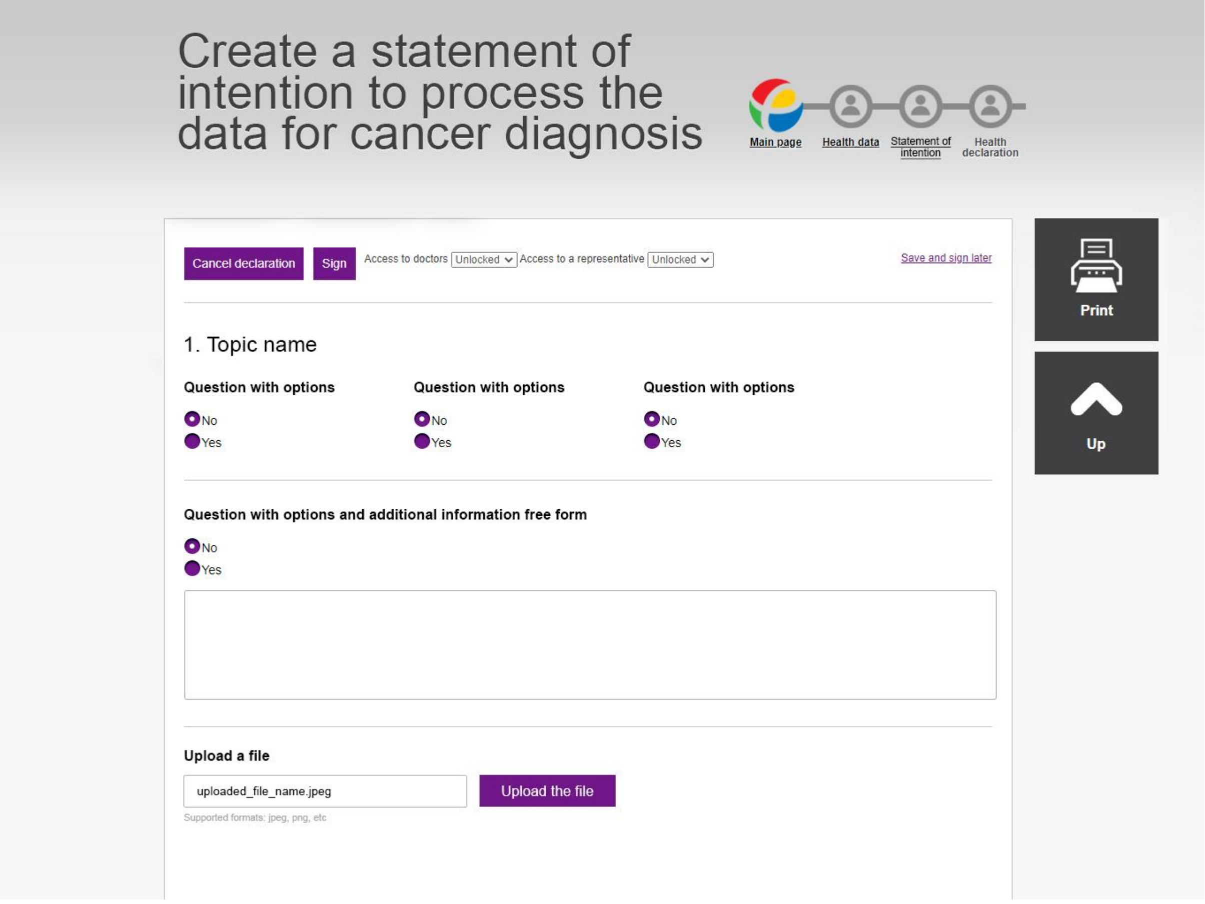
Organ donation
I agree to donating organs
07.01.2021

Blood transfusions
I agree to blood transfusions
08.01.2021

Create a statement of intention to process the data for cancer diagnosis

Figure 33 Statement of intentions (Source: Author created)

