TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Health Technologies

Monika Mariell Reinart 192536YVEM

# ANALYSIS OF DATA OVERLAP AND DATA EXCHANGE OPTIMISATION POSSIBILITIES IN ESTONIAN CANCER REGISTRY AND ESTONIAN NATIONAL HEALTH INFORMATION SYSTEM

Master´s thesis

Supervisor: Kerli Linna
Master´s degree

Tallinn 2023

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Tervisetehnoloogiate instituut

Monika Mariell Reinart 192536YVEM

# ANDMETE KATTUVUSE JA ANDMEVAHETUSE OPTIMEERIMISVÕIMALUSTE ANALÜÜS EESTI VÄHIREGISTRIS JA TERVISE INFOSÜSTEEMIS

Magistritöö

Juhendaja: Kerli Linna
magistrikraad

Tallinn 2023

# 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: Monika Mariell Reinart

4 January 2023

# Abstract

**Background:** Cancer is subject to approximately one in six deaths worldwide. Cancer prevention relies on accurate data and statistics about cancer which set base for interventional and proactive policymaking. In Estonia, cancer data is collected by health care professionals, who record data to the Estonian National Health Information System (ENHIS) and by the Cancer Registry (CR) working under the National Institute for Health Development (NIHD), who collect cancer data from health care professionals with cancer notices. This results in potential data duplication in the ENHIS´ and CR´s work processes. Additionally, it is reported that approximately 20% of all cancer cases in Estonia remain unreported to the CR on time. **The aim of the thesis** is to analyse whether a "once-only" documentation of necessary cancer datapoints is possible to cover the needs of documenting health care services´ provision and the cancer registry. **Methods:** A qualitative study was conducted by performing document analysis on documents containing cancer datapoints recorded in the ENHIS and CR. Semi-structured expert interviews were conducted with representatives of hospitals and the CR. **The results** of the document analysis revealed that out of 37 datapoints in the cancer notice, 17 datapoints were detected in the ENHIS and 20 datapoints were identified as unique to the CR, resulting in a 46% overlap of datapoints in the ENHIS and CR. The expert interviews indicated that there is interest towards exploring the opportunity of data re-use by enabling automatic data querying for the CR from the ENHIS, although experts expressed hesitation whether the free text fields in the ENHIS could be structured into specific fields to enable querying for the CR. **Conclusions:** Optimisation of data exchange and reduction of data duplication between the ENHIS and the CR could reside in the automation of CR´s data query instead of manual data collection with cancer notices.

This thesis is written in English and is 40 pages long, including 6 chapters and 4 figures.

# Annotatsioon

Andmete kattuvuse ja andmevahetuse optimeerimisvõimaluste analüüs Eesti vähiregistris ja tervise infosüsteemis

**Taust:** Maailmas sureb umbes üks kuuest inimesest vähktõve tõttu. Vähiennetus tugineb suures osas täpsetele vähiandmetele ja statistikale, mis on aluseks erinevate ennetavate sekkumiste ja regulatsioonide kujundamiseks. Eestis koguvad vähiandmeid tervishoiutöötajad, kes sisestavad andmed Tervise infosüsteemi (TIS), ja Tervise Arengu Instituudi (TAI) all töötav vähiregister (VR), mis kogub vähiandmeid tervishoiutöötajatelt vähiteatiste kaudu. Selle tulemusena leiab TIS-i ja VR-i andmestikes aset potentsiaalne andmete dubleerimine. Lisaks on teada, et kõikidest vähijuhtudest Eestis jääb VR-le õigeagselt teatamata ligikaudu 20% vähijuhtudest. **Lõputöö eesmärk** on analüüsida, kas vähiandmete ühekordne dokumenteerimine on võimalik, et katta tervishoiuteenuste osutamise dokumenteerimise ja vähiregistri vajadused. **Metoodika:** Kvalitatiivne uuring viidi läbi esmalt dokumentide analüüsi kaudu, kus võrreldi TIS-i ja VR-i teatiste andmepunkte. Haiglate ja VR-i esindajatega viidi läbi poolstruktureeritud eksperdiintervjuud. **Tulemused:** Dokumendianalüüsi tulemused näitasid, et 37 vähiteatise andmepunkti hulgast kajastuvad 17 andmepunkti TIS-is ja 20 andmepunkti on unikaalsed vähiteatisele, osutades andmete dubleerimisele 46% ulatuses. Eksperdiintervjuudest selgus, et valdkonna esindajatel on huvi andmete taaskasutamise vastu, võimaldades VR-le automaatset andmepäringut TIS-ist, kuigi väljendati kõhklust TIS-is sisalduvate vabateksti väljade struktureerimise osas. **Järeldused:** VR-i andmepäringu automatiseerimine vähiteatiste manuaalse kogumise ja töötlemise asemel võimaldaks optimeerida andmevahetust ja vähendada andmete duplikatsiooni TIS-is ja VR-is.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 40 leheküljel, 6 peatükki ja 4 joonist.

# List of abbreviations and terms

| | |
|---|---|
| CDA | Clinical Document Architecture |
| CR | Cancer Registry |
| ENHIS | Estonian National Health Information System |
| FHIR | Fast Healthcare Interoperability Resources |
| GST | General Systems Theory |
| HIS | Hospital Information System |
| HL7 | Health Level 7 |
| IG | Implementation Guide |
| ITK | East Tallinn Central Hospital *(Ida-Tallinna Keskhaigla)* |
| mCODE | Minimal Common Oncology Data Elements |
| MedMorph | Making Electronic Data More Available for Research and Public Health |
| NCI | National Cancer Institute |
| NIHD | National Institute for Health Development |
| NPCR | National Program of Cancer Registries |
| PERH | North Estonia Medical Centre *(Põhja-Eesti Regionaalhaigla)* |
| RA | Reference Architecture |
| SEER | Surveillance Epidemiology and End Results |
| TEHIK | Health and Welfare Information Systems Centre |
| TÜK | Tartu University Clinic *(Tartu Ülikooli Kliinikum)* |
| UICC | Union for International Cancer Control |
| upTIS | New Generation Health Information System (*uue põlvkonna tervise infosüsteem*) |
| USCS | United States Cancer Statistics |
| XML | Extensible Markup Language |

# Table of contents

# List of figures

# 1 Introduction

Cancer is subject to approximately one in six deaths worldwide [1], [2]. Treating cancer includes conventional methods such as surgery, radiotherapy and chemotherapy, and modern methods such as hormone treatment, stem cell therapy and immunotherapy [2]. Although effective treatment measures may reduce cancer mortality, the incidence of cancer is still widespread with about 442 new cancer cases per 100 000 people yearly, based on 2013-2017 data [3]. Reducing the incidence of cancer relies heavily on preventative measures such as avoiding tobacco usage and decreasing carcinogens from food and from the everyday living environment [4], [5], [6], [7], [8]. Prevention also relies on accurate data and statistics about cancer which set base for interventional and proactive policymaking [9], [10].

Two of the key stakeholders collecting, managing and exchanging cancer data in Estonia are the Cancer Registry (CR), operating under the National Institute for Health Development (NIHD) and the health care professionals, who record cancer data in their hospital information systems, from where it is sent to the Estonian National Health Information System (ENHIS). Recording and collecting cancer data into the ENHIS is a fundamental process of health care workers´ daily tasks. However, current Estonian legislation requires that data of the same nature must also be reported to the CR, who collects cancer data from health care professionals in the form of a cancer notice for statistical analysis and public health surveillance purposes [11], [12], [13].

Initial background research on this thematic indicated, that an overlap in the cancer datapoints collected to the ENHIS and the CR may occur, resulting in data duplication instead of enabling “once-only” data collection and secondary use of existing data, as advised in the vision document for the New Generation Health Information System (upTIS, *uue põlvkonna tervise infosüsteem*) [14], [15]. Additionally, it is reported that approximately 20% of all cancer cases in Estonia remain unreported to the CR on time, resulting in manual tracing of missing cancer cases for the CR specialists [16]. Further on, inaccurate or missing data alters the validity of cancer statistics, whereas correct

statistics are imperative for detecting trends and developing strategies to be prepared and able to reduce cancer mortality and incidence [3]. With these aspects in mind, the following problem statement, aim, objectives and research questions were formulated for this thesis.

**Problem statement**: Documentation of cancer data according to national legislation requires duplicative and separate documentation of different cancer data compositions into separate databases [11], [12], [17].

**Aim:** This research aims to analyse whether a "once-only" documentation of necessary cancer datapoints is possible to cover the needs of documenting health care services´ provision and the cancer registry.

**Objectives:**

- Identify cancer datapoints in the ENHIS and CR notices
- Identify overlapping cancer datapoints in the ENHIS and CR notices
- Determine the necessary datapoints needed for once-only data collection for the ENHIS and the CR

**Research questions:**

1. Which cancer datapoints are displayed in the ENHIS and CR notices?
2. To what extent are there overlapping cancer datapoints in the ENHIS and CR notices?
3. What are the necessary datapoints needed for once-only data collection for the ENHIS and CR?

# 2 Theoretical basis

This chapter gives an overview of the General Systems Theory and the Learning Organisation Theory as the theoretical bases for the thesis and the background information and previous research regarding cancer and cancer datasets.

## 2.1 General Systems Theory and The Learning Organisation Theory – origin and concepts

A system is defined as "*a regularly interacting or interdependent group of items forming a unified whole*" [17]. The theory of looking at the world and our surroundings as wholes instead of separate parts, dates back to early philosophers such as, for example, Aristotle [18]. He exemplified a system through a human body, explaining that the human body as a whole is more than just the aggregate of all body parts, since the function and existence of a living body cannot be compared to that of a lifeless body, although the structure and composition of both may be the same [18]. The systems theory has seen many contributions and developments by scientists and specialists from different fields, covering, for example, biology, engineering, psychology, business and health care [18], [19]. However, as there is a myriad of disciplines and specialties containing different systems, Austrian biologist Ludwig von Bertalanffy recognised the need for a universal systems theory, which he nominated as the General Systems Theory (GST) [18], [20].

The purpose of GST is to unify different sciences to determine common principles among otherwise separate systems and to provide a universal framework for specialists from diverse fields to prevent reproduction of identical principles [18], [21]. Similarly to Aristotle´s argumentation on the human body as a whole system, von Bertalanffy focused on investigating systems as a whole and the interaction between the system´s parts, instead of trying to understand the system by reducing it to isolated parts [20], [21]. Besides viewing systems as whole, another key concept of GST is the dichotomization of systems as open or closed [22]. Open systems communicate and exchange mediums with the environment (e.g., humans), whereas closed systems

remain isolated, without any material entering nor exiting the system (e.g., minority communities separated from the modern civilisation) [18], [20]. Although von Bertalanffy developed the GST emanating from his experience in biology, parallels can be drawn to systems other than living organisms as well, for example the healthcare system [22].

The GST has proven to be useful in providing insights to the healthcare system [18]. The theory enables to understand the complexities of healthcare systems and it is especially suitable when addressing information technology solutions in healthcare [23]. It is also used in public health to identify and understand issues through analysing systems as wholes but also interactions and relationships between the system´s parts [24]. Within the context of healthcare, the system consists of various parts that can be categorised into different stakeholders (e.g., patients, institutions, policy makers), different levels of care (e.g., preventative and palliative) and different professions (e.g. doctors, nurses, epidemiologists) [18]. With the healthcare system receiving energic inputs and data and producing this into services, patient volumes and transformed data, it can be said that healthcare systems are open systems [25]. This thesis focuses on the Estonian healthcare system and the integral parts of the system- institutions as key stakeholders that collect, analyse and manage cancer data for preventative public health purposes.

An adaptation from the GST is The Learning Organisation theory, which utilises the conceptual framework of systems thinking [19]. This theory has been first described by Peter Senge, who emphasizes the importance of organisations capabilities of transforming into learning organisations in order to ensure continuous improvement and goal-achievement [19]. Senge highlighted five practices an organization must maintain to be a learning organisation, which include: systems thinking, personal mastery, mental models, building a shared vision and team learning [19]. Within these disciplines, systems thinking is said to be of highest importance according to Senge, as this provides a holistic approach and a framework to detecting patterns and describing the relationships between the system´s parts [19].

Healthcare is one of the fastest developing fields today, with 30% of the data in the world generated by the healthcare industry [27]. This means that organisations in healthcare must always be ready to examine their work and learn from any mistakes in

the past [19]. According to Senge, all of the five abovementioned practices must be implemented by the organisation as a whole in order to become a learning organisation [19].

## 2.2 Cancer and cancer datasets

"Cancer" is a collective term for a combination of diseases that may develop as a result of malignant cellular changes. The cause of these cellular changes lies in the interactions between a person´s genetic and environmental factors [1]. Cancer is subject to approximately one in six deaths worldwide [1], [2]. Decreasing cancer mortality and reducing the global cancer burden is often thought to be most effective by implementing primary prevention – decreasing tobacco usage, red meat consumption, fatty foods and other potential carcinogens [4], [5], [6], [7], [8]. When complemented with screenings and effective treatment, a remarkable decrease of cancer burden could be seen [6].

