TALLINN UNIVERSITY OF TECHNOLOGY

School of Information Technologies
Department of Health Technologies

Kristjan Krass, 178251YVEM

Data quality in health information systems – completeness and timeliness

Masters Thesis

Supervisor: Janek Metsallik, MSc

Co-supervisor: Taavi Päll, PhD

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Tervisetehnoloogia instituut

Kristjan Krass, 178251YVEM

„Andmekvaliteet tervise infosüsteemides – täielikkus ja õigeaegsus“

Magistritöö

Juhendaja:	Janek Metsallik, MSc
Kaasjuhendaja:	Taavi Päll, PhD

Tallinn 2021

Author's declaration of originality

Now declare that this Master Thesis is an original investigation and achievement submitted for the Master of Science degree of Tallinn University of Technology. Present work has not published for any other academic degree.

Kristjan Krass
(signature and date)

April 30th, 2021

Supervisor: Janek Metsallik
(signature and date)

Abstract

The overall problem in the healthcare system is medical errors [1]. The topic of the thesis is the data quality in health information systems – completeness and timeliness. The thesis investigates whether Estonian Health Information System (EHIS) data quality is sufficient for the Diagnostic Match decision support system (DSS). The author compared a five-year sample of EHIS data of 244 patients against medical records in GP systems. The findings show that the EHIS database is missing 57.7% of case summaries available in GP systems. The submission inclination was dependent on diagnoses, e.g., *Hypothyroidism* sent in 95% compared to *Migraine with aura* sent 76% on average. There are also long delays in sending the case summaries to the EHIS database, e.g. Merekivi PAK OÜ, the submission time was 73 days, Jürgenson PAK OÜ 16 days, and in Pirita PAK OÜ 19 days. The observed delays by diagnoses, e.g. *Examination for a driving license* on the same day compared to *Cystitis,* sent in an average of 17 days.

The low completeness and timeliness of the data in EHIS restrict the secondary use. GPs tend to filter data by important diagnoses, which limits the DSS algorithms accuracy. Therefore, utilising Diagnostic Match should be considered with both databases in synchrony to achieve data completeness and timeliness.

This thesis is written in English and is 61 pages long, including six chapters, 16 figures and 15 tables.

Annotatsioon

„Andmekvaliteet tervise infosüsteemides – terviklikkus ja õigeaegsus“

Andmete puudumine või nende ebatäielikkus võivad patsiendi käsitluses tuua kaasa ohujuhtumi, kus tagajärjeks võib olla tervisekahjustus [1]. Lõputöö eesmärk oli uurida kas Eesti Tervise Infosüsteemi (EHIS) andmebaasil on võimalik kasutada Diagnostic Matchi otsustustoe süsteemi (DSS). Töös võrreldakse andmete täielikkust ja õigeaegsust EHIS andmebaasis Perearst2 andmebaasiga võrdlemise teel. Uurimustöö omas eetikakomisjoni heakskiitu 17.05.2018/ Nr 2324 ning järgis andmekaitse-eeskirju. Vastavalt eesmärgile viidi läbi epikriiside päringud EHIS ja Perearst2 andmebaasides. Valim koosnes 244 isikust ja arvestas andmevahemikku 01.10.2014 kuni 30.04.2019. Andmebaasi terviklikkuse tulemused näitavad, et perearstikeskused ei saatnud 57,7% epikriisidest EHIS andmebaasi. Epikriiside edastamine sõltus diagnoosist; nt diagnoos *Kilpnäärme alatalitlust* saadeti keskmiselt 95%, *Migreen auraga* keskmiselt 76% juhtudest. Diagnoos mõjutas ühtlasi andmebaasi saatmiseks kulunud ajaintervalli; nt *Juhiloa tervisetõend* saadeti samal päeval, diagnoos *Tsüstiit* keskmiselt 17 päeva jooksul. Andmebaasi õigeaegsust uurides selgus, et perearstikeskused ei ole 2019. aastal epikriiside saatmisel EHIS-i andmebaasi piisavalt kiired; nt Merekivi PAK OÜ esitamise aeg oli keskmiselt 73 päeva, Jürgenson PAK OÜ 16 päeva ja Pirita PAK OÜ-s 19 päeva. Lõputöö tulemused näitavad, et EHIS-i andmebaasi saadetud epikriiside arv piirab saadaolevat andmehulka. Puuduliku andmed vähendavad DSS töö täpsust. Andmete täielikkuse ja õigeaegsuse saavutamiseks ja osaliste vastete vältimiseks tuleks Diagnostic Match DSS juurutamist kaaluda koos EHIS-i ja Perearst2 andmebaasidega.

Lõputöö on kirjutatud Inglise keeles ning sisaldab teksti 61 leheküljel, 6 peatükki, 16 joonist, 15 tabelit.

List of abbreviations and terms

EHR	Electronic Health Records
EMR	Electronic Medical Records
GP	General Practitioner
HIV	Human immunodeficiency virus
HL7	Health Level 7 International
ICD-10	International Classification of Diseases (10th revision)
J	Jürgenson PAK OÜ
P	Pirita-Kose PAK OÜ
M	Merekivi PAK OÜ
ICF	Informed Consent Form
PAK	Perearst2 database sample
ET	The Estonian National Health Information System database sample in the current study
XML	Text-based format for structured information transport
TAI	The National Institute for Health Development
EBMEDS	The Evidence-Based Medicine Electronic Decision Support
IC	Indicator Condition
DSS	Decision Support Software
TEHIK	Health and Welfare Information Systems Centre
EHIS	The Estonian Health Information System
AHIMA	American Health Information Management Association
EHIF	Estonian Health Insurance Fund
MISP2	The mini-information system portal/ standard X-Road component
X-Road	The integrated and secure data exchange between establishments
Query#1	Perearst2 Information System query from database
Query#2	EHIS standard: Dokumendi väljavõtte päring (diagnoosid)
Medical error	Medical error is a preventable adverse effect of care, whether or not it is evident or harmful to the patient
Ohujuhtum	Soovimatu juhtum, mille tagajärjel tekkis või oleks võinud tekkida tervisekahjustus

Table of contents

1. Introduction 10
2. Background 12
   2.1. Data quality in theory and practise 12
      2.1.1. Data quality dimensions 12
      2.1.2. Issues of data quality in EHRs 14
      2.1.3. Issues in Estonian EHRs systems 16
   2.2. The landscape of healthcare information systems in Estonia 17
      2.2.1. Brief history 17
      2.2.2. Health and Welfare Information Systems Centre 18
      2.2.3. Healthcare information systems at the primary healthcare level 20
      2.2.3.1. Decision support software by Diagnostic Match OÜ 21
   2.3. Research aims, hypothesis and questions 22
      2.3.1. Research idea 22
      2.3.2. Aims 23
      2.3.3. Hypothesis and questions 23
3. Methodology and materials 24
   3.1. Description of the study 24
      3.1.1. Methods 24
      3.1.2. Study setting and patient enrollment 25
      3.1.3. Data extraction for secondary analysis 26
      3.1.4. Ethical considerations 28
   3.2. Statistical data analysis 28
      3.2.1. Software 28
      3.2.2. Data 28
      3.2.3. Sample 29
4. Results 31
   4.1. Specification 31
      4.1.1. Population 31
      4.1.2. Timeframe 31
      4.1.3. Non-Conforming Documents in EHIS 32
   4.2. Data quality 33

4.2.1. Case summaries entered to Perearst2 ... 33
4.2.2. Diagnoses entered to Perearst2 ... 35
4.2.3. Case summaries proportion submitted to EHIS ... 38
4.2.4. Submission of different diagnoses to EHIS ... 40
4.2.5. Time submission of case summaries to EHIS ... 42
4.2.6. Time of diagnoses submission to EHIS ... 43
5. Discussion ... 45
5.1. Extracting and acquiring data ... 45
5.2. Measuring the data quality ... 46
5.2.1. Perearst2 database composition ... 46
5.2.2. Data completeness ... 47
5.2.3. Data timeliness ... 48
5.3. Conclusions ... 49
5.4. Limitations ... 49
6. Summary ... 50
6.1. Future research ... 50
Acknowledgement ... 51
References ... 52
Appendixes ... 55

List of figures

Figure 1 Pyramid of healthcare data transformation to healthcare knowledge. 12

Figure 2 Monthly data queries from the EHIS 18

Figure 3 Perearst2 healthcare information system 21

Figure 4 DSS integration in Perearst2 and remainder triggered in red. 22

Figure 5 The study enrollment phase. 25

Figure 6 Data extraction from EHIS and Perearst2 s from July 2019 till January 2020. 27

Figure 7 Number of observations per year. 32

Figure 8 Fraction of non-conforming ET document id is in our raw dataset. 33

Figure 9 Increasing numbers of case summaries per month. 34

Figure 10 Number of case summaries issued per month per patient. 34

Figure 11 Common diagnoses are issued with similar proportions 37

Figure 12 Top 10 common diagnoses versus unique “other. 38

Figure 13 Documents per year, variable intercept for each PAK. 39

Figure 14 Proportion of diagnoses submitted to EHIS. 41

Figure 15 Time to submission in days of case summaries to ET 42

Figure 16 Time of submission diagnoses to ET 44

List of tables

Table 1. The five dimensions of data quality and words are often used 13

Table 2. Data Quality Management ten dimensions and analysis................................. 14