In case of managing accesses to health data, the functionality can remain the same as it is. The user clicks on the “Edit” button and then he can digitally sign to grant or remove the access for certain categories that are listed. The functionality under statements of intentions will be a bit different from the other boxes. When clicked, it will bring up a form that can be filled in which will contain the standard elements that digilugu.ee uses. The form will request the required data and will allow to attach pictures or files if needed. The form can look like the following:



Create a statement of intention to process the data for cancer diagnosis

Access to doctors Unlocked Access to a representative Unlocked

1. Topic name

Question with options
○ No
● Yes

Question with options
○ No
● Yes

Question with options
○ No
● Yes

Question with options and additional information free form
○ No
● Yes

Upload a file

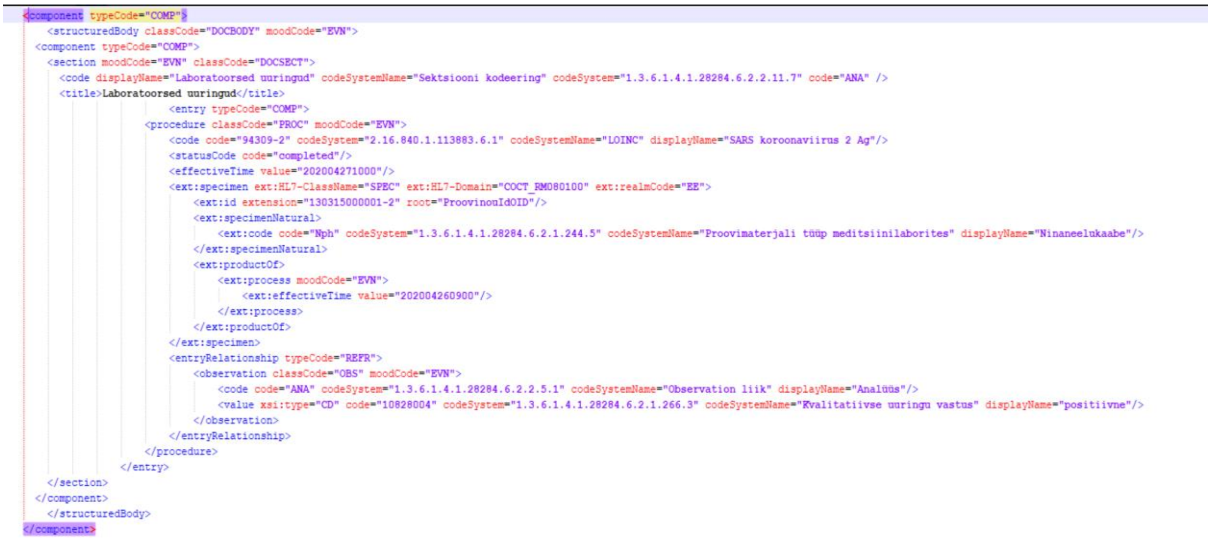
uploaded_file_name.jpeg

Supported formats: jpeg, png, etc

Figure 34 Create a statement of intention to process the data (Source: Author created)

After the form is filled in and digitally signed then the information can be sent to the appropriate places. If the information needs to go to the standard health information system database then it can be done via SQL and if it needs to go to the picture bank then it should be possible to do via the x-road system. Once the information along with the permissions and the application are in the system then the algorithm can start to fetch the data from the places where it is stored and pre-processing it from the DICOM format to the JPEG format as discussed in the previous chapters.

Once the information is there and pre-processed then the algorithm can start its work. The algorithm can be the GLCM, LBP, Colour Features classifier using Support Vector Machines from chapter 4.5. The decision on what specific algorithm or NN to use should be given to the data scientist who will be in charge of this task due to the amount of variations that are in existence. Once the result is produced then the information can be sent back to the main health information system database via the standard XML HL8 format. For this purpose slightly modified referral response can be used. The XML that is sent to the system is standardized and can be located in the standard repository which is publicly accessible. The only part that would need to be changed is the LOINC code and how TEHIK wants that LOINC code name to be displayed. LOINC code part falls under the "Laboratoorsed uuringud" component. Below is an example that was done for Coronavirus.