Primary prevention together with conventional and modern treatment methods, such as, for example surgery, chemotherapy, stem cell and targeted therapy, is an effective way to reduce cancer incidence [2]. To move towards more efficient cancer control and treatment methods, data and statistics about cancer are collected globally as a part of public health surveillance [9], [10]. Along other aspects of consideration, the collected data provides a basis for deploying interventions and strategies for prevention and patient care [26], [27]. Some examples of collected datasets include patient demographics, primary tumor site, morphology, stage of the cancer, treatment information, and death date [9]. It has been shown that a limitation to data collection is that information about comorbidities, risk factors like smoking, socioeconomic aspects or other occupational hazards is often not collected, although this data could provide additional input for implementing efficient cancer control methods [28].

## 2.3 Estonian databases and regulations

In Estonia, secondary cancer data have been registered retrospectively since 1968 [29]. The responsible institution for gathering reliable and complete cancer data is the National Institute for Health Development (NIHD) [11], [30]. It is a government established research and development institution that collects and analyses secondary data and provides reliable national information not only for cancer data, but also for

birth and abortion, drug treatment, causes of death, tuberculosis, and cancer screening data [30]. This master´s thesis focuses on NIHD´s Cancer Registry (CR) data, as the other abovementioned registries managed by NIHD do not mediate cancer data and with the cancer screening registry being separate from the CR, therefore these are not in the scope of this thesis. In addition to the CR, patients´ cancer data is also inserted to the Estonian National Health Information System (ENHIS). The ENHIS is a central national database, through which healthcare service providers can exchange data and see health data inserted by other medical professionals. Healthcare service providers, including pathologists, have an obligation to insert data to the ENHIS [31]. The authorized processor for the ENHIS is the Health and Welfare Information Systems Centre (TEHIK), which is responsible for maintenance, management, and development of the system [32].

By sending data to the ENHIS, healthcare service providers are fulfilling the "Health Services Organisation Act" *(Tervishoiuteenuste korraldamise seadus)* [33], the "Health Information System Statute" *(Tervise infosüsteemi põhimäärus)* [32], the act "Conditions and procedure for documenting the provision of health care services" *(Tervishoiuteenuse osutamise dokumenteerimise tingimused ja kord)* [34] and Regulation No. 53 of the Minister of Social Affairs of September 17, 2008 "Data composition of the documents to be transmitted to the health information system and the conditions and procedure for their submission" *(Tervise infosüsteemi edastatavate dokumentide andmekoosseisud ning nende esitamise tingimused ja kord)* [35]. Sending data to the CR, is regulated by the "Cancer Registry statute" *(Vähiregistri põhimäärus)* [11], the "Procedure for keeping the Estonian Cancer Registry" *(Eesti Vähiregistri pidamise kord)* [12] and the "Public Health Act" *(Rahvatervise seadus)* [13].

In 2021, a vision document about the New Generation Health Information System (upTIS, *uue põlvkonna tervise infosüsteem*) was published where the initial concept and an action plan on how to further develop the ENHIS over the next decade is discussed. The aim of upTIS is to offer the use of digital solutions as smartly and abundantly as possible and, among other goals, improve clinical processes. Emanating from that, one of today´s challenges presented in the vision document is related to adhering to the "once-only" data collection principle. According to this principle, data that have already been collected in national databases do not need to be duplicated and collected again,

but instead data querying must be enabled. The reuse of data collected in other databases must also be taken into account [14], [15].

## 2.4 Cancer data collection methods in Estonian databases

In the Hospital Information System (HIS), primary documentation of cancer data is digital. Data is inserted to patient case summaries and examination results in the healthcare service provider´s local HIS. When the case summary is confirmed and closed by the healthcare service provider, data is sent from the HIS to the central ENHIS [36]. For the CR, cancer data is gathered from healthcare service providers and forensic pathologists-experts either through the data exchange layer of state information systems (X-Road); by electronically filled cancer notices, that are encrypted and sent to NIHD; or by paper-based cancer notices, sent to NIHD by registered post [30]. Additionally, NIHD exchanges data with the Estonian population register (for personal data) and register of causes of death, as well as compares the CR database with databases of North Estonia Regional Hospital (PERH, *Põhja-Eesti Regionaalhaigla*) and Tartu University Clinic (TÜK, *Tartu Ülikooli Kliinikum*) [30].

Both NIHD and ENHIS are seeming to be collecting cancer data that is overlapping with similar data fields to some extent. The purposes of data collection for both establishments remain different, with NIHD gathering data for statistical purposes and ENHIS data mainly providing patient anamnesis details and examination results. According to the national Cancer Control Action Plan 2021-2030 written and finalised in 2021, the CR receives cancer notices on paper. However, the CR had the ability to link with other databases (e.g. HIS, ENHIS, health insurance fund database, etc.) for data supplementation already then [16]. Nevertheless, in a master´s thesis on the topic of "The Analysis and Design of the Electronic Cancer Registry Notification Information System at the North Estonia Medical Centre" published in May 2022, the author describes an electronic CR notice, that the author devised. By May 2022, the electronic CR notice had been in use in the PERH local HIS for about one year [37]. Question remains, whether other Estonian hospitals, that treat cancer patients – East Tallinn Central Hospital (ITK, *Ida-Tallinna Keskhaigla*) and TÜK, also use an electronic notice system for submitting cancer data to the CR or are they using paper-based notices, which is a time-consuming and poorly traceable method.

## 2.5 Problems behind data duplication and future prospects in Estonia

The ENHIS data exchange uses Health Level 7 (HL7) v3 messages based on the Extensible Markup Language (XML) syntax, and Clinical Document Architecture (CDA) document standard [14], [38]. However, these standards tend to be inconvenient, studying the standards is difficult and they are not directly aimed at developers, which is why technical development periods are long [14]. A large part of health data is unstructured and collected as free text. Therefore, much of the health data has low reuse value [14] and separate cancer notices had to be taken into use.

Another problem relates to manual cancer data submission by healthcare providers. About 20% of all cancer cases are not submitted to the CR on time, although it is required by law [13], [16]. According to the Action Plan for Cancer Control compiled and approved in 2021, there are no structured digital cancer notices and data submitted to the CR are inaccurate, partial or instructions for filling out the notices have not been followed. Finding the relevant cancer cases and entering them to the CR will thus fall on the registry workers, adding additional manual and time-consuming obligations, which is why NIHD publishes cancer data with a two-year delay [16]. According to a 2015 master´s thesis about the "Completeness of The Estonian Cancer Registry", unreported primary cancer cases from PERH and TÜK made up 5% of the total number of primary cases in Estonia in 2010-2011; and for data completeness, synchronisation and data comparison with the Estonian Health Insurance Fund and health information systems managed by TEHIK, should be enabled [39].

Yet another complication in coordinating the exchange of health data, is related to the TNM classification standard. This standard is published by the Union for International Cancer Control (UICC) and American Joint Committee on Cancer (AJCC) and is internationally recognised for classifying the spread of cancer [40], [41]. Estonian Cancer Society is an associate member of the UICC since 2004 [42]. However, the licence for TNM classification standard has never been acquired by Estonia, although the TNM staging is used exhaustively in Estonian hospitals. According to the Action Plan for Cancer Control, the 8$^{th}$ version of TNM is currently in use in Estonia by agreement [16]. Nonetheless, Systematized Nomenclature of Medicine: Clinical Terms (SNOMED CT) is free to use for everyone in the country since Estonia is a member state of the SNOMED community [38]. SNOMED CT is a coded medical terminology,

consisting of more than 350 000 concepts, that are logically connected to each other [38]. Each concept is provided with a linguistic and semantic description [38]. In July 2022, a new licensing agreement came in force between the American College of Surgeons and SNOMED International, to enable SNOMED International to incorporate updated AJCC staging concepts to SNOMED CT [43]. Outdated AJCC content will also be eliminated from SNOMED CT [43]. This agreement will enable implementation of the current evidence-based cancer staging concepts in clinical practice [43].

Based on the examples of data deficiencies and process errors described above, it can be said that justification for using cancer notices separate from ENHIS may not be the most optimal way for data exchange. Automatic data query from the ENHIS would reduce errors and relieve the workload for both NIHD workers and healthcare service providers. The upTIS project has been initiated in Estonia to update the ENHIS, with one of the main aims being erasing such inefficiencies as described above. One of the project´s aims is to transfer data exchange from the current HL7 v3 and CDA to the HL7 Fast Healthcare Interoperability Resources (FHIR) standard [14]. FHIR provides a so-called data-based exchange, where a message consists of several FHIR resources that are linked together. In FHIR messages, it is possible to send and request resources separately, which is a significantly more flexible approach when compared to CDA [38]. This data exchange standard also narrows the maximum dataset in the information model allowing to exchange exactly as much data as necessary according to different use cases of various stakeholders, while simultaneously maintaining interoperability [14].

## 2.6 Previous research on structured cancer datasets

Previous research shows an optimal dataset of cancer data collected in the United States. Cancer registries across the U.S. are obligated to gather data about cancer [44]. Cancer incidence data collection began in 1973, when the Surveillance Epidemiology and End Results (SEER) Program of the National Cancer Institute (NCI) was initiated [44]. Along with the National Program of Cancer Registries (NPCR) formed in 1992, data collection for the U.S. population is managed by the two mentioned programs [44]. SEER and NPCR also publish the yearly United States Cancer Statistics (USCS), which provides information on the cancer burden, data for research, cancer control evaluations

and future prospects [44]. Cancer data reporting depends foremost on quality data provided by hospitals and pathology laboratories [44]. Collected datasets include diagnosis, treatment, and outcomes and are used for surveillance, healthcare planning and compiling cancer control interventions [44].

Cancer data collection is recognized to be complex due to the heterogeneity of the disease, various diagnostic and prognostic aspects and the number of medical encounters of the patient, that contribute to data production from different unharmonised data sources [44]. Challenges presented are similar to those in the Estonian digital healthcare – lack of discrete data fields, use of free-text fields, problems with data flow, delay in data publication and availability and absence of standardized cancer data collection methods [44]. The current standard in the U.S. for electronic data transmission from ambulatory care to the central cancer registry is the "HL7 Implementation Guide for CDA© Release 2: Ambulatory Healthcare Provider Reporting to Central Cancer Registries, Release 1", Cancer Registry CDA IG" [44]. This standard determines the composition and format for data reporting, however it does not use HL7 FHIR [44]. The current reporting process is claimed to be time-consuming, unreliable and resource demanding, leading to some cancers and treatment information being unreported [44]. As a solution, automated electronic reporting of cancer cases is seen [44] which could be relevant in the Estonian context for cancer cases´ reporting automation as well.

Today there are no data models or structured approaches implemented in the U.S. to collect cancer data efficiently and in a reusable manner [45]. The research of the FHIR Infrastructure Work Group describes an approach to collecting cancer data through using FHIR resources [46]. An **implementation guide** (IG) is "*a set of rules about how FHIR resources are used (or should be used) to solve a particular problem, with associated documentation to support and clarify the usage*" [46]. A **FHIR resource** is a category of health care data, which defines data components; it is an indivisible unit of information [47], [48]. The Central Cancer Registry Reporting Content IG is intended to be used for automation of data capture about cancer cases and treatment. For this, it will also use the Making Electronic Data More Available for Research and Public Health (MedMorph) Reference Architecture (RA) IG and other FHIR frameworks [44], [49]. In turn, the detailed specification for MedMorph Central Cancer Registry Reporting Content IG utilises various FHIR profiles and IGs [50]. A **FHIR profile** is defined as "a

set of constraints on a resource" [51]. One of the used IGs is Minimal Common Oncology Data Elements (mCODE) FHIR IG Usage [50].

mCODE was initiated to assemble the basis of structured data elements for cancer in electronic health records [45]. It allows capturing quality data about patients´ treatments, enabling each cancer case and treatment to advance the comparative effectiveness analysis of different treatment options [45]. In addition to treatment analysis, mCODE presents a framework of datasets to be collected about cancer patients [45]. mCODE contains about 30 FHIR profiles that are categorized into six groups: Patient Information Group, Disease Characterization Group, Health Assessment Group, Genomics Group, Cancer Treatments Group and Outcomes Group [45]. A comprehensive view of mCODE profiles and their interrelationships is added to Appendix 1.

Another research, conducted on the German healthcare system´s cancer data exchange illustrates why standardisation of data elements is important, as it enables interoperability and data sharing between different organisations [52]. The research describes a problem of similar thematic, as described in previous chapters, where stakeholders of the German healthcare system, such as university hospitals, cancer centers and biobanks face a problem of maintaining interoperable cancer data exchange without coordinated standards [52]. Each of the stakeholders have different use-cases of cancer data usage, which is why creation of a modular FHIR dataset is explored, allowing structured data exchange and storage with a universal data format [52]. For a complete description of cancer and achieving an optimal dataset for recording cancer data, the adapted data model of following data profiles was described: primary diagnosis, histology, metastasis, grading, TNMc/TNMp, surgery, radiotherapy, systemic therapy, residual status, overall assessment of tumor status, follow-up, treatment case, vital status [52].