Table 3. EHIS database clinical documents and events 2019 19

Table 4 GP Centres in the study and the population 26

Table 5 Head of a table of data queries from Perearst2 information system to EHIS.... 27

Table 6 Example of Head of merged PAK and ET documents data table. 29

Table 7 Summary of the study population.. 31

Table 8 Number of ET entries with the non-conformed document number. 32

Table 9 Examples of non-conform doc_no. 32

Table 10 Top 20 most frequent diagnoses.. 35

Table 11 Description of top 20 diagnoses common at all three GP Center.................... 36

Table 12 The proportion of case summaries submitted to EHIS.................................... 38

Table 13 The proportion of diagnoses submitted to EHIS by GP Centres..................... 40

Table 14 Average time to submit case summaries to ET 42

Table 15 Top 20 diagnoses of time to submission of to EHIS in days........................... 43

1. Introduction

The insufficiency in healthcare data completeness and timeliness can lead to information unavailability in the care process. The problem identified in 2016 by Makary *et al.* is that nearly 250,000 patients die each year due to medical errors in the United States [1]. One approach to reducing medical errors involves the use of decision support software. A famous proverb in Latin says: *Errāre hūmānum est* – meaning everyone makes mistakes – when it is a random human fault, then it is understandable, but if it is a systemic and possible discrepancy in the setting or procedures, then it is costly [2]. The literature review in 2015 from Davoudi *et al.* shows that accurate data usage in decision support software means that the databases must have data without contradictions, uniformly use classifications, authentic and up to date [3]. In 2017, The National Institute for Health Development (TAI) reported a study about GP Centres outpatient consultations data transmitted to The Estonian Health Information System (EHIS) and cross-examined 2015-year documents records for data quality. However, the conducted secondary analysis failed to estimate yearly trends and the study report misses the organised information on how data were linked and compiled [4].

In 2016, a team called "Diagnostic Match" took part in the HIV-Digital hackathon. The group proposed an innovative way to find hidden Human Immunodeficiency Virus (HIV) positive patients. Estonia has a big problem with HIV, with the highest incidence rates of HIV in Europe and over 14 new cases per 100,000 population each year [5]. The idea was to make a decision-support tool for healthcare workers to detect hidden HIV positive patients with automatic HIV indicator condition disease algorithm, or other words, to make decision support software (DSS). The application analyses patient's health data and displays reminders for a doctor to make an HIV test if the patient is in the at-risk group. In 2019, the proof-of-concept study was initiated in the HIV indicator disease algorithm "Enhancing HIV Indicator Disease-Guided Testing Strategy Implementation in Estonia by Using HIV Clinical Decision Support System - a Pilot Study in Primary Care" [6].
The prior mentioned study population was the starting point for the current thesis secondary analysis about completeness and timeliness of data quality.

The overall goal of this thesis is to measure the electronic health records (EHR) completeness and timeliness for use in decision support software. It is interesting to investigate whether The Estonian Health Information System (EHIS) and GP Centres information system software "Perearst2" database case summaries data submitted on October 1st, 2014, till April 30th, 2019 varied for the use of Diagnostic Match DSS. This thesis aims to test the suitability of the EHIS data for Diagnostic Match DSS by examining the impact of querying, linking, and combining EHRs datasets for establishing the data available for secondary analysis from the sample of three general practitioners (GP) Centres in Tallinn with a cohort of 244 study participants. The study estimates the data completeness and timeliness in EHIS by comparing it to the original records at GP Centres. With this aim in mind, the objective is to present a method for query, link, and merge EHRs datasets for the secondary analysis.

One of the primary benefits of DSS in healthcare is improved outcomes in treatment and healthcare processes. Therefore, the healthcare data quality analysis and investigation of if the data is usable, accessible, and interoperable for a healthcare worker in any place or distance to make correct medical decisions. In 2017 Wagner *et al.* stated the necessity for reliable and fast information to provide quality care, containing costs and guarantee adequate access to healthcare. Therefore, efficient data flow is vital for healthcare organisations to establish secondary data use in decision support software. The healthcare data have to be available to the right people at the right place and at the right time [2].

Notably, the Estonian Health Insurance Fund has launched a nationwide DSS covering 1,600 algorithms and 45,000 messages for national wide use in 2020 May [7]. The nationwide DSS uses EHIS data. Hence the importance to develop research methods to measure the data quality of EHIS even raises.

2. Background

2.1. Data quality in theory and practise

2.1.1. Data quality dimensions

The healthcare information system data are a good source for research in epidemiology, drug surveillance, public health, and decision support. With the availability of EHR databases, the researchers are exceedingly interested in secondary data use. Although EHRs data holds a great promise, unfortunately, the datasets have various data quality issues [2]. The healthcare information systems must represent the up-to-date data for a given task. The data is expected to be 100% present. The missing values are not an option. The data must have a good range and depth for a particular task from a specific user's viewpoint. The data has to be complete and relevant to support reliable and safe services for patients. In healthcare information systems, the data transforms into information and filter to new knowledge (see Figure 1), meaning that knowledge is a combination of logics, best practice guidelines, relationships, concepts, and experience [2] [8].

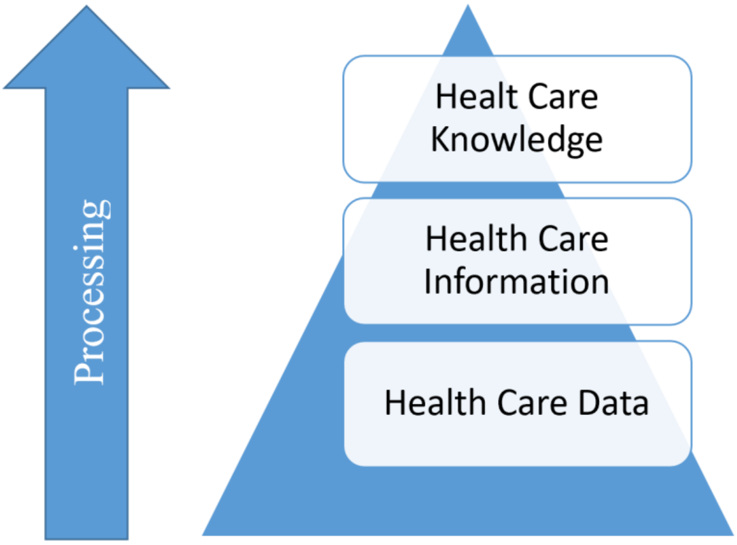


Figure 1 Pyramid of healthcare data transformation to healthcare knowledge [2].

In 2013 Weiskopf *et al.* reviewed the question of how to evaluate healthcare data. It would be good to have one single data quality evaluation measure or standard that all can use. It would be easier if the purpose of the data is determined, but with complex information systems, different data dimensions are present. Weiskopf *et al.* developed a framework for organisations to evaluate healthcare data five quality dimensions:

- Completeness: a review of the presence or absence of the data elements.
- Correctness: a review of the data elements that have predictive value.
- Concordance: a review agreement between data elements.
- Plausibility: a review of data validity and integrity.
- Currency: a review within a set time limit.

However, Weiskopf *et al.* considered that only three are fundamental - correctness, completeness, and currency describe the data quality core concepts related to EHRs secondary usage. In the review the Weiskopf *et al*. found that the terminology describing the dimensions often overlaps in the scientific literature (see Table 1), making harmonisation difficult when evaluating data quality [9].

Table 1. The five dimensions of data quality and words are often used [9].

Completeness	**Correctness**	**Concordance**	**Plausibility**	**Currency**
Accessibility	Accuracy	Agreement	Accuracy	Recency
Accuracy	Corrections made	Consistency	Believability	Timeliness
Availability	Errors	Reliability	Trustworthiness	
Missingness	Misleading	Variation	Validity	
Omission	Positive predictive value			
Presence	Quality			
Quality	Validity			
Rate of recording				
Sensitivity				
Validity				

The dimensions help to choose the right assessment tool or methodology to assess health data quality. For example, methods could be the gold standard, element agreement, source agreement, element presence validity checks, log review and distribution comparison. When evaluating, it is essential to be aware of task-dependence and interoperability of data quality and knowledge of the dataset and the study objectives [2] [9].

In 2015 the American Health Information Management Association (AHIMA) developed another guideline to include all health data in the data quality evaluation model. The six most common dimensions in the Data Quality Management Model are:

- Data accuracy: a review correctly reflected data and valid values.
- Data accessibility: a review of missing data.
- Data comprehensiveness: a review of the required data
- Data consistency: a review for the data consistency.
- Data_currency: a review of data availability in time.
- Data definition: a review of data elements within precise definitions.

Overall, The AHIMA has identified ten dimensions that help define the process and understand data quality analysis (see Table 2) in healthcare organisations for management processes [3].