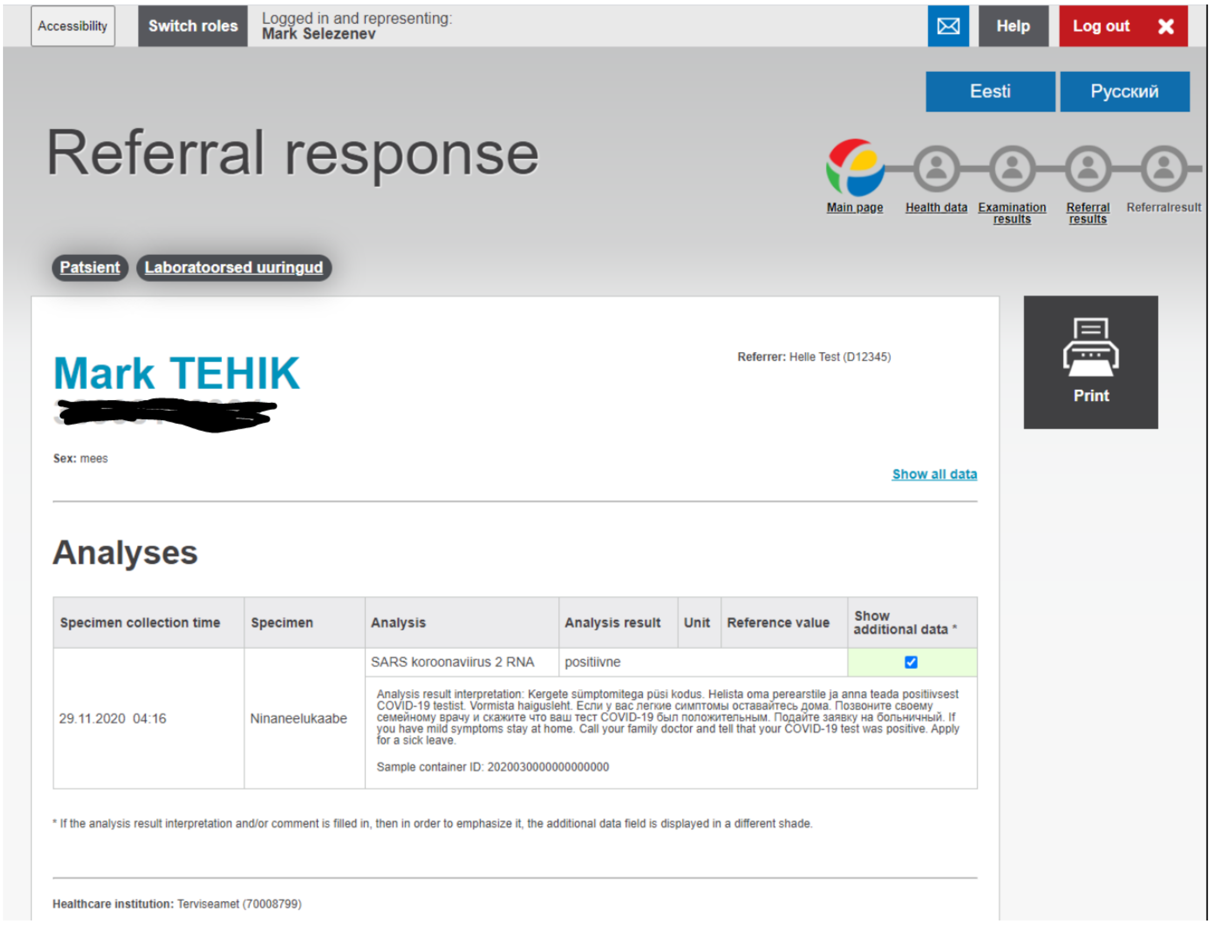
```xml
<component typeCode="COMP">
    <structuredBody classCode="DOCBODY" moodCode="EVN">
  <component typeCode="COMP">
    <section moodCode="EVN" classCode="DOCSECT">
      <code displayName="Laboratoorsed uuringud" codeSystemName="Sektsiooni kodeering" codeSystem="1.3.6.1.4.1.28284.6.2.2.11.7" code="ANA" />
      <title>Laboratoorsed uuringud</title>
                        <entry typeCode="COMP">
                    <procedure classCode="PROC" moodCode="EVN">
                        <code code="94309-2" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC" displayName="SARS koroonaviirus 2 Ag"/>
                        <statusCode code="completed"/>
                        <effectiveTime value="202004271000"/>
                        <ext:specimen ext:HL7-ClassName="SPEC" ext:HL7-Domain="COCT_RM080100" ext:realmCode="EE">
                            <ext:id extension="130315000001-2" root="ProovinouIdOID"/>
                            <ext:specimenNatural>
                                <ext:code code="Nph" codeSystem="1.3.6.1.4.1.28284.6.2.1.244.5" codeSystemName="Proovimaterjali tüüp meditsiinilaborites" displayName="Ninaneelukaabe"/>
                            </ext:specimenNatural>
                            <ext:productOf>
                                <ext:process moodCode="EVN">
                                    <ext:effectiveTime value="202004260900"/>
                                </ext:process>
                            </ext:productOf>
                        </ext:specimen>
                        <entryRelationship typeCode="REFR">
                            <observation classCode="OBS" moodCode="EVN">
                                <code code="ANA" codeSystem="1.3.6.1.4.1.28284.6.2.2.5.1" codeSystemName="Observation liik" displayName="Analüüs"/>
                                <value xsi:type="CD" code="10828004" codeSystem="1.3.6.1.4.1.28284.6.2.1.266.3" codeSystemName="Kvalitatiivse uuringu vastus" displayName="positiivne"/>
                            </observation>
                        </entryRelationship>
                    </procedure>
                </entry>
    </section>
  </component>
    </structuredBody>
</component>
```