## 2.7 Conclusion of theoretical basis

In conclusion, duplicative cancer data collection by healthcare service providers, who are obliged to submit data to the ENHIS and the CR separately, presents as an inefficiency in the Estonian healthcare system. In some cases, data submission to the CR may even take place on paper-based cancer notices, which is a time-consuming and

error prone method. Secondary use of collected cancer data is of great importance in research, epidemiological surveillance, and evaluation [52]. It is established, that control of the cancer burden could be supported by extensive implementation of verified preventative interventions [9]. As cancer control is a data-driven matter, it is only possible with precise, timely and complete data [9], [52]. Previous research indicates that a structured dataset for automatic cancer reporting improves data completeness and this is demonstrated through exploring the HL7 FHIR mCODE initiative [45], as Estonia´s healthcare system is also moving towards implementing the FHIR specification [14], [38]. Research from Germany affirmed that organisations participating in cancer data exchange face difficulties without having a common data standard [52]. Hence, analysis of primary data collection and its usage for NIHD´s secondary data collection is performed in the thesis on the basis of the General Systems Theory and the Learning Organisation Theory. These theories enable to research the current internal work processes of healthcare professionals and CR specialists as subsystems of the healthcare system and analyse the stance of the organisations toward adopting more optimal data exchange methods. The theories also provide a basis for the analysis of the interdependence and relationships between the subsystems as they form a unified whole of the healthcare system´s cancer data management entity. To research this theme further and provide potential opportunities for improvement, next chapters will present a methodology overview of document analysis and expert interviews, results of this research, discussion and proposals for further research. The topic out of scope for this thesis is data quality of cancer data in the CR and ENHIS.

# 3 Methodology

This thesis is written on the basis of the advocacy and participatory worldview, applied within a qualitative research method, as described by J. W. Creswell, B. Atweh, S. Kemmis and P. Weeks [53], [54]. Key features of this worldview include targeting a policy reform or a plan for transformation, with the aim of improving the lives of the research participants or target groups [53]. Hence, during document analysis for this thesis, regulatory policies were processed along with inputs from interviewees, describing the current inefficiencies in the cancer data collection system and targeting potential methods for improvement. Within the advocacy and participatory worldview, attention is also aimed at including participants as collaborative parties, as opposed to marginalised individuals [53], which is why consideration was given to implement four principles while compiling the scheme for the interview. Firstly, clear communication with the participants was established by interacting in a manner that conveys trust and transparency [54]. Secondly, an accurate and goal-oriented intent was conveyed to participants, by explaining the background of the research to the interviewees, but not expecting them to share a solution to the problem right away [54]. Thirdly, jargon was avoided in the interview scheme, which was shared with the participants, by writing out any abbreviations to avoid misunderstandings [54]. Lastly, appreciation of the participants´ time constraints was expressed, by designing the interview scheme in an optimal manner to retrieve as much of quality information as possible in an efficient timeframe [54].

Triangulation strategy was used for quality improvement of the qualitative research [55]. In triangulation, combination of methodologies is used to study the same topic [56], for example, two qualitative research methods [55]. By collecting information from different sources with different methods, the author can confirm the findings and reduce any biases [56]. Data collection was conducted by a review of cancer data documentation and selection of appropriate documentation for further document analysis; followed by individual unstandardised semi-structured expert interviews.

Previous knowledge from document analysis was collected prior to performing the interviews. A combination of inductive and deductive content analysis was chosen for analysing the interviews´ transcripts. The purpose of the inductive approach enables avoidance of any biases and allows new potential perspectives to emerge from interviews and the deductive approach enables to remain within limits of the specific research topic by following the semi-structured interview scheme [57].

## 3.1 Document analysis

For answering the first research question "Which cancer datafields are displayed in the ENHIS and CR notices?" and the second research question of the thesis "To what extent are there overlapping cancer datapoints in the ENHIS and CR notices?" a document analysis approach was chosen. Document analysis is used for electronic or printed materials consisting of text and images, to interpret data into meaning and gain knowledge as explained by Glenn A. Bowen [56]. This method is often used in healthcare researches due to a large number of healthcare documentation being produced constantly [58]. The benefit of document analysis manifests in developing original information and knowledge in fields that cannot be researched by observational or experimental methods [58]. It is also a stable method as in the researcher´s examination of documents does not alter the target under study [56]. As a disadvantage to document analysis accessibility to documents can be brought out or the process of retrieving the documents can be difficult [56]. Additionally, the documents are initially produced for aims other than having them researched which may result in the researcher not being able to retrieve sufficient details from the documents [56].

As the first step of document analysis, Estonian regulations and laws were reviewed to identify legislation covering health data processing, management and exchange in the ENHIS and the CR. From the included legislation, information regarding cancer data management was retrieved. Thirdly, data collectors, data exchangers and data fields of cancer datasets were extracted from legislation. In the next steps, standards and guidelines for healthcare professionals collecting patient and cancer data were retrieved from public sources. From the standards and guidelines, all specific datapoints related to cancer patients and the process of cancer case management in the ENHIS were extracted and listed in a Microsoft Excel table. Also, from public sources cancer notice

datapoints were retrieved and juxtaposed to the Excel table along the datapoints from standards and guidelines.

The following documents for datasets´ analysis were retrieved from public sources:

- Health information system data exchange standard: Outpatient case summary. TEHIK *(Tervise infosüsteemi andmevahetuse standard: Ambulatoorne epikriis. TEHIK)* [59]

- Instructions for filling out an outpatient case summary *(Ambulatoorse epikriisi täitmise juhend)* [60]

- Health information system data exchange standard: Case summary. Patient portal. *(Tervise infosüsteemi andmevahetuse standards: Epikriis. Digilugu)* [61]

- Instructions for filling out an inpatient discharge summary *(Statsionaarse epikriisi täitmise juhend)* [62]

- Health information system data exchange standard: Referral reply letter. TEHIK *(Tervise infosüsteemi standard: Vastus saatekirjale TEHIK) I* [63]

- Instructions for filling out a reply to the reference letter *(Saatekirja vastuse täitmise juhend)* [64]

- Instructions for completing the form "Notice to the Cancer Registry" *(“Teatis vähiregistrile” täitmise juhend)* [65]

- Instructions for completing the form "Notification of the Department of Pathology to the Cancer Registry" *(“Patoloogiaosakonna teatis vähiregistrile” vormistamise juhend)* [66]

Based on the comparison conducted between these datasets, a list of overlapping and differing datapoints was compiled. The retrieved documents all contained large amounts of datapoints in a written and numerically ordered manner. Identification of relevant data fields was done to isolate and extract all information related to cancer patients´, including general data of medical documents, analysis results, patient data, referral data, diagnosis data, treatment data etc. All datapoints were then organized to provide an overview of recurring and relevant datapoints.

To follow the “once-only” data collection principle, a dataset consisting of all necessary datapoints was identified through three steps:

1. Identifying overlapping datapoints in the ENHIS and CR
2. Identifying the datapoints needed by NIHD for the CR, that are not represented in the ENHIS
3. Identifying the necessary data used and needed by both stakeholders during expert interviews and obtaining additional input for the most optimal data exchange principles

As a final step, all content from the results of the document analysis and results of interviews´ content analysis was compared and investigated.

## 3.2 Individual unstandardised semi-structured expert interviews

With hospitals´ and NIHD representatives, individual unstandardised semi-structured expert interviews were performed. Unstandardised semi-structured interviews allow participants to answer questions in their own words [67]. This contributes to potential development of new lines of thought, that may be left unmentioned in case of standardised and structured interviews. Answering the interviewer´s questions individually provides more time and a calmer pace for the interviewee to elaborate, whereas in focus groups the pace may be faster and there is less time for reasoning and answering [67]. By conducting individual interviews with different experts, potential disputes can also be avoided that may otherwise arise in focus group interviews and which may occupy time unnecessarily. Expert interviews enable to gather factual knowledge about the field of interest, while ordinary in-depth interviews with any participants are aimed at gathering personal opinions or attitudes towards the topic [67], [68]. Considering the detailed structure of cancer datasets, the usual practices of hospitals´ and NIHD representatives and their knowledge about the field, individual unstandardised semi-structured expert interview was chosen as the most appropriate method.

The target participants for the interviews were gathered by purposive sampling, whereas using the criterion sampling strategy. This strategy allows to determine and select a

limited number of target participants that are most likely to provide relevant and practical information on a deeper and more profound level [69]. The interviews were conducted online through a Microsoft Teams web-call. The limitation of online interviews can be the possibility of poor internet connection, technical errors or poor audio/visual quality of the call, which may cause loss of information or loss of time for fixing the problems. Nonetheless, advantages for conducting online interviews include the interactive communication method, where spontaneous topics may add additional value and insight to the topic under research. Additionally, unexpected themes may also help overcome researcher-centred bias, in which addressed conversation subjects are expected to be of importance to and already known by the author [70]. The questions for the interviews were developed alongside document analysis. Each stakeholder group was asked different questions emanating from their specific pathway of managing cancer patients and/or the patients´ cancer data. The author contacted participants who would represent each stakeholder holding and participating in exchange of cancer data or regulating this exchange and who would meet the following criteria: be knowledgeable in their work field, demonstrate experience in their work field, be in a leading role in their work field, be able to express and communicate their opinions and knowledge in an articulate manner [71].

Expert interviews were conducted with:

- oncologists from each of the three hospitals in Estonia, where cancer patients are treated – PERH, ITK and TÜK
- specialist working with the CR, and
- an analytic from Pärnu Hospital.

Participants were contacted via emails and phone calls. Interview schemes were sent to the interviewees beforehand via email, to enable preliminary preparation and fact checking. All interviews were conducted virtually during an online web-call in the Microsoft Teams environment, to save time at the expense of transportation, since some participants were located in other Estonian cities than Tallinn. Additionally, given the local rising trend of COVID-19 infections, online web-calls were preferred. Interviews took place in October-November 2022.

Interviews´ transcriptions were analysed during a combination of inductive and deductive content analysis. Content analysis is defined as "*a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use*" by the method author K. Krippendorff [72]. "*Other meaningful matter*" is defined by Krippendorff as interviews, art, images, sounds, maps or even symbols, that can be used as data for content analysis [72]. Inductive content analysis is used for qualitative research by creating code categories for similar patterns based on the raw data [73], [57]. Any pre-set codes are avoided to prevent any biases and to allow new data and insights to emerge spontaneously [57]. This can also be seen as an advantage of the inductive content analysis method – preconceived angles are prevented and participants can provide direct information [57]. Deductive content analysis entails pre-set codes that are deduced from previous research or an initial theory [57]. In the current thesis, the codes are verified against the content of the interviews´ transcriptions by firstly applying deductive coding by developing codes based on previous document analysis. During deductive analysis, additional codes emerged, leading to an inductive coding of transcriptions in parallel, in order to develop codes based on repetitive patterns.

Audio recordings of interviews were reproduced into transcripts non-verbatim. Content analysis aims to identify patterns and similar ideas in interviews´ transcripts [74], therefore verbatim transcription was not seen necessary as only general ideas were extracted. Transcriptions were produced using a speech recognition software. Produced transcripts were then proof-read along the recordings and sent back to interviewees for fact-checking and quality control. As a limitation of the speech recognition software, it was identified that manual transcription of audio recordings could be more accurate in cases where the audio is unclear during the recording. Whereas an advantage to speech recognition is the speed of transcription, when compared to manual transcription. For the next step, content of the transcripts was coded with labels to enable identification and categorization of patterns and general ideas in the conducted interviews. Finally, the codes are grouped to three topics: "Current work process with cancer data management" under which the description of current cancer data management is provided by the participants; "Problems in the current work process" where participants highlight problems in their daily work processes with cancer data management; and "Ideas for improvement" where participants were able to describe their ideas for a more efficient

management of cancer data and express any topics of interest regarding cancer data management optimisation (see Figure 1). The coding was conducted in the NVivo programme [76].