Table 2. Data Quality Management ten dimensions and analysis [3]

Dimensions	**Definition**
Data Accuracy	Free of identifiable errors.
Data Accessibility	Legality to user access processes with controls
Data Comprehensiveness	All of the data is collected
Data Consistency	Data reliability across applications
Data Currency	The data have to be up to date on the specific time
Data Definition	The meaning of the data element
Data Granularity	Detail of the attributes and values defined
Data Precision	Data values should support the purpose
Data Relevancy	The data collected should be helpful for the purposes
Data Timeliness	The data should be available within a practical time frame and up to date in the context

2.1.2. Issues of data quality in EHRs

Nowadays, when almost all healthcare providers adopt EHRs, the emphasis must shift towards leveraging to achieve better care quality, efficiency, and safety. However, research suggests that the EHRs high prevalence of unstructured data leads to health data quality issues for the computer systems to support process optimisation. In contrast, designated data fields in medication administration systems are a clear example of data quality improvement to prevent medication side effects and coordination. Process optimisation requires additional investments into information systems to fix the issues with data quality. EHRs need tools for error checking, e.g. structured data entry, drop-down lists, and templates for accuracy, consistency, and completeness for better data quality [2].

The occurrence of incomplete data in EHRs has reviewed by Chan *et al.* for the estimated data accuracy, meaning that the information system accurately reflects the number, dose, and specific drugs the patient is currently taking, including the timeliness. The study data showed that accuracy was only 66%. Another aspect of the review was the completeness of data, and the missing data element level was 57%. The lack of data points in EHRs and other data errors significantly reduce the reliability and validity of the DSS [10].

Prior research by Wagner *et al.* suggests that DSS has to fit into the workflow and access all diagnostic data available in different electronic health records sources, e.g., hospitals, GP Centres. It ensures the proper processes and patient safety, and treatment quality. Most importantly, the correctness, completeness, and currency of the exchanged data. Increased use of electronic records pushes for bigger and better data analytics and quality checks to ensure healthcare data and knowledge [2].

In 2013 Bayley *et al.* accessed four healthcare information systems to extract data from EHRs and got five different issues during the study:

- Data errors, e.g. incorrect data from the blood pressure measurements.
- Missing data, e.g. ICD-10 code for the disease, can identify patients with uncontrolled blood pressures: but the actual blood pressure measurements are missing.
- Uninterpretable data, e.g. blood pressure data, is not consistently collected.
- Unstructured text data, e.g. the information is written to the note field and not into the coded field.
- Inconsistent data, e.g. software developers collect data in different coding or older versions or standards.

When reusing the data of healthcare information systems, the DSS outcomes may end up being problematic and unstable. In the study, Bailey *et al.* noted that the data quality of EHR must improve for secondary data usage, and it must be developed separately from clinical care [11].

In 2015 the AHIMA guideline instructed that data quality management for healthcare providers must define the processes for the data collection and application in databases. Using healthcare data standards helps the data exchange and interoperability, e.g. ICD-10-CM/PCS, SNOMED CT, LOINC. The outcome of data quality management by implementing structured data entry, drop-down lists, and templates for accuracy, consistency, and completeness improve data quality leads the way to the knowledge that it is applicable to use for all the processes [3].

2.1.3. Issues in Estonian EHRs systems

The National Institute for Health Development (TAI) in 2016 conducted a study that analysed EHIS for hospital inpatient and outpatient case summaries submitted in the year 2015. The researchers looked for data accuracy, data accessibility, and missing data. Data analysis showed that in 2015 hospitals reported only 24% of inpatient cases within a day or two and 67% of inpatient case summaries in a week. In total, hospitals submitted 96% of discharge letters to EHIS in a month [12].

In 2017 TAI organised a study for the primary healthcare centres. The study included the outpatient case summaries submitted to the EHIS by 468 general practitioners (GP) centres in 2015. All GP centres had electronic medical records system and were able to send EHRs data to EHIS. Nevertheless, outpatient data submitted was unsatisfactory – GPs reported only 22% of the case summaries; therefore, 78% of data was not available in EHIS. Researchers also found other issues with the data quality during the study [4].

Common mistakes found by TAI (2017) on case summaries that lead to data errors:

- Data accuracy issues, e.g. repeated visits and minor or recurring diagnoses, are often not transferred with case summaries.
- Data relevancy issues, e.g. in case summaries and medical examination notices, had the duplicated data.
- Data accessibility issues, e.g. doctors and nurses, are reflected together on the same case summary.
- Data consistency issues, e.g. 0.6% of the case summary entries were with incorrect dates.

In 2016 and 2017, studies by TAI researchers highlighted the need to reduce input errors by implementing automatic error checks and coding the controls, e.g. date values or duplicates of case summaries that result in data errors in information systems. The critical finding was that data timeliness is not satisfactory, meaning that data is not available at the right time to be reviewed within a set time limit [4] [12].

In 2015 Ross and Metsallik analysed the EHIS, Estonian Health Insurance Fund database and Estonian Genome Centre of the University of Tartu database for the feasibility of national comprehensive decision support software for personalised medicine. They found that some of the data is standardised and structured, but a substantial part of the data is unstructured and difficult to use for DSS. The investigation showed that the developed DSS is applicable in the Estonian national-wide health information system in personalised medicine [13]. In 2017 a follow-up analysis by Ross and Metsallik concluded that implementing DSS algorithms at the GP Centres information system database and the hospital information system database is technically viable. However, the existing databases often miss essential data, limiting DSS to provide reliable results and challenging using DSS in real-time clinical decision-making [14]. Regardless of the results of previous studies, in 2020, EHIF has launched a nationwide DSS that uses EHIS data and covering 1,600 algorithms and 45,000 messages in clinically vital topics [7].

2.2. The landscape of healthcare information systems in Estonia

2.2.1. Brief history

From 1990 to 2000, healthcare service providers, e.g. hospitals, GP Centres, started using electronic medical records systems (EMR). Many software companies were founded at that time, focusing on developing health information systems, e.g. Medisoft, Gennetlab, Nortal. At the end of the nineties, the planning of a nationwide electronic healthcare system started. In 2001, the Estonian Health Insurance Fund (EHIF) launched a digital invoicing system and, since 2002, signed contracts for electronic data transmission with 76% of healthcare service providers and 54% of pharmacies. By 2005 already 100% of the invoices and reimbursement information was submitted electronically. The main purpose of advanced software was to submit medical claims and share patient administration information to the EHIF. [15] [16] [17].

2.2.2. Health and Welfare Information Systems Centre

2005 the Estonian Government initiated an all-inclusive nationwide electronic health record system as a part of the Estonian Information Society Strategy and in 2006 founded The Estonian E-Health Foundation. The established foundation meant to initiate and implement e-health activities and develop and manage healthcare providers and other healthcare providers to cooperate better with faster communication. Since 2017 foundation was merged under the Ministry of Social Affairs IT-department and renamed into Tervise ja Heaolu Infosüsteemide Keskus (Eng.: Health and Welfare Information System Centre) or, in short, TEHIK and regulated by law Healthcare Organization Act ("Tervishoiuteenuste korraldamise seadus") in §59^{1}. Nowadays, by estimation, more than 10,000 healthcare workers use the system daily [15] [16] [18] [19] [20].

Since the EHIS initiation in 2008, the usage of its data has steadily increased. The queries from patients and healthcare professionals have reached more than two million per month (see Figure 2). The ability to submit data by healthcare providers to EHIS is technically sufficient as at the end of 2015, all GP Centres (468) were able to transmit case summary to the EHIS [4].

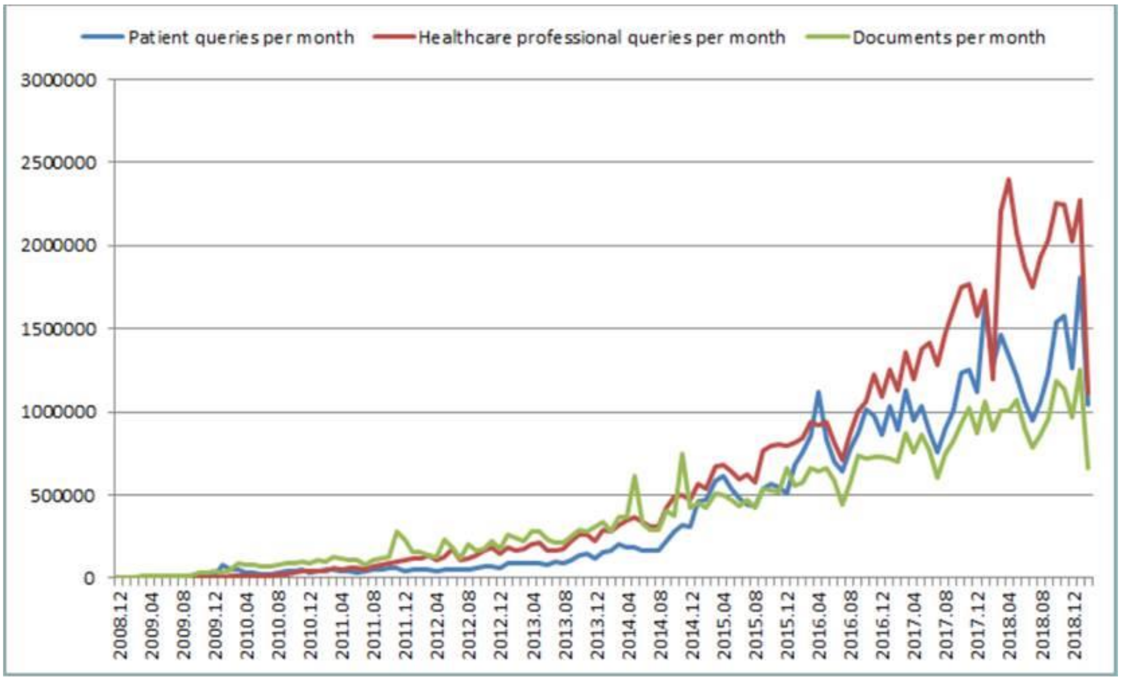


Figure 2 Monthly data queries from the Health and Welfare Information System Centre database (Data Source: TEHIK 02.2019)

Health and Welfare Information System Centre (TEHIK) develops and manages nationwide standards for health information exchange. All healthcare service providers must submit data to EHIS. The data submission is regulated with 24 different clinical document standards [19] [21]. Secure communication is an essential cornerstone of health data communication. Access to the EHIS is granted only to licensed medical professionals. The EMR systems communicate with EHIS through X-road. Users must use e-ID for authentication and digital signature for integrity. Both the medical personnel and the patients are required to use secure authentication [19].