Figure 35 XML example (Source: Author created)

LOINC codes are internationally standardized so there are 2 approaches that can be taken regarding this. One is that TEHIK can ask the new codes to be added internationally into the registry or alternatively TEHIK can do that just for the Estonian systems and update our Estonian LOINC code repository with the explanations that is located here: https://elhr.digilugu.ee/data/algandmedList.html

The reason the author has chosen LOINC code as the medium for displaying the result is because they are already supported by almost all of the portals that are used in Estonia in healthcare and because this is the main goal of the LOINC standard. LOINC is supposed to be a common language for identifying health measurements, observations, and documents. If you think of an observation as a “question” and the observation result value as an “answer”.[40]

After the document is sent to the system and saved then it will be displayed in the following format on the portal. Current example is with Coronavirus again but instead of Coronavirus TEHIK can use the display name for the new LOINC code and then write the analysis results along with any other additional information.



Logged in and representing:
Mark Selezenev

Referral response

Mark TEHIK

[redacted]

Referrer: Helle Test (D12345)

Sex: mees

Analyses

Specimen collection time	Specimen	Analysis	Analysis result	Unit	Reference value	Show additional data *
29.11.2020 04:16	Ninaneelukaabe	SARS koroonaviirus 2 RNA	positiivne			☑
		Analysis result interpretation: Kergete sümptomitega püsi kodus. Helista oma perearstile ja anna teada positiivsest COVID-19 testist. Vormista haigusleht. Если у вас легкие симптомы оставайтесь дома. Позвоните своему семейному врачу и скажите что ваш тест COVID-19 был положительным. Подайте заявку на больничный. If you have mild symptoms stay at home. Call your family doctor and tell that your COVID-19 test was positive. Apply for a sick leave. Sample container ID: 2020030000000000000				

* If the analysis result interpretation and/or comment is filled in, then in order to emphasize it, the additional data field is displayed in a different shade.

Healthcare institution: Terviseamet (70008799)

Figure 36 Referral response (Source: Author created)

6 Conclusion

When the author set out to write this work then his goal was to create a possible framework for change, prototype, and requirements for such a system. The limitations that were considered to be a problem were related to technological limitations, access to data from both technical and legal standpoints and the questions of data quality and locality.