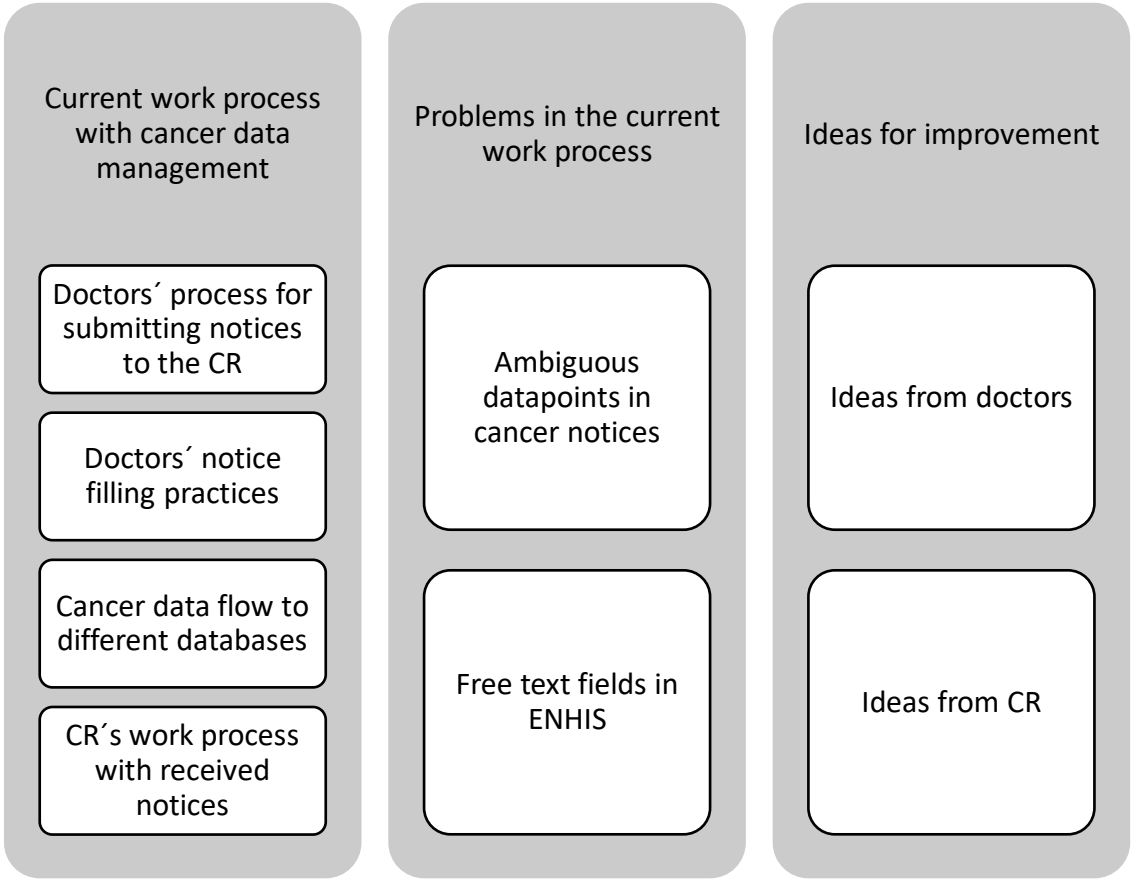



Figure 1. Interview data coding categorisations

## 3.3 Ethical considerations

Permission for recording the interview was requested at the beginning of each interview. The author initiated the interview by explaining the aim of the thesis and principles for ensuring confidentiality by pseudonymization, in case of including any citations to the thesis. Interviewees were then asked for permission to record the audio and cite any quotes from the interview. All permissions to record and cite were orally obtained and recorded on the recording of each interview. After conducting the interviews, transcriptions were produced using a web-based speech recognition and transcription service. Recording files were named according to pseudonyms assigned to the participants and then uploaded to the transcription environment, where the author

registered an account for personal use only. All interviewees received a transcript of their interview for any fact checking to assure the quality and truthfulness of the data, and for final approval to allow usage of the transcriptions for analysing the content and writing the thesis.

Citations were generalised and pseudonymised to a level which did not allow identification of any details related to the participant. Pseudonyms were assigned to participants based on the order of interviews (P1, P2, … etc.). Interview recordings and transcriptions were stored in the password-protected TalTech OneDrive cloud server, allowing access only to the author of the study and the supervisor. Audio recordings were destroyed after transcription of the interviews. Transcripts will be stored for 3 months after the defence of the thesis in January 2023.

# 4 Results

The following section provides the findings of document analysis and individual unstandardised semi-structured expert interviews. Document analysis was performed on documents containing information regarding cancer datasets in the ENHIS and CR. Content analysis of the interviews was conducted, where the transcripts´ text was grouped into three topics: “Current work process with cancer data management” under which the description of current cancer data management is provided by the participants; “Problems in the current work process” where participants highlight problems in their work process with cancer data management; and “Ideas for improvement”.

## 4.1 Document analysis

Document analysis was conducted with the aim of answering the first research question “Which cancer datapoints are displayed in the ENHIS and CR notices?” and the second research question of the thesis “To what extent is there overlapping data in the ENHIS and CR notices?”.

Datapoints from the standards and guidelines were juxtaposed in an Excel table for identification of overlapping and differing datapoints in the two databases. From the 37 datapoints required in cancer notices, 18 datapoints were detected in the ENHIS and 19 datapoints were identified as unique to the CR (marked in grey colour in Figure 2). Therefore, approximately 49% of the datapoints needed by the CR are recorded in the ENHIS. It is however possible, that since some data in the ENHIS may be recorded in a different form and under a differently titled section, than those in the notice, then the share of the notice´s datapoints available in the ENHIS may be greater than 49%.

| ENHIS | CR |
|---|---|
| Patient data | |
| Identification code | |
| First and last name | |
| Place of residence | |
| Diagnosis | |
| Metastasis data | |
| Previously diagnosed malignant tumors | |
| Tumor site of previously diagnosed malignant tumors | |
| | Date of previously diagnosed malignant tumors |
| Detailed site of current tumor | |
| | Laterality |
| | Date of diagnosis of current tumor |
| Research methods that confirmed the diagnosis | |
| Morphological diagnosis | |
| Degree of malignancy (G1-G4) | |
| TNM and stadium during diagnosis (TNM+c/p) | |
| Treatment data | |
| Surgical treatment | |
| Operation name as free text | |
| Date of operation | |
| | Radiation therapy |
| | Radiation dose |
| | Date of treatment initiation |
| | Chemotherapy |
| | Healthcare institution |
| | Name of drug or course of treatment |
| | Date of treatment initiation |
| | Hormone therapy |
| | Healthcare institution |
| | Drug name |
| | Date of treatment initiation |
| | Other treatment (e.g., biological, stem cell treatm |
| | Did not receive special treatment + justification |
| Time of death | |
| Cause of death (free text or RHK-10 code) | |
| Health care service provider (health care institution) and data related to data submission | |
| Medical record or medical card number | |

Figure 2. Overlapping datafields in the ENHIS and CR

## 4.2 Individual unstandardised semi-structured expert interviews

The transcriptions of the interviews were coded and distributed into three categories, where participants describe the current work process with cancer data management,

problems in the current work process and finally propose ideas for improvement. For a better understanding of the interdependence of hospitals and NIHD, the current work process was in turn categorised into doctors´ work process, doctors´ notice filling practices, cancer data flow to different databases and CR´s work process. The participants highlighted problems with ambiguous datapoints required in the notices and free text fields in the ENHIS, while also providing ideas for improvement from both hospitals´ perspectives and the CR´s perspective.

### 4.2.1 Doctors´ process for submitting notices to the CR

The interviews revealed that the general concept and requirements for submitting the cancer notices to the NIHD are similarly understood by all representatives of the hospitals.

> "*The thing about this notice is that it should be completed by the oncologist who made the initial cancer diagnosis. Is it a surgeon, a chemotherapy doctor, a radiotherapy doctor… generally, a surgeon does it.*" (P2)

The healthcare professional who records the initial diagnosis of the patient is generally obliged to submit the notice to the NIHD as well. Additionally, every "*change in the treatment schema or treatment plan"* (P2) must be recorded on a new cancer notice and submitted. Another participant specified that in addition to initial diagnosis and treatment, a notice must also be submitted "*in case of the cancer patient´s death"* (P3).

Although in theory the notice should be submitted by the oncologist who made the initial diagnosis and, in most interviews, it was evident that this is done in practice as well, it may be that the notice is sometimes submitted by another healthcare professional, for example, a nurse:

> "*Sometimes the nurse also fills the notice. It depends on how the work is organised in hospitals."* (P2)

In some cases, the patient already has a diagnosis but has transferred from one healthcare institution to another because of treatment availability purposes or for a second opinion. In this case, doctors usually do not submit a notice with a diagnosis, however, they cannot be sure, whether the previous institution already submitted a notice with the diagnosis or not:

> "*For example, I do not see the notices made in other hospitals. Similarly, I do not see the register of the National Institute of Health Development and whether a notice about the patient has already been made."* (P3)

Although it is possible for a healthcare professional to make a query to NIHD about a specific patient and whether a notice has been submitted, the written query must be submitted manually and the response from NIHD may require time, due to it being processed manually as well. Another participant also mentioned that perhaps the reason behind some unreported notices relies in the "*movement of the patient from one institution to another. And for some reason, neither institution submits the notice, I don´t know whether they think that the other institution already reported or…"* (P4).

All participants confirmed that it is very likely, that one patient has more than one notice related to them:

> "*The same patient could undergo surgery, radiation, and chemotherapy. And based on that, he may have had five notices actually, because each doctor did something in the treatment there*." (P1)

Besides different purposes for submitting several notices about one patient, a realistic scenario also includes human error, where "*if I accidentally sent a flawed notice, then I will make another one again"* (P1).

### 4.2.2 Doctors´ notice filling practices

The interviews revealed that in PERH, ITK, TÜK and Pärnu hospitals, cancer data is collected digitally and notices are also filled and forwarded to the NIHD digitally over the X-road.

> *"The cancer notice is one A4 sheet, at least it used to be, now it is a digital notice. We have definitely had a digital notice option for about over a year. But we also enter the same data in the digital notice, which used to be written on paper notices - the digital notice has exactly the same basis."* (P1)

With the digital notices, a pre-filling option is available, enabling the automatic transfer of some datapoints from the structured fields epicrises in the hospital information system to the digital cancer notices:

> *"We have had digital notices for about two years /.../ The hospital information system pre-fills the cancer notice with elementary personal data – name, gender, place of residence, and diagnosis. In case of breast cancer, laterality is also included."* (P2)

Although the pre-filled information only contains basic personal data and diagnosis of the patients in most cases, capability for such a pre-filling functionality is developed, providing reduction of manual data entry to some extent.

Regarding the general obligation of submitting notices, doctors admit, that they are aware of the problem with notices being left unreported, be it due to forgetfulness or uncertainty regarding whether notices about a patient have already been submitted or not. One participant highlighted, that as the NIHD "*collects diagnoses, not treatment methods"* (P5) then the treatment section on the notice is left unfilled by the participant. A different participant elaborated on discussions that the participant has had with colleagues, about "*whether it even makes sense to enter treatment data there, because it is so chaotic and not all doctors enter it, and the cancer registry doesn't really evaluate it later either"* (P2). Similarly, another participant illustrated that although the participant submits the notices as required, it remains unclear what is done with this data further on:

> "*It is not clear what is being done with this data, because I can find out from the cancer registry how many people with this initial diagnosis there have been this year. Nothing more can be learned from there. Maybe I can get the age composition as well, how old, how many, which cancers were there and that's it. But the rest of the information entered in the cancer notice - what was the surgery, what type of chemotherapy, what was the histological diagnosis, I can't find it anywhere anymore. Or what stage it was, for example*." (P1)

As detailed statistics of aggregated cancer data cannot be retrieved from the CR nor the ENHIS, the same participant explained having to collect patients´ cancer data on paper, to be able to submit this data yearly for statistical reporting to a specialty society.

### 4.2.3 Cancer data flow to different databases

All of the participants from hospitals confirmed, that cancer data is currently forwarded to two main systems – the ENHIS and the CR. One participant added that the cancer data collected by the participant must also be submitted to a specialty society regarding "*how much and what exactly has been done, with what method, what has been cut and how much chemotherapy has been administered, what stages have there been"* (P1). As the participant expresses that "*I don't think it is possible to get this information from the hospital's information system*" (P1) and the "*specialty society probably cannot get this information from the cancer registry either*" (P1) then some cancer data must be recorded on paper in order to provide accurate statistics for specialty societies, besides ENHIS and CR. Another participant included that due to the participant´s specialty, there is an obligation to submit cancer data to a site-specific registry but additionally because detailed information cannot be retrieved for statistics neither from CR nor ENHIS, site-specific registries fulfil this gap:

> "*The site-specific databases have a bit more detailed information. There should be aggregated data somewhere, that's why such a site-specific database exists, where it is written exactly which drugs were used, when they were started, when they were stopped, plus survival data, when the disease progresses. Then we can also evaluate the effectiveness and quality of treatment. Based on the current collected data, including the epicrisis, we can't really do that, or we have to do it manually - take out all the patients one by one. But if they are entered in the site-specific database, it is possible to make queries from there."* (P2).

Therefore, in order to be able to make queries, receive aggregate statistical data and also get an overview of treatment efficacies, some doctors must also submit cancer data to additional databases, besides ENHIS and CR. Nevertheless, this practice is only done within specific specialities and not for all tumor sites.