At the beginning of the year 2019, the EHIS contains 49 million clinical documents (see Table 3) of clinical work, e.g. referral answers and ambulatory cards. The data in documents contain structured values, e.g. diagnosis with ICD-10 coding and unstructured text fields. Structured data are using international classification of, e.g. ICD-10, LOINC, NCSP, ATC and has universally documented in all information systems.

Estonian citizens have a unique identification number that makes data interoperability possible between information systems and allows the developing nationwide DSS. Overall, it is valuable input for secondary data usage and analysis [22].

Table 3. EHIS database clinical documents and events 2019 (data source: TEHIK 02.2019)

DOCUMENT TYPE	**COUNT**
Ambulatory case summary	24,193,514
Referral answer	13,393,873
Dental card	3,615,404
Referral	2,969,933
Hospital case summary	2,054,293
Immunisation notification	1,130,684
Ambulance card	798,300
Children health check notification	691,857
Health certificate (for drivers license)	423,808
Day-care case summary	396,018
Home nursing summary	11,334
Total	**49,679,018**

EHIS information system uses XML-based HL7 v3 messages that facilitate the exchange of medical information for healthcare provider systems. The clinical documents in EHIS are stored in XML format according to the HL7 CDA R2 standard. The data identifiers have OID-s. Only final versions of clinical documents are submitted into EHIS after the health provider closed medical record from their healthcare information system, e.g. Liisa, HEDA, Perearst2 [19].

The broad objective of TEHIK is to endorse the patient-centred healthcare system, and by initiating a patient portal (https://www.digilugu.ee/) all patients have an overview of healthcare services received. The patient has the right to conceal (right to opt-out) the data collected into EHIS. The objective is that collected healthcare data is always high quality and transparent and controllable and safe, and linked with relevant state registers and other databases to facilitate secondary data cross-analysis [18] [19].

2.2.3. Healthcare information systems at the primary healthcare level

Healthcare service providers have the responsibility to transfer medical data to EHIS. The user experience serviceability of the information from the health information system depends on the developed EHRs software. There are four general practitioners' information systems Perearst2, Perearst3, 5D -MED, Watson and Arstiportaal + [21] [23].

The leading health information developer at the primary healthcare level is Medisoft Perearst2. The program uses Microsoft SQL 2005 or higher database engine and Delphi XE programming language for providing EMR (see Figure 3). Medisoft claims that 85% of Estonian GP Centres use their software, mainly the health information system Perearst2 provides for GP Centres the timeliness and truthfulness of the data by storing the data correctly and securely [24]. However, in 2014 Vanker examined that the user experience of the Perearst2 software is cumbersome by describing the difficulty to find information and too many clicks and scrolls to navigate the menus [13] [25].

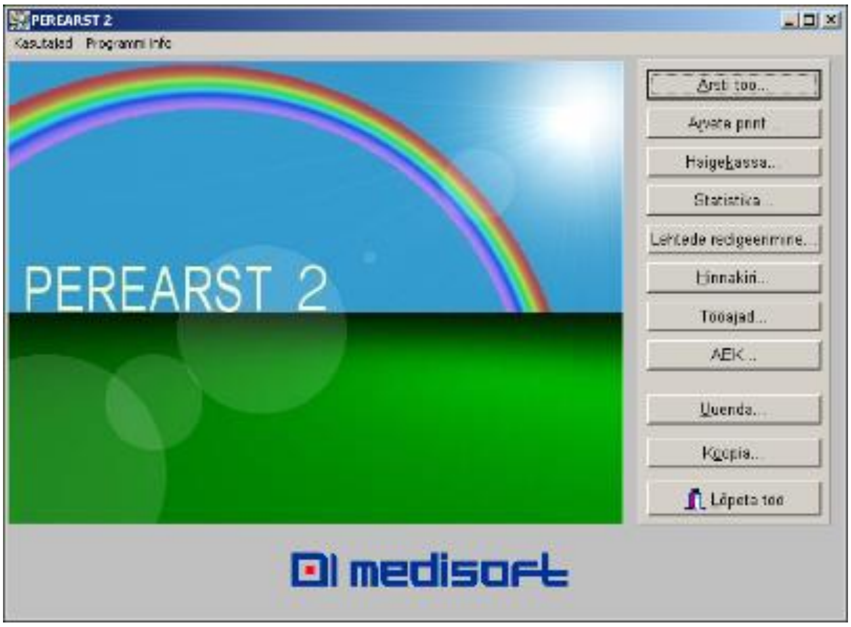
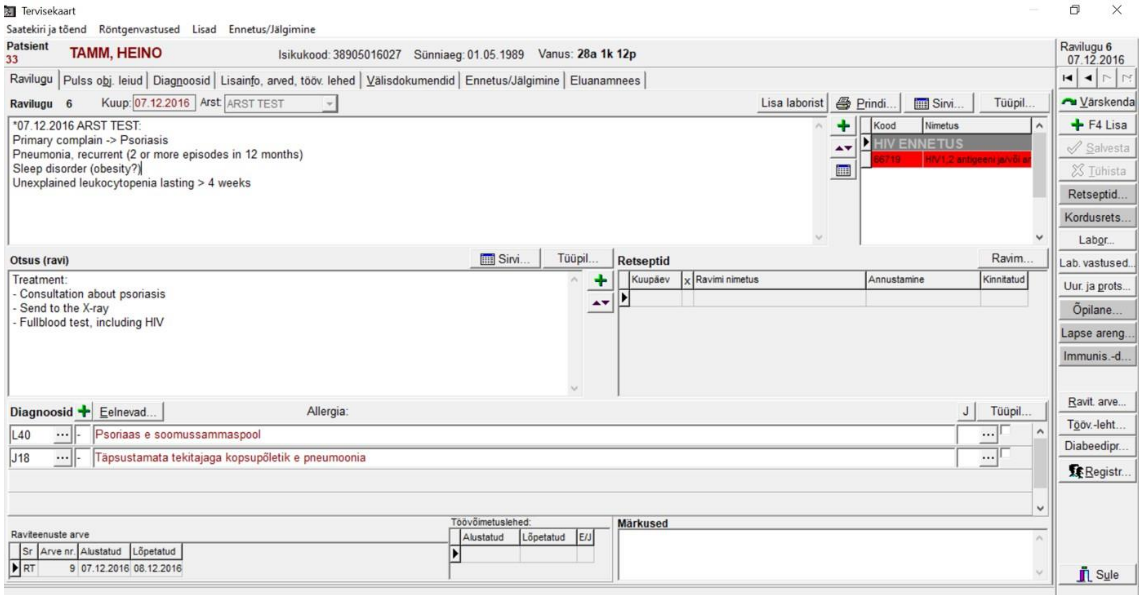


Figure 3 Perearst2 healthcare information system [24].

2.2.3.1. Decision support software by Diagnostic Match OÜ

Diagnostic Match OÜ provided decision support software (DSS) is integrated on Perearst2 Software to remind family doctors of the possible HIV indicator condition disease, and the DSS algorithm triggers clinical reminder suggesting to do HIV testing (see Figure 4) [26] [27] [28]. The DSS clinical remainder improves diagnostic accuracy by detecting hidden HIV positive people from the history of EHRs [6].

In 2019, there was a proof-of-concept study of the HIV indicator disease algorithm "Enhancing HIV Indicator Disease-Guided Testing Strategy Implementation in Estonia by Using HIV Clinical Decision Support System - a Pilot Study in Primary Care". The results showed that detecting HIV indicator condition diseases at the primary healthcare level increased HIV-testing four-fold in Estonia [6].