When it comes to this work being a framework then the Healthcare Analytics Adoption Model that was found by the author seems to be able to answer the question of what can be done and why it should be done pretty well. This model was able to explain the system readiness for change and outlined the possible paths that can be taken. As for the question of how then the situation is a bit more debatable. The specific case of using pictures to predict lung cancer is a bit less straightforward compared to using specific numeric or alphabetic data for analysis. Although the ideas behind the how would remain the same if the analysed data would be numeric or alphabetic.

When it comes to the prototype then the author believes that the prototype provided is quite clear. A very important part of this prototype is that it tries to use the existing systems and procedures to reduce the work that needs to be done to finish the project. One additional part that could have been added to the prototype is the output or attempt to use the neural network to process some of the data that Estonian systems have in the Picture bank but the author gave up on this idea after finding out that this process will take quite a lot of time due to this process needing to go through the Research Ethics Committee which could go beyond the final date for the master’s thesis.

When it comes to the requirements, process analysis and architectural analysis then the situation is complicated. The author had full freedom when discussing the future process and explaining why it needs to be that way. The requirements part that came out of the interviews were much harder to put in writing because the majority of stakeholders had very few opinions on the system. The only exception was the data scientist. The architectural analysis had problems with the information that the author could discuss. Container view and beyond are protected by secrecy which cannot be simply removed by making the presentation private although those views were created.

The limitations that were presented at the start of this work were indeed problematic but the author believes that he was able to overcome them. The technological limitation was the easiest to solve. The system follows already established norms, practices, and components in the Estonian IT healthcare field. Limitations connected to data requirements for the technology was also resolvable. Since the use case that was chosen was lung cancer then the requirement that came from the methodology was images of CT scans. Estonian Picture bank follows the international guidelines of storing medical data in DICOM format which can be re-processed and extracted if needed. Since the pictures are in DICOM format this means that it is possible to sort the images based on the other data stored in the DICOM file and it is possible to remove the unwanted images. The legal limitations on the data are more complicated. While it is possible to get the data that is needed through the Research Ethics Committee, it is not a smooth process. Because this would need to be done for every increment of new data to make the re-training of the algorithm possible. Also there are additional limitations on how the system can function and permissions need to be granted before the data can be processed which reduces the overall possible benefit of the system.

In the authors opinions the stated goals of the master’s thesis have been achieved and the limitations overcome.

References

[1] Sotsiaalministeerium, November 2020. [Online material] Estonian eHealth Strategic Development Plan 2020, page 29 paragraph 3. Available: https://www.sm.ee/sites/default/files/content-editors/sisekomm/e-tervise_strateegia_2020_15_en1.pdf [Used in October 2020]

[2] Majandus- ja Kommunikatsiooniministeerium. [Online material] Eesti infoühiskonna arengukava 2020, page 1 point 5. Available: https://www.mkm.ee/sites/default/files/elfinder/article_files/eesti_infouhiskonna_arengukava.pdf [Used in October 2020]

[3] Majandus- ja Kommunikatsiooniministeerium. [Online material] Eesti infoühiskonna arengukava 2020, page 2 point 1. Available: https://www.mkm.ee/sites/default/files/elfinder/article_files/eesti_infouhiskonna_arengukava.pdf [Used in October 2020]

[4] Sotsiaalministeerium, November 2020. [Online material] Estonian eHealth Strategic Development Plan 2020, page 29 paragraph 1 and 2. Available: https://www.sm.ee/sites/default/files/content-editors/sisekomm/e-tervise_strateegia_2020_15_en1.pdf [Used in October 2020]

[5] Majandus- ja Kommunikatsiooniministeerium. [Online material] Eesti infoühiskonna arengukava 2020, page 17 point 13. Available: https://www.mkm.ee/sites/default/files/elfinder/article_files/eesti_infouhiskonna_arengukava.pdf [Used in October 2020]