### 4.2.4 CR´s work process with received notices

The NIHD´s CR is composed of two databases – a database for notices and a database for cancer cases, as explained by NIHD´s specialist. Notices from both, oncologists and pathologists are collected to the notices´ database but the main activity performed by the

registry is "*compilation of the cancer case"* (P4) which takes place in the cancer cases´ database:

> "*The database of cancer cases is person-based, all notices received at different times are linked to that person. And then we have this compilation of the cancer case, which actually means that we review all the information and create a separate cancer case record based on the information we have from notices. The basis for the cancer case registration is the cancer notice, which must be there, because otherwise we will not register the cancer case. We will not register the cancer case with only a pathologist´s notice. If we only have a pathologist´s notice, then we send a request to the healthcare institution and ask for the doctor´s notice as well. The doctor´s notice contains the main data, based on which the cancer case is registered*." (P4)

For ensuring the completeness of data, CR has two-year contracts with PERH and TÜK, under which CR can send patient records with missing RHK-10 diagnosis codes to PERH and ask for the missing notices; and with TÜK, incomplete records are also sent but since CR has access to TÜK´s database, missing notices are filled in by the CR itself. Another participant noted that their hospital gets a query from NIHD every two years, where "*approximately 15 patients are included in the query about whom notices have not been submitted to NIHD*" (P3).

### 4.2.5 Ambiguous datapoints in cancer notices

In addition to it sometimes remaining unclear to some doctors, why certain datapoints (e.g., treatment) is collected by CR, another problem resides in the ambiguity of some datapoints required in the notices:

> "*When we collect the dates of diagnosis, which is the most important thing… actually the date of diagnosis comes from that pathologist's notice. Well, the date of the result is there. According to international rules, I guess the diagnosis is still the date of taking the sample. And not even the time of the histology result, but precisely when the sample is taken. And theoretically, I think that this is exactly what we should record there. I don't know if it has been agreed upon at all… actually it is a good question if we have really systematically agreed on what we are going to record as the diagnosis date."* (P5)

The date of the diagnosis presents as one of the confusing datapoints in the notice, with which doctors often hesitate as to what is expected to be written:

> "*What I still don't understand is that there is a cancer diagnosis time- a date, when this cancer is diagnosed. If we read the guide for compiling the cancer notice, it states that the time of cancer diagnosis is when the patient turns to the doctor, which isn´t really a correct statement. The time of diagnosing the cancer is when you take its histology, and the histology result comes back. Then you diagnose cancer when you get that answer. Because before getting the result, I don't know what it is. Some people record the diagnosis date as the date when the histology is taken. I mark the date when the answer is specifically answered, and I have received the diagnosis from the histology. Another thing is that sometimes the results arrive after two months, some genetic research was done or something. Then it is not appropriate to mark the diagnosis date like that.*" (P1)

As misunderstanding regarding the diagnosis date arose already among the participants of this research, it can be presumed that there is wider misapprehension regarding the correct date to be marked on the notice.

Although mentioned more rarely than the problem with correct diagnosis dates, TNM classification was also brought out as a vague area, that could be understood differently, when reading the guideline for filling the notice:

> "*TNM is also an interesting thing in a sense that, well, how different people understand it. You can read this guide from one perspective or another. And if, for example, a head and neck surgeon or an otolaryngologist records it, then all those TNMs are correct, which are put there under the primary diagnosis. But, for example, if a urologist records it, I often re-do the TNM. Because I feel like it should be different.*" (P5)

In other cases, the problem may lie more in the practice of different doctors, with some doctors even "*not describing the TNM classification at all"* (P2).

A concern related to the specific oncological speciality of a participant was related to recording the spread of the tumor, that could not be marked on the notice in a sufficient manner:

> "*In the cancer notice, it is necessary to select the spread of the tumor. For example, in the case of ovarian cancer, there is no option to choose the spread that is characteristic of this tumor. They are usually not localized, but they have not metastasized either. Instead, they have spread along the peritoneum, but it cannot be written there. Some cells are missing that we would like to have. It's just specifically for this tumor, there's no problem with other tumors, but I don't know, maybe other doctors feel the same way about their tumor, that something is missing, and you can't write what you want. Then you have to mark "undefined", but in fact it is quite defined, but in the case of this tumor, the right cell is simply not given, which needs to be marked."* (P1)

### 4.2.6 Free text fields in ENHIS

Interviewees from the hospitals and NIHD explained, that inefficient and manual work processes around cancer notices are caused by unstructured free text fields in patient epicrises.

> *"[In the cancer notice] there is a cell for histological diagnosis, /.../. This is a free text field. But for free text, you can't make statistics about how many tumors there were."* (P1)

Several participants expressed how they feel that they are not getting as much data back from the CR, as they are inserting to the notices, but at the same time pointed out that data collected in free text fields on the notices cannot be efficiently processed in order to produce any sort of statistics.

> *"Each healthcare service provider has procured specific developments [to their hospital information systems] and they send the data from the health information system, but the data is not in the designated fields where it should be."* (P4)

Data written on free text fields in the hospital information system does not allow automatic integration of this data, either to pre-fill cancer notices or for NIHD´s data processing purposes.

### 4.2.7 Ideas from doctors

A continuous line of thought throughout the interviews was that within the current practice of submitting cancer notices, errors are prone to happen, when a human component is mediating data submission and that automatic data querying could reduce errors:

> "*If data is submitted automatically, let's say that a diagnosis is made somehow or a tick has to be marked somewhere, then the more automatic the process is, the fewer errors and the more normal the data actually is."* (P5)

A fundamental thought provided by the same participant explained, that in order to move from collecting unstructured and often seemingly unnecessary data with notices, to collecting justified and complete data, we must "*answer the question, why and what do we collect? The cancer registry should articulate very well why and what they collect and understand the value of this data and that if this data is not perfect, it is not possible to produce any statistics from it. And if these goals are in place, we can collect this data"* (P5).

Emanating from the fact that doctors are only able to see notices that have been compiled in the same hospital, a participant elaborated on the possibility of having the cancer notice move along with the patient, in case the patient transfers from one healthcare institution to another:

> "*Of course, it would be good if the patient went, for example, from TÜK to ITK, that there would be such a pre-filled notice, where, for example, surgical treatment would already be marked down, if the patient has been operated on. In a sense, they could see it in real time. And then they put their data, which they document in their hospital - whether, for example, the stage has changed, whether the treatment has changed, for example, they operate on metastases in the spine or perform some kind of systemic treatment there, which was not done at the TÜK. In some ways, of course, it would be good if it were like that.*" (P2)

At the end of the interviews, representatives of the four hospitals were asked, whether they would prefer to continue submitting cancer notices or rather prefer a scenario where it would be possible for the CR to query necessary datapoints automatically from the ENHIS. The participants preferred the automatic query possibility, although they expressed hesitation whether the free text fields in the ENHIS today could be structured into specific fields to enable querying for the CR.

### 4.2.8 Ideas from CR

In the light of the fact that many cancer cases yearly go unreported to the CR and flawed notices may be received, the representative of the CR was also asked whether they would prefer to continue working with manually compiled notices or would like to see an option for automatic data query from ENHIS. Initially the response reflected a rather negative stance toward automatic querying, as the participant expressed the understanding that ENHIS contains documents in PDF format, and it would be complicated to retrieve necessary information in a structured manner. However, further elaboration conceded that in case the necessary datapoints were made available in a structured manner, then theoretically it could be possible to query cancer data directly from the ENHIS. Even though, it was expressed that the preferred form of receiving data would still be in the format of a notice, whether it be manual or automatic data retrieval.

As seen from the interviews, all four hospitals are filling and submitting cancer notices digitally. Although the Action Plan for Cancer Control states, that creating a structured digital cancer notice improves the completeness and accuracy of data submission, which in turn reduces the number of return requests made by the registry staff, helps speed up the routine work of the registry and the publication of cancer statistics [16], this cannot be confirmed yet, as the CR is working with a two-year delay, meaning that they have not yet started processing the digital notices, which have only been in use for about 1-2 years.

# 5 Discussion

This section of the thesis discusses the results, the current state, and practices of cancer data management in the ENHIS and CR and provides a look into data exchange optimisation possibilities and answers to the research questions. Limitations and further research opportunities are presented at the end of the chapter.

## 5.1 Key stakeholders collecting and managing cancer data

The thesis focused on key stakeholders collecting, managing, and analysing cancer data in Estonia, how current data exchange between them takes place, and potential data exchange optimisation possibilities. The network of key stakeholders comprises of the central ENHIS and the CR operating under the NIHD. Additionally, as became eminent from the interviews- a relevant stakeholder group are specialty societies and site-specific databases, to where cancer data is also submitted.

Decreasing cancer incidence relies on preventative measures, among other cancer-countering methods [5], [6], [7], [8]. To be able to see the root causes for cancer incidence but also seek for effective treatment methods, accurate data and aggregate statistics are necessary [9], [26]. With a common goal of decreasing cancer incidence, all parts of the healthcare system are required to cooperate. This means, that while healthcare service providers are actively diagnosing and treating cancer patients, they must also record cancer and treatment data for secondary use and provide this information to the CR for statistical analysis. Although the CR relies heavily on healthcare service providers to submit accurate data, CR is also responsible for producing correct overviews of cancer statistics. These statistics are the basis for policy-makers but also for moving towards a general health-conscious lifestyle, be it at the population level or at individual level. It is eminent that with a field requiring collaboration at such a demanding level, all parts of the healthcare system are expected to cooperate with mutual understanding and for a collective goal of reducing cancer incidence. By grouping together all the inefficiencies and errors in the doctors´ and CR´s work processes, it is possible to look at the problems as failures of the whole system and identify potential patterns, as advocated by the General Systems Theory [19]. This provides input for developing the healthcare system into a more efficient

network consisting of capable organisations as sub-systems. However, with the hospital representatives being confused of, for example, the reason behind CR collecting treatment information while not publishing statistics about cancer treatments later on; or the CR having to search for 20% of unreported cancer cases yearly [16], it is clear, that the stakeholders may not have a common understanding of how the mutual goal should be achieved.

The doctors´ perception of cancer data management entails justified reasoning behind each datapoint required to be collected. The current dataset collected with cancer notices includes information that currently is not justified to an extent where doctors would assent to collecting it. Such datapoints are included in the treatment section of the notice. It was also expressed in expert interviews that the level of structuration for the treatment data on the notice is not reflected in the ENHIS, where information regarding the cancer patient´s treatment is mostly recorded in free text fields. Therefore, if treatment data collection is justified by the CR, as it is for specialty and site-specific registries, then it would be preferred to see published statistics regarding this treatment from the CR by the healthcare professionals´ community. If there are no published data on treatment and the oncologists do not feel like they are getting back the data that they submit to the CR, there is no motivation to fill treatment information in the notice. According to the Learning Organisation theory, the system with its sub-systems must implement the systems thinking, along with other important disciplines such as building a shared vision and team learning [19]. If the hospitals´ and CR´s representatives are able to learn from the experience of cancer notices´ usage, gather and analyse the inefficiencies of the current cancer data management system and build a vision of an ideal scenario of cancer data management, then meaningful action can be taken to improve the system as a whole [19].

The CR´s vision of cancer data collection would be to continue gathering data in the form of notices. However, with one of the biggest problems for the CR today being unreported cancer cases, there is willingness to consider automation for data querying from the ENHIS. In this case, structured fields for surgical treatment, radiation therapy, hormonal or other treatment should be considered as necessities in the development of hospital information systems and ENHIS. Structuration of data and implementing a common data format for data exchange is supported by previous research in the field of cancer data management [45], [52].

## 5.2 How and what cancer data is collected in Estonia?

The participants of the expert interviews explained that the most detailed data composition on a cancer patient is inserted to the hospital information system, from where it is forwarded to the ENHIS. Specialty and site-specific registries also contain detailed cancer and cancer treatment data submitted by the health care professionals, whereas it is possible to retrieve aggregated data queries from these databases as well. The CR collects cancer and treatment data on a more general level but retrieving the processed and aggregated data back from the CR will often produce no results, as the representatives of the four hospitals explained.

The CR´s cancer notice contains required datapoints categorised into five different sections. The first section with general patient data, such as personal identification code, first and last name and place of residence is also available in the ENHIS. Section two of the cancer notice requires information about the diagnosis, diagnosis method and details about the spread of the tumor. The section also contains information regarding previously diagnosed malignant tumors, the site of the previous tumors and date of diagnosis, with the diagnosis date of the previous tumors being the only datapoint which cannot be retrieved from the ENHIS dataset. Regarding the current tumor, the site, research methods that confirmed the diagnosis, morphological diagnosis, degree of malignancy, TNM and stadium during diagnosis are datapoints that are presented in the ENHIS as well, however the laterality and diagnosis date are not directly available in ENHIS as structured data fields. Cancer treatment data on the cancer notice is expected to be filled out for surgical treatment, radiation therapy and chemotherapy. For surgical treatment, operation name and date are included in the ENHIS dataset. Regarding radiation therapy dose and date, chemotherapy drug name and date, hormone therapy drug name and date or other treatment methods, no structured fields are available in the ENHIS. From the expert interviews it became evident that this data is inserted to the specialty and site-specific cancer databases, from where it can be queried, and aggregate statistics can be compiled for treatment. The notice´s fourth section contains information regarding time of death and cause of death of the patient, with the death date being available in the ENHIS, but cause of death not being structured among ENHIS datafields. The last section on the cancer notice is related to data about the healthcare service provider who is managing the patient, which is also available in the

ENHIS. Additional information about cancer data can be found in ENHIS, which is not required on the notice.