Patsient 33 TAMM, HEINO Isikukood: 38905016027 Sünniaeg: 01.05.1989 Vanus: 28a 1k 12p

Ravilugu 6 07.12.2016

Ravilugu 6 Kuup: 07.12.2016 Arst: ARST TEST

*07.12.2016 ARST TEST:
Primary complain -> Psoriasis
Pneumonia, recurrent (2 or more episodes in 12 months)
Sleep disorder (obesity?)
Unexplained leukocytopenia lasting > 4 weeks

Kood	Nimetus
HIV ENNETUS	
66719	HIV1,2 antigeeni ja/või a

Otsus (ravi)

Treatment:
- Consultation about psoriasis
- Send to the X-ray
- Fullblood test, including HIV

Retseptid

Kuupäev	x	Ravimi nimetus	Annustamine	Kinnitatud

Diagnoosid Allergia:

L40	-	Psoriaas e soomussammaspool
J18	-	Täpsustamata tekitajaga kopsupõletik e pneumoonia

Raviteenuste arve

Sr	Arve nr.	Alustatud	Lõpetatud
RT	9	07.12.2016	08.12.2016

Töövõimetuslehed:

Alustatud	Lõpetatud	E/J

Märkused

Figure 4 DSS integration in Perearst2 and remainder triggered in red [6].

Preferably, DSS have to have a complete overview of patient conditions by accessing all diagnostic data available in multiple sources nationwide of EHRs, e.g. hospitals, GP Centres and EHIS. However, the developed Diagnostic Match OÜ DSS algorithm queries the data from a Perearst2 database. The definitive goal is to create the application where the program can extract data and analyse it throughout nationwide EHR by evaluating HIV indicator condition diseases for HIV-testing [6].

2.3.Research aims, hypothesis and questions

2.3.1. Research idea

The thesis idea originated from a proof-of-concept study of the HIV indicator disease algorithm “Enhancing HIV Indicator Disease-Guided Testing Strategy Implementation in Estonia by Using HIV Clinical Decision Support System - a Pilot Study in Primary Care” [4]. During the pilot-study data collection in 2019, it was increasingly evident that DSS algorithms are dependent on the quality of data sources, and there is a need for methods to evaluate them. Also, in 2017 Wagner *et al.* literature review emphasised that the DSS requires the quality of data in EHRs. It is essential to have timely access to the data in clinical documents. The information should be accurate for a given task in hand. It is vital to establish fair use of secondary data in healthcare organisations [2]. According to studies in 2015 and 2017 conducted by TAI, healthcare providers have discrepancies in case summary document submission to EHIS, resulting in data loss [4] [12].

2.3.2. Aims

The thesis aims to compare the data completeness and timeliness in EHIS and Perearst2 databases to understand whether it is possible to utilise Diagnostic Match DSS on the EHIS data.

2.3.3. Hypothesis and questions

The thesis hypothesis is that the data in EHIS and Perearst2 databases vary significantly due to the limited submission and availability within a reasonable time of the case summaries from GP Centres.

Question #1: What is data completeness in the EHIS database when compared to the Perearst2 database?

Question #2: Does the EHIS database have data available from GP Centres within the set time limit?

3. Methodology and materials

3.1. Description of the study

3.1.1. Methods

The study's formal preparation started in 2018 with the description of processes and an application to the ethical committee from Diagnostic Match OÜ. The study received formal ethical approval from Tallinn Medical Research Ethics Committee decision No. 2324; 17.05.2018.

The study cohort for secondary data analysis originates from a proof-of-concept study of the HIV indicator disease algorithm "Enhancing HIV Indicator Disease-Guided Testing Strategy Implementation in Estonia by Using HIV Clinical Decision Support System - a Pilot Study in Primary Care" [6]. The HIV-testing exposure triggered by DSS with HIV indicator condition disease remainder in real-life GP Centres in October 2018 till April 2019, with patient enrolment with an informed consent form (ICF) to the sample (see Appendix 2).

The objective is to find a solution for linking the EHIS and Perearst2 databases and compare EHRs for data completeness and timeliness by implementing the EHIS repository and Perearst2 database queries for case summaries with Medisoft Perearst2 software. With this objective in mind, the focus is to present a method for query, link up, and merge EHRs datasets of ICD-codes, case summaries for the secondary analysis.

It is a retrospective cohort study of the data collected from the two databases. The data is extracted by querying the five-year data from EHIS and Perearst2, October 1st, 2014 to April 30th, 2019. The extraction resulted in a combined database for further analyses. Multi-level logistic modelling of counts and time to submission is the statistical methods used to describe medical document submission patterns.

In 2017, Bridge *et al.* described that the binomial regression technique could correctly model over-dispersed count data and account for the grouping of individual patient-level data within GP Centres. Multi-level binomial regression modelling and analysis can serve as an illustration to count different important outcomes in primary and other healthcare research [29]. A similar retrospective cohort study conducted previously in 2015 by TAI compared the EHIF and EHIS databases for inpatient case summaries submitted by hospitals. The study analysed the data quality of EHIS [12]. Another retrospective cohort study in 2017 by TAI evaluated data quality by comparing outpatient case summaries submitted to the EHIS and reported directly to TAI by GP Centres [4].

3.1.2. Study setting and patient enrolment

The study is a continuation of a before-mentioned HIV indicator disease study [4]. The same three GP Centres from the referred study were used as data collection sites also in this study. For the earlier study, the sites agreed to install Diagnostic Match OÜ DSS to the Perearst2 information system (see Figure 5). The GP Centres were highly computerised and operated daily without written notes and keep data records only electronically. The three GP Centres represent a sample of 17,162 people in Tallinn urban areas.

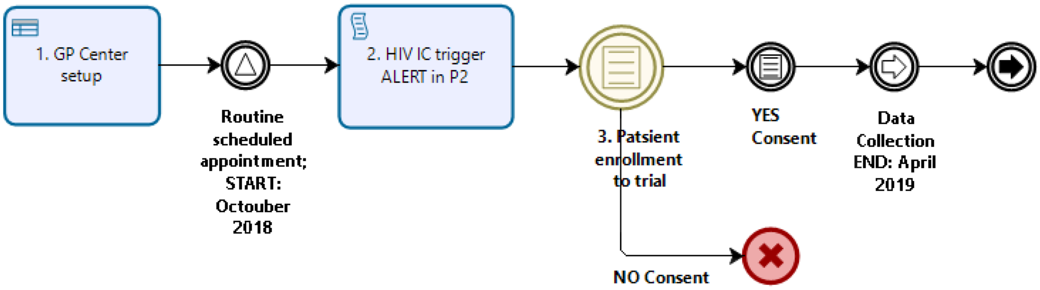

1. GP Center setup
Routine scheduled appointment; START: Octouber 2018
2. HIV IC trigger ALERT in P2
3. Patsient enrollment to trial
YES Consent
Data Collection END: April 2019
NO Consent

Figure 5 The study enrolment phase.

GPs enrolled the study subjects in real-life primary healthcare settings during routine visits. The DSS triggered for inclusion of a study participant. The algorithm for the remainder for HIV-testing used criteria of 216 ICD-10 codes (seeAppendix 3). Study patient’s inclusion/exclusion criteria for the 6-month study period was: a) aged ≥18 or ≤ 65 years at the study’s time; b) at least one HIV Indicator Condition (IC); c) no HIV-test in the past six months. Patients have excluded if: a) age >65 or <18 years at the time; b) no HIV IC in five years; c) done HIV-test in six months; d) HIV positive; e) did not sign ICF.

The patients with HIV indicator condition disease agreed to be enrolled with ICF to study at the GP Centre between October 1st, 2018 and April 30th, 2019. A total of 244 (1,4%) patients agreed to enrol in the study from three GP Centres (see Table 4).

Table 4 GP Centres in the study and the population (data source: EHIF, 01.01.2019)

GP Center	**Software**	**Population (01/01/2019)**	**Participants (01/10/18-30/04/19)**	**% of patients in the study**
Merekivi PAK OÜ	Perearst2	7,244	117	1,6%
Jürgenson PAK OÜ	Perearst2	5,200	62	1,2%
Pirita-Kose PAKOÜ	Perearst2	4,718	65	1,4%
Total		**17,162**	**244**	**1,4%**

In 2019, Kikas, in her master thesis, described in detailed the patient enrolment phase that preceded secondary analysis [6].

3.1.3. Data extraction for secondary analysis

For the secondary analysis, the data extraction and connecting multiple data sources included matching (linkage) data and safeguarding that records refer to the correct patient [30]. The study patient's data from EHIS and Perearst2 was collected using primary healthcare software Perearst2 from July 2019 till January 2020. All the processing took place on the premises of the GP Centres and by the healthcare worker.

- First, GP Centres queried data from the Perearst2 (Query#1). The database supported extraction to the *CSV* file. The extraction included manual removal of text fields with patient personal data: names, addresses, telephone numbers, e-mail address. (see Figure 6).
- Second, GP Centres queried data from the EHIS database. The query was run via Perearst2 software, which used standard diagnoses query or EHIS - "Tervise infosüsteemi standard: (diagnoosid)" [31] (Query#2).