[6] Majandus- ja Kommunikatsiooniministeerium. [Online material] Eesti infoühiskonna arengukava 2020, page 17 points 14-17. Available: https://www.mkm.ee/sites/default/files/elfinder/article_files/eesti_infouhiskonna_arengukava.pdf [Used in October 2020]

[7] Gregor Polancic, November 2014. [Online material] The Growing Popularity of the Business Process Model and Notation. Available: https://blog.goodelearning.com/subject-areas/bpmn/popularity-bpmn-rising/#:~:text=The%20Popularity%20of%20BPMN%20in%20Industry&text=As%20can%20be%20seen%20from,release%20of%20BPMN%20v%202.0. [Used in November 2020]

[8] Simon Brown, November 2020. [Online material] The C4 model for visualizing software architecture. Available: https://c4model.com/ [Used in November 2020]

[9] Visual Paradigm, November 2020. [Online material] What is use case Diagram? Available: https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-use-case-diagram/ [Used in November 2020]

[10] Health Catalyst, December 2020. [Online material] A Framework to Develop Analytics Maturity. Available: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/ [Used in December 2020]

[11] Health Catalyst, December 2020. [Online material] A Framework to Develop Analytics Maturity. Available: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/ [Used in December 2020]

[12] Health Catalyst, December 2020. [Online material] A Framework to Develop Analytics Maturity. Available: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/ [Used in December 2020]

[13] Health Catalyst, December 2020. [Online material] A Framework to Develop Analytics Maturity. Available: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/ [Used in December 2020]

[14] Health Catalyst, December 2020. [Online material] A Framework to Develop Analytics Maturity. Available: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/ [Used in December 2020]

[15] Health Catalyst, December 2020. [Online material] A Framework to Develop Analytics Maturity. Available: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/ [Used in December 2020]

[16] Health Catalyst, December 2020. [Online material] A Framework to Develop Analytics Maturity. Available: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/ [Used in December 2020]

[17] Health Catalyst, December 2020. [Online material] A Framework to Develop Analytics Maturity. Available: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/ [Used in December 2020]

[18] Health Catalyst, December 2020. [Online material] A Framework to Develop Analytics Maturity. Available: https://www.healthcatalyst.com/healthcare-analytics-adoption-model/ [Used in December 2020]

[19] Healthcare Catalyst, Dale Sanders PhD, February 2016. Healthcare Analytics Adoption Model. Available: https://www.youtube.com/watch?v=gTPfx6NycHk&ab_channel=HealthCatalyst [Used in December 2020]

[20] Dale Sanders PhD, David Burton MD, November 2013. [Online material] Healthcare Analytics Adoption Model: A Framework and Roadmap(White paper) Available: https://www.healthcatalyst.com/white-paper/healthcare-analytics-adoption-model/2/ [Used in December 2020]

[21] Health Catalyst. [Online material] Healthcare analytics. Available: https://www.youtube.com/watch?v=gTPfx6NycHk&ab_channel=HealthCatalyst [Used in December 2020]

[22] Sotsiaalministeerium, August 2019. Tervise ja Heaolu Infosüsteemide Keskuse põhimääruse kinnitamine. Available: https://www.tehik.ee/sites/default/files/2020-12/Tervise_ja_Heaolu_Infosuesteemide_Keskuse_pohimaeaerus.pdf [Used in December 2020]

[23] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy. [Printed literature] Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34 Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html [Used in December 2020]

[24] National Healthcare Service, Crown copyright 2019. [Online material] Diagnosis Lung Cancer. Available: https://www.nhs.uk/conditions/lung-cancer/diagnosis/ [Used in January 2021]

[25] National Healthcare Service, Crown copyright 2019. [Online material] Diagnosis Lung Cancer. Available: https://www.nhs.uk/conditions/lung-cancer/diagnosis/ [Used in January 2021]

[26] National Healthcare Service, Crown copyright 2019. [Online material] Diagnosis Lung Cancer. Available: https://www.nhs.uk/conditions/lung-cancer/diagnosis/ [Used in January 2021]