With Estonia moving towards implementing the HL7 FHIR data exchange standard and thereby re-mapping the datasets of various services that should be exchanged through the ENHIS, this presents as a good opportunity to organize the cancer dataset so that it can be used for both primary and secondary data use. With previous research in the example of U.S., mCODE presents an optimal dataset for collecting cancer data while also allowing interoperability between different organisations [45] that is currently lacking in the Estonian healthcare setting.

## 5.3 Optimisation possibilities

According to the expert interviews, there are inefficiencies in the current work process of cancer data collection and management. One of the main concerns of the hospitals´ representatives was that the composition of cancer dataset required by the CR lacks justification, as information regarding treatment is submitted to the CR but is not published in the statistics later on. The CR also revealed that information submitted to them is often incomplete, especially regarding treatment data. With the treatment data being available in site-specific and specialty registries, an opportunity to be explored could include the CR linking their database with the site-specific and specialty databases in order to have a control measure for their data and complement it with data from specialty registries. It is also stated in the vision document for upTIS, that the data already collected in national databases do not need to be duplicated but instead data querying must be enabled for relevant stakeholders [14], [15]. The vision document highlights, that reuse of data in different databases must be taken into account [14], [15].

Participants of the expert interviews were all asked whether they would prefer to continue submitting cancer notices or rather prefer a scenario where it would be possible for the CR to query necessary datapoints automatically from the ENHIS. The representatives of hospitals preferred the option for automatic data query from ENHIS, as it would reduce the amount of manual work for them and expressed hope for a more complete data composition in the CR through this option. The hospitals´ representatives also elaborated, that most of the datapoints required on the cancer notice, could be

recorded in structured datafields in the ENHIS, if such developments for structured fields would be made to the hospital information systems. With 49% of the datapoints overlapping in the notice and ENHIS, there are 19 unique datapoints in the notice that would require structuration in the ENHIS, to cover the needs for CR´s potential data query from ENHIS or find out whether any of the 19 datapoints are documented under some other characteristic in the ENHIS. Likewise, for the reason of free text fields and unstructured data in the ENHIS, the CR´s representative initially stated, that they would prefer to continue receiving notices in the same way. However, during further discussions it became evident that if all the required datapoints would be structured and if the automatic query could be compiled in the shape of the current notice, then the CR does not rule out the option of transferring their data collection methods to automatic querying from the ENHIS.

Automatic data query from the ENHIS would potentially enable the CR to retrieve a more complete data composition to their database. Since doctors are filling out the epicrises and anamnesis of their patients on a daily basis and as also explained by experts during the interviews, the data composition in ENHIS is detailed and comprehensive. However, as the cancer notice does not serve a purpose for the doctors directly and they do not see the result of this collected data, then it is a secondary duty for them to manually compile and submit it, resulting in the CR receiving incomplete notices or in many cases not receiving notices at all. An automated query would eliminate the human factor to an extent, where retrieving data would not depend on the doctors remembering to compile and send the notices to the CR. Transferring to an automatic data query would also require the re-work of current national legislation, as this currently demands for duplicative and separate documentation of cancer datapoints.

An advantage to the CR collecting data through notices is adaptability. There is an opportunity to change their required data composition more flexibly if CR is regulating the data assembly by themselves. In case the required data would be assembled through the ENHIS, a standardisation process would have to be followed, reducing the flexibility of the CR´s data composition. Therefore, if the treatment section of the notice is partially filled or not filled at all by a large number of health care professionals, the CR has the capability today to, for example, remove or rearrange this section of the notice at any time. Although the quality of data remains out of scope for this thesis, the experts additionally explained that the quality of data could be higher, in case data is

submitted directly to the CR, as the data quality recorded in the ENHIS varies by different doctors. This is also the reason behind speciality and site-specific registries collecting information separately and not through querying data from the ENHIS. Nonetheless, as highlighted in the interview, the CR has many manual and time-consuming obligations today, to ensure the quality of data in their databases, which results in the CR publishing their collected data with a two-year delay.

## 5.4 Limitations

A limitation of this thesis is that two key stakeholders´ cancer data management process was analysed, while there are more relevant stakeholders behind the development of the current cancer data management system. Such additional stakeholders are, for example, policy-makers in the Ministry of Social affairs and the Estonian Health Insurance Fund who processes the treatment bills of the cancer patients. Likewise, other healthcare professionals like, for example, nurses, physiotherapists and psychologists who may also come in contact with cancer patients may provide further perspectives as to what data they collect and how they manage and exchange it. As it was not possible to get the representative of the Ministry of Social Affairs to agree in participating for the interview, this stakeholder was excluded from the selection. Additionally, as the focus was intended to be kept solely on the data compositions of the CR´s cancer notice and the ENHIS, other stakeholders were excluded as well. The participants for the expert interviews were gathered by purposive sampling, being prone to researcher bias, as the author of the thesis is making general assumptions when choosing participants for the interviews. Although there may be variations in working methods in the different departments of the hospitals and selection bias cannot be averted, general work practices of the hospitals are represented by the participants.

## 5.5 Future research

For seeking justification to the datapoints required for the CR, future research on the organisational and policy-makers´ contributions on the current situation could provide useful information as to why the CR collects exactly these datapoints today and what was the reasoning behind this. Further on, a side-theme could be gathering information on the reasons behind why doctors tend to not submit cancer notices today. An

additional research topic would be analysing the current problems in the data exchange between the cancer screening registry and the ENHIS, as the screening registry already works by querying data from the ENHIS. However, there are known to be problems with this data collection method for the screening registry, which could also be taken into account for the CR. Data quality of cancer data when submitting notices manually to the CR could be compared to a scenario where data is queried from the ENHIS, where data quality can vary among doctors as well, as this could provide additional reasoning when making decisions regarding the automation of CR´s work process. After the implementation of upTIS, a research theme could be to identify whether duplication of cancer data collection is erased, "once-only" data collection principle has been rooted and whether databases consisting of cancer data enable secondary use of this data.

## 5.6 Final conclusions

This research contributes to opening the topic of inefficiencies rooted in the current national legislation, where duplicative and separate documentation of cancer data is required [11], [12], [17]. With previous international research in the field highlighting the importance of structured cancer data, no similar study has been detected in the Estonian context. This research expanded the current inefficiencies and problematic patterns in the process of cancer data collection and management by healthcare professionals and the CR.

Based on the results of the thesis, following answers can be provided to the initially established research questions:

1. The ENHIS contains cancer datapoints that are unstructured and recorded as free text to a large extent. The CR´s notice contains sections of datafields for general patient data, diagnosis, diagnosis method and tumor spread details, treatment, time of death and information regarding the healthcare service provider

2. The overlapping datapoints in ENHIS and CR notices extend to 49% of the datapoints.

3. The necessary datapoints for the CR are included in the sections of general patient data, diagnosis, diagnosis method and tumor spread details, time of death and information regarding the healthcare service provider. As speculated by the

experts, the currently collected cancer treatment data would be better excluded from the notices.

Optimisation of data exchange and reduction of data duplication between the ENHIS and the CR could reside in the automation of CR´s data query instead of manual data collection with cancer notices. Although there are factors that argue for data collection with cancer notices, such as flexibility in changing the notice´s data composition, compelling arguments against the use of notices emerged as well. Transferring the CR´s data collection to automatic querying from the ENHIS requires re-working the current legislation, which currently requires duplicative and separate data collection. An automatic data query option could result in a more complete cancer dataset for the CR and reduce manual workload of the registry workers, who today are tracking missing information and unreported cancer cases through time-consuming processes.

# 6 Summary

The aim of the thesis was to analyse whether the needs of the secondary data use for NIHD can be covered with primary data collection by healthcare professionals. The author of the thesis conducted a document analysis and individual unstandardised semi-structured expert interviews with key stakeholders managing, collecting and exchanging cancer data, being hospital representatives as ENHIS stakeholders and the CR representative as the NIHD stakeholder.

The results of the document analysis revealed that out of 37 datapoints required in the cancer notices, 18 datapoints were detected in the ENHIS meaning that approximately 49% of the datapoints are overlapping in the cancer notice and the ENHIS.

Semi-structured expert interviews revealed the current work processes for the hospitals´ representatives and registry workers and also the preferences of potential data exchange options. Throughout four of the hospitals involved in the interviews, cancer notices are submitted digitally to the CR. Among hospitals´ representatives there is a tendency to prefer a potential automatic data query option, where the doctors insert cancer data to their hospital information system from where it is shared to the ENHIS, and the CR would be able query this data from the ENHIS automatically over the X-road. The CR´s representative expressed, that if this scenario entails structured cancer data, that could be queried from the ENHIS, then this is a solution to be considered.

In conclusion, the rate of unreported cancer cases and missing data in the cancer notices submitted to the CR could see improvements through automation of data querying. Current advantages to the CR collecting cancer data with manually submitted notices manifests in flexibility to alter the required data composition according to need, without having to comply to standardisations required in the ENHIS data composition. Nevertheless, if there is a fundamental agreement between stakeholders and justification for the dataset required to be collected, the need to change the composition of the cancer dataset from time to time could remain minimal, and a standardised dataset of cancer data could be collected and re-used by all stakeholders.

# References


[1] "Cancer," World Health Organization, 3 February 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed 15 October 2022].

[2] D. T. Debela, S. G. Muzazu and K. D. Heraro, "New approaches and procedures for cancer treatment: Current perspectives," 2021.

[3] National Cancer Institute, "Cancer Statistics," 25 September 2020. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/statistics. [Accessed 25 October 2022].

[4] B. Alberts, A. Johnson and J. Lewis, "Cancer Treatment: Present and Future," in *Molecular Biology of the Cell. 4th edition.*, New York: Garland Science, 2002.

[5] I. Valle, D. Tramalloni and N. L. Bragazzi, "Cancer prevention: state of the art and future prospects," 2015.

[6] P. Vineis and C. P. Wild, "Global cancer patterns: causes and prevention," *The Lancet,* vol. 383, no. 9916, pp. 549-557, 2014.

[7] B. K. Dunn and S. K. Barnett, "Cancer Prevention: Lessons Learned and Future Directions," 2016.

[8] H. A. Loomans-Kropp and A. Umar, "Cancer prevention and screening: the next step in the era of precision medicine," 2019.

[9] M. C. White, F. Babcock, N. S. Hayes and A. B. Mariotto, "The history and use of cancer registry data by public health cancer control programs in the United States," *Cancer,* vol. 123, no. S24, pp. 4969-4976, 2017.

[10] Tervise Arengu Instituut, "Vähitõrje," [Online]. Available: https://www.tai.ee/et/tervis-ja-heaolu/vahitorje. [Accessed 15 October 2022].

[11] Minister of Health and Labour, "Vähiregistri põhimäärus," 2019. [Online]. Available: https://www.riigiteataja.ee/akt/112032019022. [Accessed 15 October 2022].

[12] Minister of Social Affairs, "Eesti Vähiregistri pidamise kord," [Online]. Available: https://www.riigiteataja.ee/akt/13124858?leiaKehtiv. [Accessed 15 October 2022].

[13] Riigikogu, "Rahvatervise seadus," [Online]. Available: https://www.riigiteataja.ee/akt/104012021013?leiaKehtiv. [Accessed 20 October 2022].

[14] upTIS juhtrühm, "Uue põlvkonna tervise infosüsteem. Visioon tervise infosüsteemile," 2021.

[15] "upTIS tehnilised põhimõtted järgimiseks," Tervise ja Heaolu Infosüsteemide Keskus, 2021. [Online]. Available: https://www.tehik.ee/uptis-tehnilised-pohimotted-jargimiseks. [Accessed 16 October 2022].

[16] Sotsiaalministeerium, "Vähitõrje tegevuskava 2021-2030," Tervise Arengu

] Instituut, Tallinn, 2021.

[17] Merriam-Webster, 2022. [Online]. Available: https://www.merriam-webster.com/dictionary/system. [Accessed 24 November 2022].

[18] C. P. Cordon, "System Theories: An Overview of Various System Theories and Its Application in Healthcare," *American Journal of Systems Science,* vol. 2, no. 1, pp. 13-22, 2013.

[19] B. R. Anderson, "Improving health care by embracing Systems Theory," *The Journal of Thoracic and Cardiovascular Surgery,* vol. 152, no. 2, pp. 593-594, 2016.