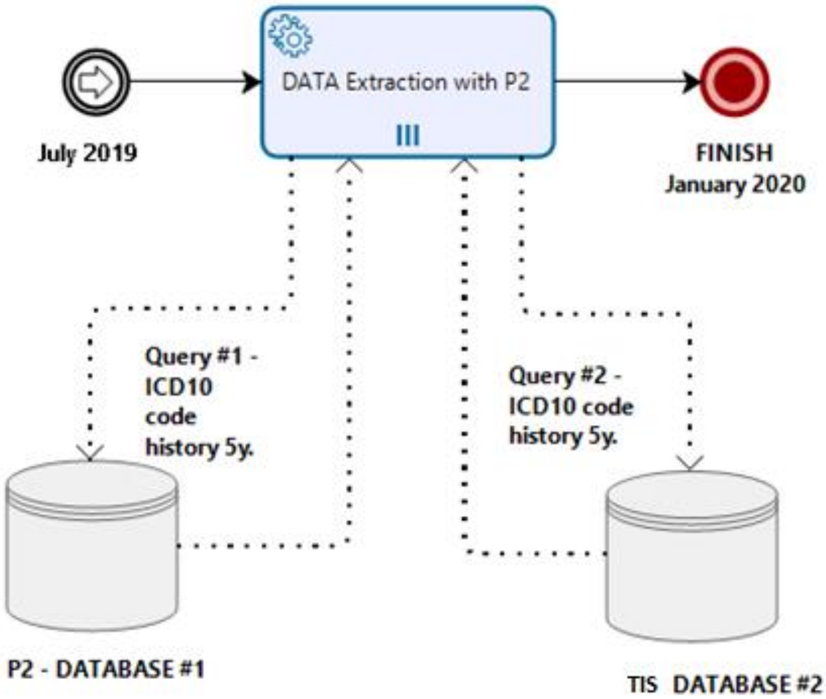
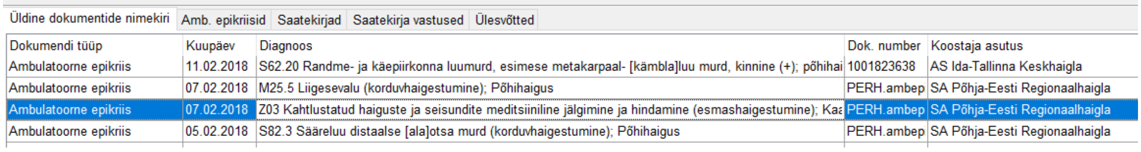


Figure 6 Data extraction from EHIS and Perearst2 s from July 2019 till January 2020.

The patient data queried with Query#1 from Perearst2 database with the following structural fields: a) Medical Document Type; b) The Timestamp - Of Medical Document Creation in Perearst2; c) Diagnosis – ICD-10 code. Data queried with Query#2 from EHIS database with the following structural fields: a) Medical Document Type; b) The Timestamp - Of Medical Document Creation in EHIS; c) Diagnosis – ICD-10 code; d) Medical Document identifier; e) Owner of the Medical Document (see Table 5). Extracted EHR datasets included all medical diagnoses five-years historically recorded in Perearst2 and EHIS.

Table 5 Head of a table of data queries from Perearst2 information system to EHIS



Dokumendi tüüp	Kuupäev	Diagnoos	Dok. number	Koostaja asutus
Ambulatoorne epikriis	11.02.2018	S62.20 Randme- ja käepiirkonna luumurd, esimese metakarpaal- [kämbla]luu murd, kinnine (+); põhihai	1001823638	AS Ida-Tallinna Keskhaigla
Ambulatoorne epikriis	07.02.2018	M25.5 Liigesevalu (korduvhaigestumine); Põhihaigus	PERH.ambep	SA Põhja-Eesti Regionaalhaigla
Ambulatoorne epikriis	07.02.2018	Z03 Kahtlustatud haiguste ja seisundite meditsiiniline jälgimine ja hindamine (esmashaigestumine); Kaa	PERH.ambep	SA Põhja-Eesti Regionaalhaigla
Ambulatoorne epikriis	05.02.2018	S82.3 Sääreluu distaalse [ala]otsa murd (korduvhaigestumine); Põhihaigus	PERH.ambep	SA Põhja-Eesti Regionaalhaigla

The progression of data extraction from Query#2 had stalled by the absence of any data export functionality by the Perearst2 information system. Therefore, screenshots images from Perearst2 data tables healthcare workers took with Windows 10 Snipping application. The data tables resulting in images were converted to Microsoft EXCEL with ABBY FineReader15 from January 2020 till August 2020.

The GP Centre's healthcare worker handled all data extraction process. Data extraction included also removing text fields with patient personal data: names, addresses, telephone numbers, e-mail address. Data were de-identified before being handed for secondary analysis.

3.1.4. Ethical considerations

The study followed the firm data protection rules – no personalised information was shared outside the healthcare service providers. Researchers used the safe harbour protocol to de-identify healthcare info before sharing the dataset with an outside party.

3.2.Statistical data analysis

3.2.1. Software

Statistical analysis was done in *R* vers. 4.0.3 [32]. Tools from R tidyverse package were used for data wrangling [33]. *readxl R* package was used to import data from MS Excel spreadsheets [34]. Bayesian modelling was performed with *R* packages *rstan* vers. 2.21.2 [35] and *brms* vers. 2.13.3 [36]. Models were processed and visualised with *tidybayes* [37] and *modelr* [38]. *ggplot2* vers. 3.3.1 [39] and *cowplot* [40] R packages were used for graphics. *glue* [41] R package was used as a helper to generate dynamic strings. *here* [42] R package was used to track project files from scripts.

3.2.2. Data

Raw data was imported from *MS Excel* spreadsheets using *readxl R* package [34] and saved to *CSV* file using import *R* script. However, the raw data from Query#2 resulting in screenshot images extraction text contained several observations with missing characters and spelling errors. Data cleaning and validation were done programmatically by using a clean R script. The cleaned-up dataset was saved to a separate *CSV* file. The cleaned dataset contained observations from both Perearst2 (PAK) and EHIS (ET) databases.

The author assumes that all diagnoses given to a patient at PAK dataset on a single day belong to one case summary.

In the ET dataset, the document number used for grouping diagnoses belonging to the same case summary with a date, and the PAK dataset contains diagnoses assigned to a date. Each ET document includes a set of one or more diagnoses.

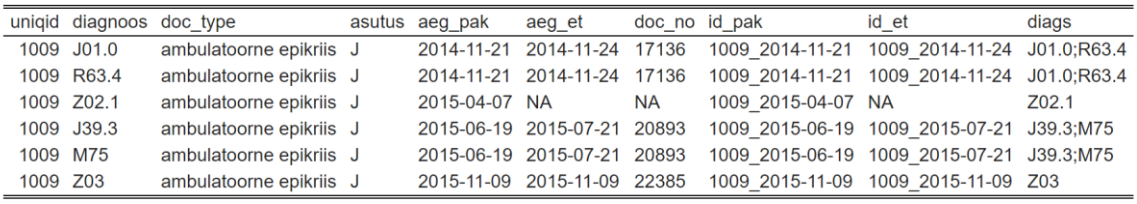
The sets of diagnoses generated for both PAK and ET datasets, based on patient id and diagnosis date and then the PAK dataset was joined to ET dataset by patient id and by matching sets of diagnoses into the merged database.

Observations belonging to PAK and ET databases were saved to *csv* file *pak_data.csv* and *et_data.csv*, respectively. Document type and institution name were considered but not used for matching the observations—the merged data saved to *CSV* file *merged_documents.csv*.

Original variable column names and values were normalised and converted to syntactically correct form but not translated. Generated new variables and values are using English notation.

Imported data of merged PAK and ET case summaries and diagnoses include the following variables (see Table 6): a) unique: patient id; b) diagnoos: diagnose ICD code; c) doc_type: document type in PAK; d) asutus: PAK name; e) aeg_pak: diagnosis date in PAK database; f) aeg_et: document date in ET database; g) doc_no: document number in ET database; h) id_pak: document id generated by linking patient id and PAK diagnosis date; i) id_et: document id generated by linking patient id and ET diagnosis date; j) diags: diagnoses included in id_pak and id_et.

Table 6 Example of Head of merged PAK and ET documents data table.



uniqid	diagnoos	doc_type	asutus	aeg_pak	aeg_et	doc_no	id_pak	id_et	diags
1009	J01.0	ambulatoorne epikriis	J	2014-11-21	2014-11-24	17136	1009_2014-11-21	1009_2014-11-24	J01.0;R63.4
1009	R63.4	ambulatoorne epikriis	J	2014-11-21	2014-11-24	17136	1009_2014-11-21	1009_2014-11-24	J01.0;R63.4
1009	Z02.1	ambulatoorne epikriis	J	2015-04-07	NA	NA	1009_2015-04-07	NA	Z02.1
1009	J39.3	ambulatoorne epikriis	J	2015-06-19	2015-07-21	20893	1009_2015-06-19	1009_2015-07-21	J39.3;M75
1009	M75	ambulatoorne epikriis	J	2015-06-19	2015-07-21	20893	1009_2015-06-19	1009_2015-07-21	J39.3;M75
1009	Z03	ambulatoorne epikriis	J	2015-11-09	2015-11-09	22385	1009_2015-11-09	1009_2015-11-09	Z03

3.2.3. Sample

The resulting sample contains observations where one PAK case summary and diagnosis is matched to one ET document. Patient case summaries in the PAK dataset can include multiple diagnoses in the merged PAK and ET data table (see Table 6).

It implies that the time between case summary creation at PAK database and submission to ET database is the same for all diagnoses in the document. Consequently, one diagnosis was selected randomly from each document for further analysis.