[27] National Healthcare Service, Crown copyright 2019. [Online material] Diagnosis Lung Cancer. Available: https://www.nhs.uk/conditions/lung-cancer/diagnosis/ [Used in January 2021]

[28] RocketReach.co, January 2021. [Online material] Health and Welfare Information Systems Centre(TEHIK). Available: https://rocketreach.co/health-and-welfare-information-systems-centre-tervise-ja-heaolu-infosusteemide-profile_b45eb036fc741563 [Used in January 2021]

[29] European Union, January 2021. Estonian Medical Digital Image Bank. Available: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/2019/07/29/Estonian+Medical+Digital+Image+Bank [Used in January 2021]

[30] Tanzila Saba, August 2020. [Online material] Journal of Infection and Public Health Volume 13, Issue 9, Pages 1274-1289. Recent Advancement in cancer detection using machine learning: Systematic survey of decades, comparisons, and challenges. Available:

https://www.sciencedirect.com/science/article/pii/S1876034120305633#bib0435 [Used in January 2021]

[31] Mohammed Ismail, November 2019. [Online material] Lung Cancer Prediction Using Mining Techniques. Avilable: https://www.researchgate.net/publication/341592012_Lung_Cancer_Prediction_using_Data_Mining_Techniques [Used in January 2021]

[32] Darcy Mason and pydicom contributors, February 2021. [Online material] Working with Pixel Data. Available: https://pydicom.github.io/pydicom/stable/old/working_with_pixel_data.html [Used in February 2021]

[33] Darcy Mason and pydicom contributors, February 2021. [Online material] Writing DICOM files. Available: https://pydicom.github.io/pydicom/stable/old/writing_files.html [Used in February 2021]

[34] P. Muthamil Selvi, DR.B. Ashadevi, International Journal of Advanced Science and Technology Vol. 29. No. 3s. 2020, pages 1823-1832. [Online material] Elimination of Noise in CT Images of Lung Cancer using Image Preprocessing Filtering Techniques. Availabe: http://sersc.org/journals/index.php/IJAST/article/view/6991/4163 [Used in March 2021]

[35] P. Muthamil Selvi, DR.B. Ashadevi, International Journal of Advanced Science and Technology Vol. 29. No. 3s. 2020, pages 1826-1828. [Online material] Elimination of Noise in CT Images of Lung Cancer using Image Preprocessing Filtering Techniques. Availabe: http://sersc.org/journals/index.php/IJAST/article/view/6991/4163 [Used in March 2021]

[36] P. Muthamil Selvi, DR.B. Ashadevi, International Journal of Advanced Science and Technology Vol. 29. No. 3s. 2020, page 1831. [Online material] Elimination of Noise in CT Images of Lung Cancer using Image Preprocessing Filtering Techniques. Availabe: http://sersc.org/journals/index.php/IJAST/article/view/6991/4163 [Used in March 2021]

[37] US Department of Health and Human Services, March 2021. [Online material] Definition of benign tumor. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/benign-tumor [Used in March 2021]

[38] US Department of Health and Human Services, March 2021. [Online material] Definition of malignant tumor. Available: https://www.dictionary.com/browse/malignant-tumor [Used in March 2021]

[39] Riigiteataja.ee, March 2019. [Online Material] Isikuandmete kaitse seaduse rakendamise seadus. Available: https://www.riigiteataja.ee/akt/113032019002 [Used in March 2021]

[40] Regenstrief Institute Inc, April 2021. [Online material] What is LOINC? Available: https://loinc.org/get-started/what-loinc-is/ [Used in April 2021]

Appendix 1 File locations

Some of the interviews, pictures and original tables can be located in the google drive which the author used for storing the data: https://drive.google.com/drive/folders/1CP2Z8em8h1YO4-hghLAHxlMNDtMgqJ7d?usp=sharing

Appendix 2 Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I, Mark Selezenev, grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis Improvement of Data Processing in Estonian National E-health Information System on the Basis of Lung Cancer analysis supervised by Priit Raspel:

1) to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
2) to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

20.05.2021

[1] *The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1 and 2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.*