[20] B. D. Friedman and K. N. Allen, "Systems Theory," SAGE Publications Inc, 2014.

[21] L. v. Bertalanffy, General System Theory, New York: George Braziller, Inc., 1968.

[22] F. E. Kast and J. E. Rosenzweig, "General Systems Theory: Applications for Organization and Management," *The Journal of Nursing Administration,* vol. 11, no. 7, pp. 32-41, 1981.

[23] D. K. Fahey, E. R. Carson, D. G. Cramp and J. A. Muir Gray, "Applying Systems Modelling to," *Systems Research & Behavioral Science,* vol. 21, no. 6, pp. 635-649, 2004.

[24] W. M. Trochim, D. A. Cabrera, B. Milstein, R. S. Gallagher and S. J. Leischow, "Practical Challenges of Systems Thinking and Modeling in Public Health," *American Journal of Public Health,* vol. 96, pp. 538-546, 2006.

[25] R. M. Meyer and L. L. O´Brien-Pallas, "Nursing Services Delivery Theory: an open system approach," *Journal of Advanced Nursing,* vol. 66, no. 12, pp. 2828-2838, 2010.

[26] U. E. Bauer, P. A. Briss, R. A. Goodman and B. A. Bowman, "Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA," *The Lancet,* vol. 384, no. 9937, pp. 45-52, 2014.

[27] A. K. Goel, W. S. Campbell and R. Moldwin, "Structured Data Capture for Oncology," *JCO Clinical Cancer Informatics,* vol. 5, pp. 194-201, 2021.

[28] V. Arndt, B. Holleczek, H. Kajüter, S. Luttmann, A. Nennecke, S. R. Zeissig, K. Kraywinkel and A. Katalinic, "Data from Population-based Cancer Registration for Secondary Data Analysis: Methodological Challenges and Perspectives," *Das Gesundheitswesen,* 2020.

[29] K. Innos, "Vähitõrje tegevuskava elluviimine vajab teadlaste tuge," 2022.

[30] "Research and development: Estonian Cancer Registry," National Institute for Health Development, 2020. [Online]. Available: https://en.tai.ee/en/r-and-d/registers/estonian-cancer-registry. [Accessed 15 October 2022].

[31] "Health information system," [Online]. Available: https://www.tehik.ee/en/health-information-system. [Accessed 15 October 2022].

[32] Vabariigi Valitsus, "Tervise infosüsteemi põhimäärus," [Online]. Available: https://www.riigiteataja.ee/akt/103082022004?leiaKehtiv. [Accessed 15 October 2022].

[33] Riigikogu, "Tervishoiuteenuste korraldamise seadus (lühend - TTKS)," [Online].

] Available: https://www.riigiteataja.ee/akt/110102022003?leiaKehtiv. [Accessed 15 October 2022].

[34] Minister of Social Affairs, “Tervishoiuteenuse osutamise dokumenteerimise tingimused ja kord,” [Online]. Available: https://www.riigiteataja.ee/akt/124032022006?leiaKehtiv. [Accessed October 15 2022].

[35] Minister of Social Affairs, “Tervise infosüsteemi edastatavate dokumentide andmekoosseisud ning nende esitamise tingimused ja kord,” [Online]. Available: https://www.riigiteataja.ee/akt/13029628?leiaKehtiv. [Accessed 15 October 2022].

[36] R. Ojavee, *Tervise infosüsteemiga liidestumine haigla infosüsteemi LIISA näitel,* Tartu: Tartu Ülikool. Matemaatika-informaatikateaduskond, 2013.

[37] A. Kikas, “Elektroonilise vähiregistri teatiste infosüsteemi analüüs ja kavandamine Põhja-Eesti Regionaalhaiglas,” Tallinna Tehnikaülikool, Tallinn, 2022.

[38] “Teabekeskus,” Tervise ja Heaolu Infosüsteemide Keskus, [Online]. Available: https://www.tehik.ee/teabekeskus. [Accessed 20 October 2022].

[39] J. Põder, “EESTI VÄHIREGISTRI ANDMETE TÄIELIKKUSE UURING,” 2015. [Online]. Available: http://dspace.ut.ee/bitstream/handle/10062/85691/P%c3%b5der2015.pdf?sequence=1&isAllowed=y.

[40] “What is the TNM cancer staging system?,” UICC.org, [Online]. Available: https://www.uicc.org/resources/tnm. [Accessed 20 October 2022].

[41] “AJCC staging system,” National Cancer Institute, [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/ajcc-staging-system. [Accessed 23 October 2022].

[42] “Member Organisation,” UICC.org, [Online]. Available: https://www.uicc.org/membership/estonian-cancer-society. [Accessed 20 October 2022].

[43] “Cancer staging content from the American Joint Committee on Cancer will appear in SNOMED CT under new licensing agreement,” SNOMED International, 6 July 2022. [Online]. Available: https://www.snomed.org/news-and-events/articles/SNOMED-ACS-AJCC-licensing-agreement. [Accessed 23 Otober 2022].

[44] Public Health Work Group, “Central Cancer Registry Reporting Content IG,” HL7 International, [Online]. Available: https://build.fhir.org/ig/HL7/fhir-central-cancer-registry-reporting-ig/usecases.html. [Accessed 16 October 2022].

[45] “minimal Common Oncology Data Elements (mCODE) Implementation Guide,” HL7 FHIR, [Online]. Available: http://hl7.org/fhir/us/mcode/STU2/. [Accessed 17 October 2022].

[46] “Resource ImplementationGuide - Content,” HL7 FHIR, [Online]. Available: https://www.hl7.org/fhir/implementationguide.html. [Accessed 16 October 2022].

[47] The Office of the National Coordinator for Health Information Technology, “Introduction to FHIR® Resources,” [Online]. Available: https://www.healthit.gov/sites/default/files/page/2021-04/Intro%20to%20FHIR%20Resources%20Fact%20Sheet.pdf. [Accessed 16 October 2022].

[48] E. Yesakov, “WHAT ARE FHIR PROFILES?,” Kodjin Interoperability Suite, 2022. [Online]. Available: https://kodjin.com/blog/what-are-fhir-profiles/.

[Accessed 16 October 2022].

[49] "MedMorph Home Page," HL7 FHIR, [Online]. Available: http://build.fhir.org/ig/HL7/fhir-medmorph/. [Accessed 16 October 2022].

[50] "Central Cancer Registry Reporting Content IG: Detailed Specification," HL7 FHIR, [Online]. Available: https://build.fhir.org/ig/HL7/fhir-central-cancer-registry-reporting-ig/spec.html. [Accessed 16 October 2022].

[51] "Profiling FHIR," HL7 FHIR, [Online]. Available: http://hl7.org/fhir/R4/profiling.html. [Accessed 16 October 2022].

[52] I. Zachary, S. A. Boren and E. SImoes, "Information Management in Cancer Registries: Evaluating the Needs for Cancer Data Collection and Cancer Research," *Online J Public Health Inform,* 2015.

[53] J. W. Creswell, Research design: Qualitative, quantitative, and mixed methods approaches, SAGE Publications, Inc, 2009.

[54] B. Atweh, S. Kemmis and P. Weeks, Action research in practice: partnerships for social justice in education, Routledge, 2002.

[55] U. Flick, An introduction to qualitative research. Fourth edition, SAGE Publications Ltd, 2009.

[56] G. A. Bowen, "Document Analysis as a Qualitative Research Method," *Qualitative Research Journal,* vol. 9, pp. 27-40, 2009.

[57] F. Moretti, L. Vliet and J. Bensing, "A standardized approach to qualitative content analysis of focus group discussions from different countries," *Patient Education and Counseling,* vol. 82, no. 3, pp. 420-428, 2011.

[58] T. Moilanen, M. Sivonen, K. Hipp, H. Kallio, O. Papinaho, M. Stolt, R. Turjamaa, A. Häggman-Laitila and M. Kangasniemi, "Developing a Feasible and Credible Method for Analyzing Healthcare Documents as Written Data," *Global Qualitative Nursing Research,* 2022.

[59] G. Kriiska, *Tervise infosüsteemi standard: Ambulatoorne epikriis. Digilugu,* Tervise ja Heaolu Infosüsteemide Keskus, 2020.

[60] "Ambulatoorse epikriisi täitmise juhend," Tervise ja Heaolu Infosüsteemide Keskus, 2021. [Online]. Available: https://pub.e-tervis.ee/manuals/Ambulatoorse%20epikriisi%20t%C3%A4itmise%20juhend. [Accessed 15 October 2022].

[61] Tervise ja Heaolu Infosüsteemide Keskus, "Tervise infosüsteemi standard: Epikriis. Digilugu," 2020.

[62] "Statsionaarse epikriisi täitmise juhend," Tervise ja Heaolu Infosüsteemide Keskus, 2021. [Online]. Available: https://pub.e-tervis.ee/manuals/Statsionaarse%20epikriisi%20t%C3%A4itmise%20juhend. [Accessed 15 October 2022].

[63] *Tervise infosüsteemi standard: Vastus saatekirjale. TEHIK,* Tervise ja Heaolu Infosüsteemide Keskus, 2021.

[64] Tervise ja Heaolu Infosüsteemide Keskus, 2022. [Online]. Available: https://pub.e-tervis.ee/manuals/Saatekirja%20vastuse%20t%C3%A4itmise%20juhend. [Accessed 15 October 2022].

[65] ""Teatis vähiregistrile" täitmise juhend," Tervise Arengu Instituut, [Online]. Available: https://www.tai.ee/sites/default/files/2022-05/Teatis%20v%C3%A4hiregistrile%20vormistamise%20juhend.pdf. [Accessed 15

October 2022].

[66] ""Patoloogiaosakonna teatis vähiregistrile" vormistamise juhend," Tervise Arengu Instituut, [Online]. Available: https://www.tai.ee/sites/default/files/2021-03/Patoloogiaosakonna_teatise_vormistamise_juhend.pdf. [Accessed 15 October 2022].

[67] K. Lepik, H. Harro-Loit, K. Kello, M. Linno, M. Selg and J. Strömpl, "Sotsiaalse analüüsi meetodite ja metodoloogia õpibaas," 2014. [Online]. Available: https://samm.ut.ee/intervjuu. [Accessed 22 October 2022].

[68] B. DiCicco-Bloom and B. F. Crabtree, "The qualitative research interview," *Wiley Online Library,* 2006.

[69] S. Campbell, "Purposive sampling: complex or simple? Research case examples," 2020.

[70] L. Busetto, W. Wick and C. Gumbinger, "How to use and assess qualitative research methods," 2020.

[71] L. A. Palinkas, S. M. Horwitz and C. A. Green, "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research," 2015.

[72] K. Krippendorff, Content Analysis: An Introduction to Its Methodology. Fourth Edition, SAGE Publications, 2019.

[73] A. J. Kleinheksel, N. Rockish-Winston, H. Tawfik and T. R. Wyatt, "Demystifying Content Analysis," 2020.

[74] E. J. Halcomb and P. M. Davidson, "Is verbatim transcription of interview data always necessary?," *Applied Nursing Research,* vol. 19, no. 1, pp. 38-42, 2006.

[75] A. Olev and T. Alumäe, "Estonian Speech Recognition and Transcription," *Baltic J. Modern Computing,* vol. 10, p. 409–421, 2022.

[76] Tervise Arengu Instituut, "Vähiregister," [Online]. Available: https://www.tai.ee/et/statistika-ja-registrid/vahiregister. [Accessed 25 November 2022].