For reproducibility, the sample seed was set to 2021 with *set.seed* command. The final sample was saved to a *CSV* file (*diags_sample.csv)*.

4. Results

4.1.Specification

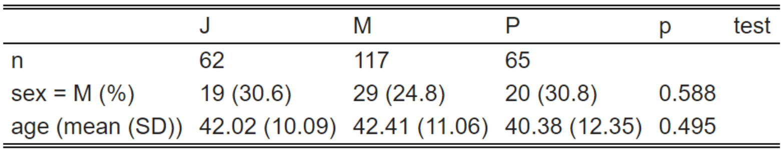
4.1.1. Population

The study population has enrolled at the three GP Centres between October 1st, 2018 and April 30th, 2019. The study population for statistical analysis was 244 patients, from which 68 (28%) were men and 176 (72%) were women. The mean age was 41.6 years (see Table 7).

The study population includes patients from three GP Centres:

a) Jürgenson PAK OÜ (J);
b) Merekivi PAK OÜ (M);
c) Pirita-Kose PAK OÜ (P).

Table 7 Summary of the study population.



	J	M	P	p	test
n	62	117	65		
sex = M (%)	19 (30.6)	29 (24.8)	20 (30.8)	0.588	
age (mean (SD))	42.02 (10.09)	42.41 (11.06)	40.38 (12.35)	0.495	

4.1.2. Timeframe

The full dataset includes PAK case summaries from the period 30/06/2004 to 02/08/2019. The ET case summaries are from the period 10/08/2014 to 27/08/2019 (see Figure 7). All ET and PAK database case summaries have diagnosis code present. The study time frame was used to filter the dataset, resulting in 6,599 observations from period October 1st, 2014 to April 30th,2019.

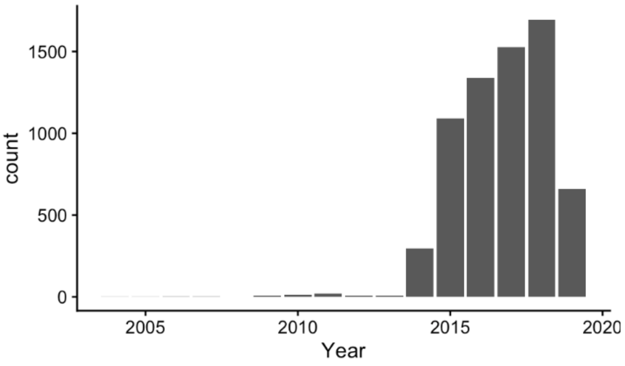


Figure 7 Number of observations per year.

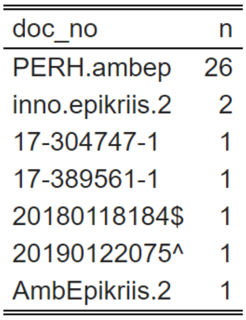
4.1.3. Non-Conforming Documents in EHIS

Documents in the ET database should carry unique identification numbers consisting of only numeric characters. However, some of the ET documents (n=32) merged with PAK data have non-conforming identification fields in ET (see Table 8). Examples of non-confirmed entries are 20180118184$; 75019371VIIM (see Table 9).

Table 8 Number of ET entries with the non-conformant document number.

GP Centres	Non-conforming	Total
P	6	990
J	5	930
M	22	2,141
Total	**32**	**4,061**

Table 9 Examples of non-conform doc_no.



doc_no	n
PERH.ambep	26
inno.epikriis.2	2
17-304747-1	1
17-389561-1	1
20180118184$	1
20190122075^	1
AmbEpikriis.2	1

A simple binomial model shows that all three GP Centres have a similarly small amount of mislabelled case summaries entered ET (see Figure 8).

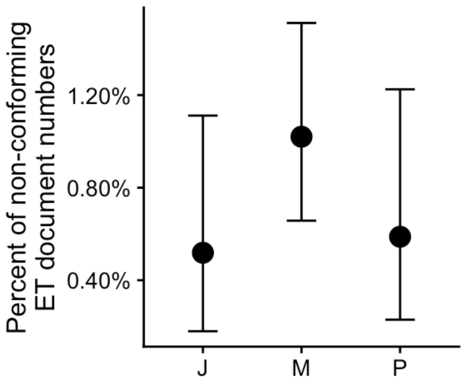


Figure 8 Fraction of non-conforming ET document id is in our raw dataset. (n | trials(total) ~ PAK, binomial response distribution). N = 3. Points denote the model best fit. Error bars denote 95% credible interval.

Considering the evidence, the non-conforming documents were filtered out from the dataset. Together, a final sample includes 4,061 PAK case summaries, from which 1,718 (42.3%) were submitted to ET. The final sample includes ICD-10 diagnoses for 647 unique diseases.

4.2.Data quality

4.2.1. Case summaries entered to Perearst2

The study sample in the PAK database had a total number of 4,061 case summaries entered. To understand if the total number of case summaries issued by GP Centres has remained the same during the study period, the fitted negative binomial model to the number of case summaries per month. First, the number of case summaries was naively modelled to PAK per month and not taking the number of patients into account (see Figure 9).

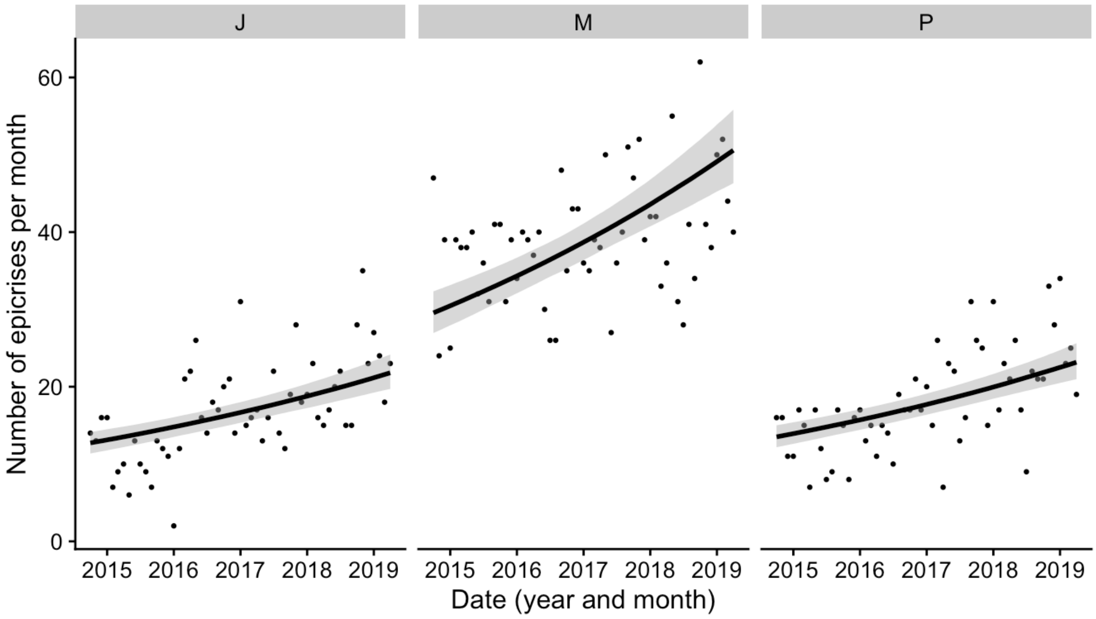


Figure 9 Increasing numbers of case summaries per month (negative binomial model, number of case summaries ~ year month + PAK). Points denote original data, N = 165. Line denotes model best fit, and a grey area denotes 95% credible region.

From examining the findings, it is essential to understand the number of case summaries entered per month, including patient exposure, as the total number of case summaries is dependent on the number of patients. The simple negative binomial model that takes the number of patients (exposure) in each PAK into account reveals that the number of case summaries issued by PAK has remained stable with a possible marginal increase during the study period (see Figure 10).