# Appendix 1 – Scope of mCODE and relationships between mCODE profiles

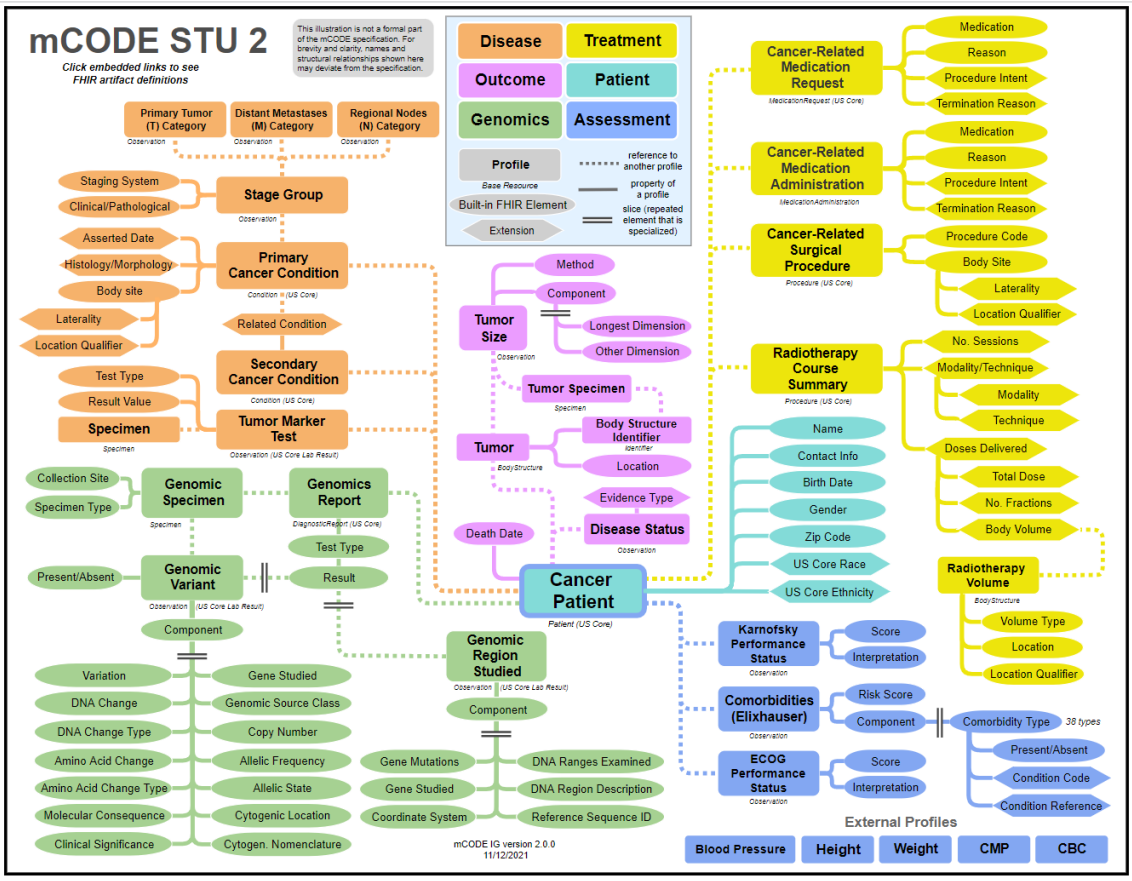



Figure 3. Minimal Common Oncology Data Elements [45]

# Appendix 2 – Questionnaire to hospitals´ representatives

| **Question** | **Purpose** |
|---|---|
| 1. Please briefly describe your current role in dealing with cancer patients; what activities do you generally encounter in your daily work? | Background and introduction |
| 2. Do you only collect cancer patients´ data (personal data, diagnosis, treatment, examination results, etc.) digitally in the hospital´s information system, or have there been situations, where in-hospital paper forms have been used for data collection?<br>a. does the information on the paper forms end up in the health information system later? | Destinations where data is recorded |
| 3. Where do you send cancer data today? | Databases where cancer data is managed |
| 4. Is the transfer of information to the cancer registry done on paper or digitally? | Digital notices vs paper notices<br>Work process for compiling notices |
| 5. Who compiles and sends the cancer notice to the cancer registry?<br>a. is there a potential risk or known occurrence of duplicate cancer notices being sent to the cancer registry for one patient when | Work process for compiling notices |

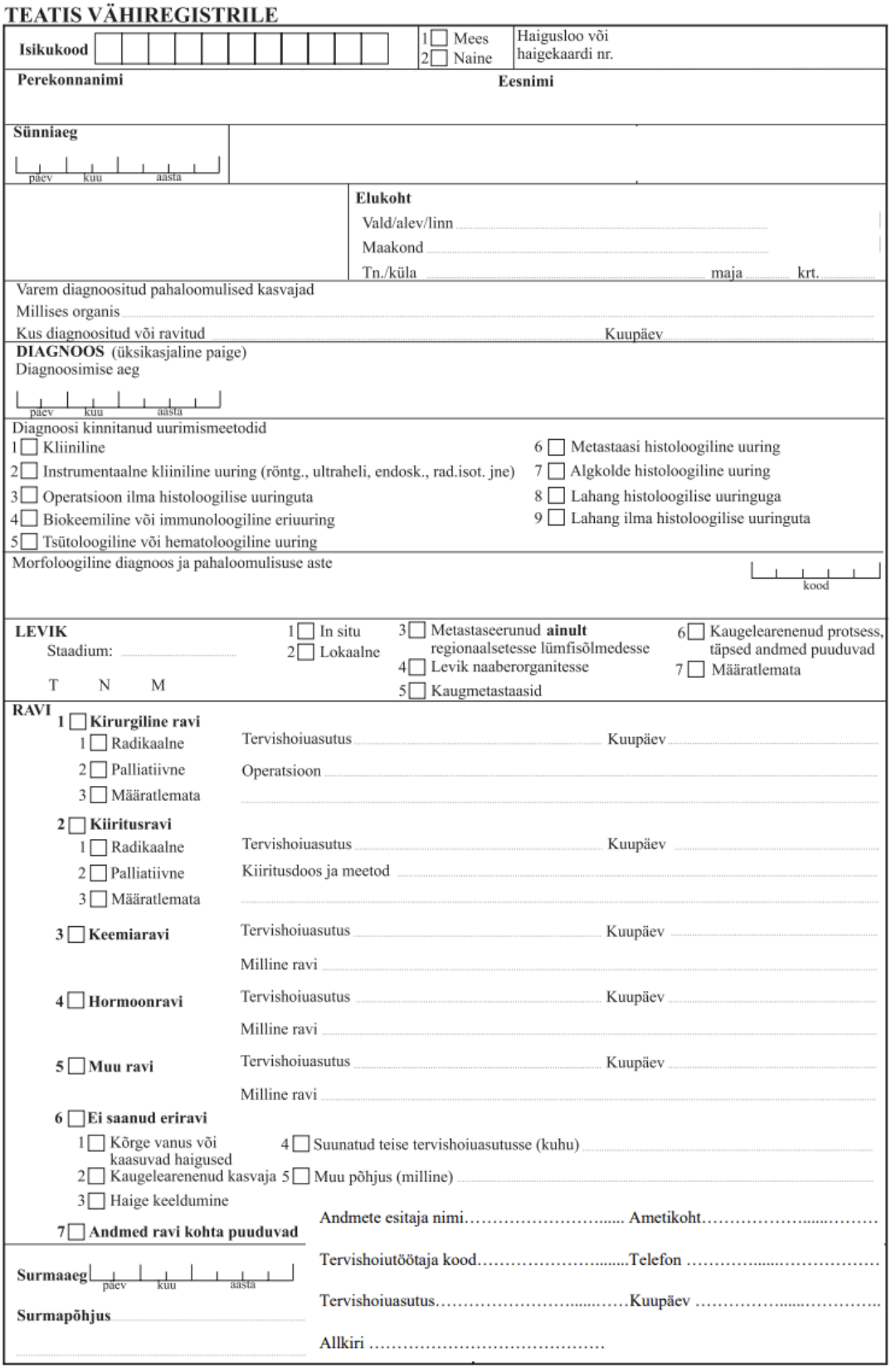
| | |
|---|---|
| multiple doctors are treating the patient?<br><br>b. is there any way for the doctor to determine that a cancer notice has already been sent for a specific patient and tumor? | |
| 6. How much time does it approximately take to compile and send the cancer notice to the cancer registry? | Work process for compiling notices |
| 7. What software do you use as your hospital information system, and do you feel that this software solution supports the documentation of cancer data to a sufficient extent; what is missing? | Work process for compiling notices |
| 8. Which data collected in the cancer registry is missing from ENHIS documents today? | Necessary datapoints |
| 9. Is the dataset you receive from other doctors through ENHIS sufficient for you to organize treatment? If not, what is missing? | Necessary datapoints |
| 10. What is your feedback on the composition of cancer-related data, can everything important be clearly documented so that it can be used by others? | Ideas for improvement<br><br>Feedback on current work process |
| 11. In an ideal scenario, how would the collection and secondary use of cancer care data work in your eyes? | Ideas for improvement<br><br>Feedback on current work process |

# Appendix 3 – Questionnaire to the Cancer Registry

| Question | Purpose |
|---|---|
| 1. Please briefly describe how the journey of a cancer notice goes today, starting with the preparation of the notice in the health care institution and ending with the information on the notice reaching the cancer registry.<br>a. who should compile the notice? | Background<br><br>Work process for compiling notices |
| 2. Do you receive notices in paper form today as well, or are they all digital? | Work process for managing notices |
| 3. In addition to the data fields included in the forms "Notice to the Cancer Registry" and "Notice from the Pathology Department to the Cancer Registry", do you also map any additional data points related to cancer patients? | Necessary datapoints |
| 4. Where and how else does the cancer registry request information?<br>a. do you also link to the databases/information systems of the big 3 hospitals? For what purpose? | Necessary datapoints |
| 5. Does the cancer notice information reach the registry automatically or does the information have to be entered manually?<br>a. entering a paper vs a digital notice<br>b. if the information is entered into the databases manually, approximately how long does it take? (both for paper and for the digitally received notice) | Work process for managing notices |

| | |
|---|---|
| 6. How many duplicate cancer notices do you get for the same tumor in the same patient? But also, how many faulty or incomplete notices? How is the sorting of OK/faulty notices organized?<br>a. for one patient, do all the doctors dealing with him fill out one notice, or does everyone fill out their own notice? | Work process for managing notices<br><br>Work process for compiling notices |
| 7. Do all the data fields filled in the notice reach the statistics? | Work process for managing notices |
| 8. At what scale are you working together with the Cancer Screening Registry today?<br>a. to what extent do the datasets of the two registers overlap? | Necessary datapoints |
| 9. What should be done differently to transfer the data collection of the Cancer Registry to the health information system?<br>a. which data reflected in the Cancer Registry are missing from the documents of the health information system today? | Ideas for improvement<br><br>Feedback on current work process |
| 10. What are currently the biggest problems in the work and data exchange of the Cancer Registry in general, and what are the causes of these problems? | Ideas for improvement<br><br>Feedback on current work process |

# Appendix 4 – Cancer notice



TEATIS VÄHIREGISTRILE

Isikukood | 1 ☐ Mees 2 ☐ Naine | Haigusloo või haigekaardi nr.

Perekonnanimi | Eesnimi

Sünniaeg (päev kuu aasta)

Elukoht
Vald/alev/linn
Maakond
Tn./küla ... maja ... krt.

Varem diagnoositud pahaloomulised kasvajad
Millises organis
Kus diagnoositud või ravitud ... Kuupäev

DIAGNOOS (üksikasjaline paige)
Diagnoosimise aeg (päev kuu aasta)

Diagnoosi kinnitanud uurimismeetodid
1 ☐ Kliiniline
2 ☐ Instrumentaalne kliiniline uuring (röntg., ultraheli, endosk., rad.isot. jne)
3 ☐ Operatsioon ilma histoloogilise uuringuta
4 ☐ Biokeemiline või immunoloogiline eriuuring
5 ☐ Tsütoloogiline või hematoloogiline uuring
6 ☐ Metastaasi histoloogiline uuring
7 ☐ Algkolde histoloogiline uuring
8 ☐ Lahang histoloogilise uuringuga
9 ☐ Lahang ilma histoloogilise uuringuta

Morfoloogiline diagnoos ja pahaloomulisuse aste (kood)

LEVIK
Staadium: ...
T N M
1 ☐ In situ
2 ☐ Lokaalne
3 ☐ Metastaseerunud **ainult** regionaalsetesse lümfisõlmedesse
4 ☐ Levik naaberorganitesse
5 ☐ Kaugmetastaasid
6 ☐ Kaugelearenenud protsess, täpsed andmed puuduvad
7 ☐ Määratlemata

RAVI
1 ☐ **Kirurgiline ravi**
1 ☐ Radikaalne
2 ☐ Palliatiivne
3 ☐ Määratlemata
Tervishoiuasutus ... Kuupäev ...
Operatsioon ...

2 ☐ **Kiiritusravi**
1 ☐ Radikaalne
2 ☐ Palliatiivne
3 ☐ Määratlemata
Tervishoiuasutus ... Kuupäev ...
Kiiritusdoos ja meetod ...

3 ☐ **Keemiaravi**
Tervishoiuasutus ... Kuupäev ...
Milline ravi ...

4 ☐ **Hormoonravi**
Tervishoiuasutus ... Kuupäev ...
Milline ravi ...

5 ☐ **Muu ravi**
Tervishoiuasutus ... Kuupäev ...
Milline ravi ...

6 ☐ **Ei saanud eriravi**
1 ☐ Kõrge vanus või kaasuvad haigused
2 ☐ Kaugelearenenud kasvaja
3 ☐ Haige keeldumine
4 ☐ Suunatud teise tervishoiuasutusse (kuhu) ...
5 ☐ Muu põhjus (milline) ...

7 ☐ **Andmed ravi kohta puuduvad**

Andmete esitaja nimi..... Ametikoht.....

Surmaaeg (päev kuu aasta)
Surmapõhjus

Tervishoiutöötaja kood.....Telefon .....

Tervishoiuasutus.....Kuupäev .....

Allkiri .....

Figure 4. Cancer Notice of the National Institute for Health Development [76].