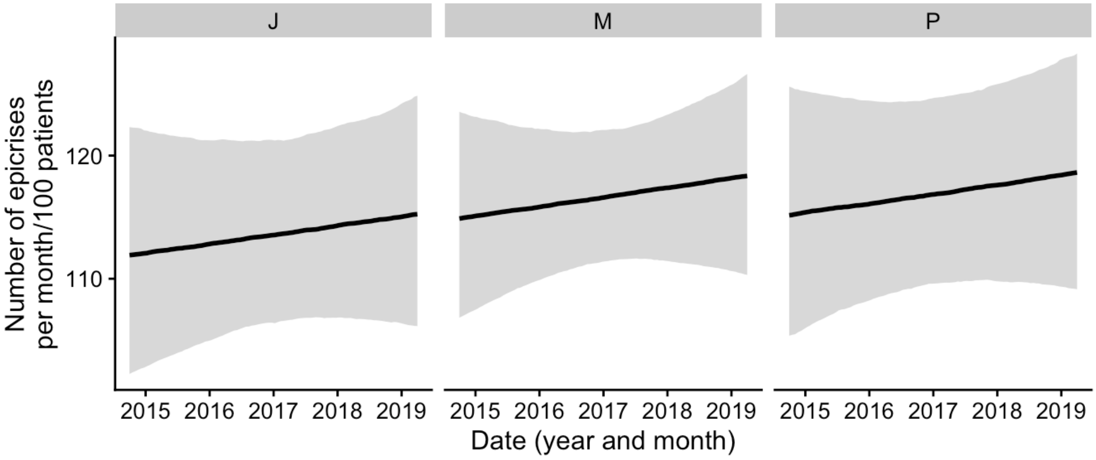
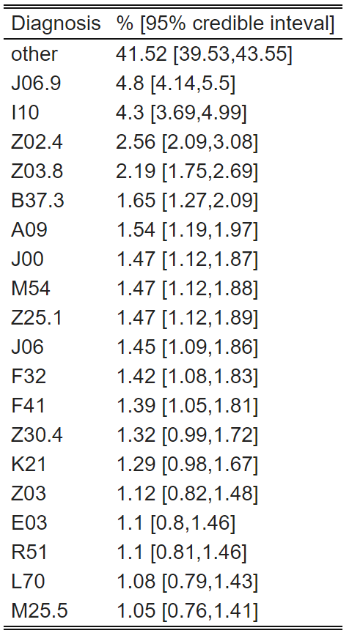


Figure 10 Number of case summaries issued per month per patient. The number of case summaries entered per month with patient exposure. (negative binomial model, number of case summaries ~ year month + PAK + offset (log(n patients))). Line denotes model best fit, N = 165. A grey area denotes a 95% credible region.

4.2.2. Diagnoses entered to Perearst2

The study sample in the PAK database had a total number of 647 unique diagnoses entered. For entered diagnosis analysis, only informative ICD-10 codes present more than once in all three PAK-s were analysed separately. All other unique or fewer common diagnoses were lumped to class “other”. The top 20 most frequent diagnoses (ICD-10 code) common to all three PAK are shown in Table 10 and 11.

Table 10 Top 20 most frequent diagnoses. Class “other” includes unique and less common diagnoses. Values are from binomial model (n | trials(total) ~ diagnosis + PAK).



Diagnosis	% [95% credible inteval]
other	41.52 [39.53,43.55]
J06.9	4.8 [4.14,5.5]
I10	4.3 [3.69,4.99]
Z02.4	2.56 [2.09,3.08]
Z03.8	2.19 [1.75,2.69]
B37.3	1.65 [1.27,2.09]
A09	1.54 [1.19,1.97]
J00	1.47 [1.12,1.87]
M54	1.47 [1.12,1.88]
Z25.1	1.47 [1.12,1.89]
J06	1.45 [1.09,1.86]
F32	1.42 [1.08,1.83]
F41	1.39 [1.05,1.81]
Z30.4	1.32 [0.99,1.72]
K21	1.29 [0.98,1.67]
Z03	1.12 [0.82,1.48]
E03	1.1 [0.8,1.46]
R51	1.1 [0.81,1.46]
L70	1.08 [0.79,1.43]
M25.5	1.05 [0.76,1.41]

Table 11 Description of top 20 diagnoses common at all three GP Centres. Values are rounded to two decimal places.

ICD-10	Diagnosis Codes	P	J	M	Mean
Others	Others	41,52%	41,51%	41,53%	41,52%
J06.9	Acute upper respiratory infections of multiple and unspecified sites	4,80%	4,80%	4,80%	4,80%
I10	Essential (primary) hypertension	4,30%	4,30%	4,30%	4,30%
Z02.4	Encounter for examination for driving license	2,55%	2,56%	2,56%	2,56%
Z03.8	Encounter for observation for other suspected diseases	2,19%	2,18%	2,19%	2,19%
B37.3	Candidiasis of vulva and vagina	1,65%	1,65%	1,65%	1,65%
A09	Infectious gastroenteritis and colitis	1,54%	1,54%	1,54%	1,54%
Z25.1	Need for immunization against influenza	1,47%	1,47%	1,47%	1,47%
J00	Acute nasopharyngitis	1,47%	1,47%	1,47%	1,47%
M54	Other dorsalgia	1,47%	1,47%	1,47%	1,47%
J06	Acute upper respiratory infections of multiple and unspecified sites	1,45%	1,45%	1,45%	1,45%
F32	Major depressive disorder	1,42%	1,42%	1,42%	1,42%
F41	Other anxiety disorders	1,39%	1,39%	1,39%	1,39%
Z30.4	Encounter for surveillance of contraceptives	1,32%	1,32%	1,32%	1,32%
K21	Gastro-esophageal reflux disease	1,30%	1,29%	1,29%	1,29%
Z03	Encounter for observation for other suspected conditions ruled out	1,12%	1,12%	1,12%	1,12%
R51	Headache	1,10%	1,10%	1,10%	1,10%
E03	Other hypothyroidism	1,10%	1,10%	1,10%	1,10%
L70	Acne	1,08%	1,08%	1,08%	1,08%
M25.5	Pain in joint	1,05%	1,05%	1,05%	1,05%
Z02.7	Encounter for issue of medical certificate	1,05%	1,05%	1,05%	1,05%
Z24.1	Need for immunization against viral encephalitis	1,03%	1,03%	1,03%	1,03%

Modelling the number of diagnoses using logistic regression reveals that common diagnoses are issued with similar proportions at three GP Centres (see Figure 11 and 12). Unique and less common diagnoses, assigned to class “other” (see Appendix 4), were issued at similar proportions (41.5%) at the three GP Centres.

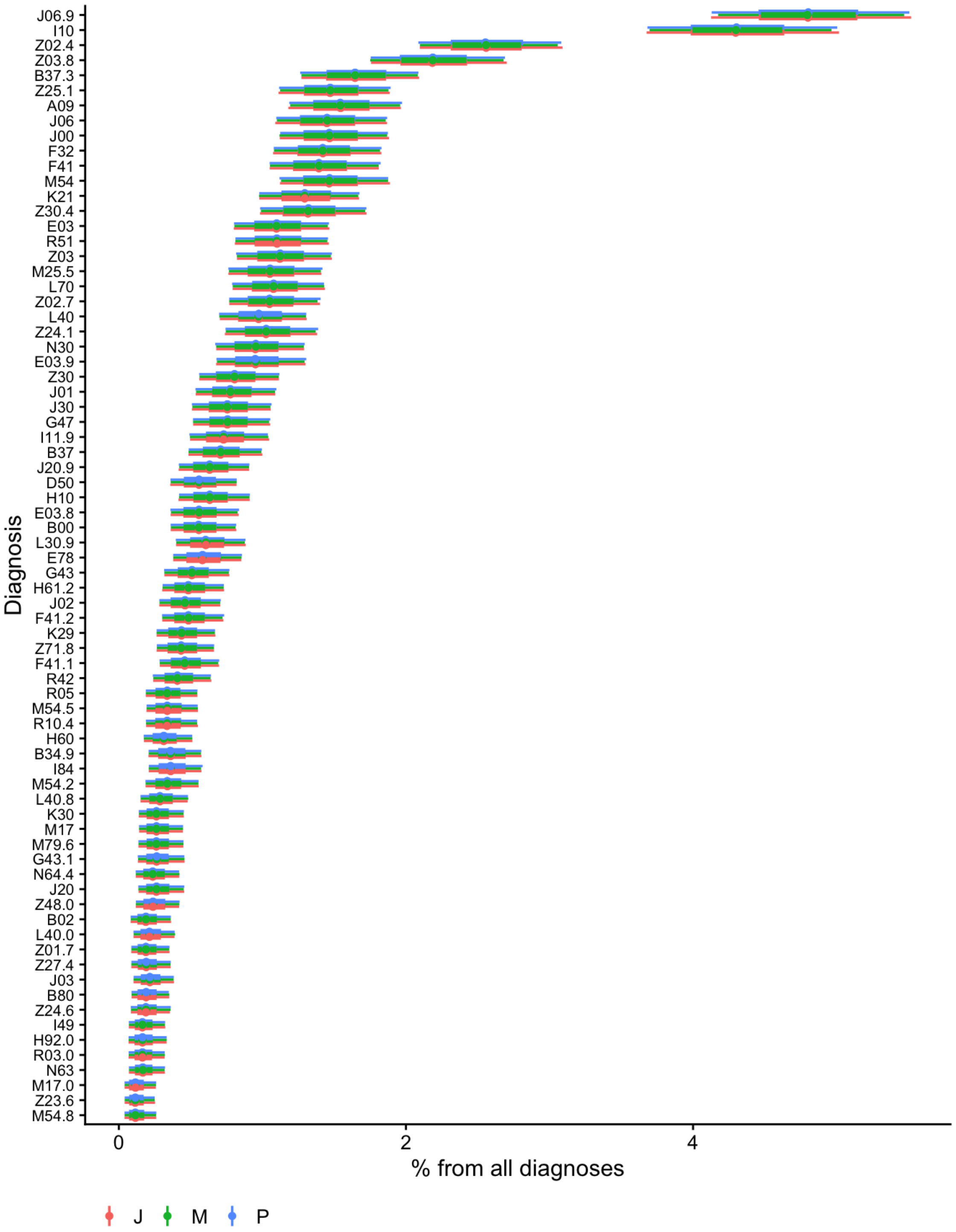


Figure 11 Common diagnoses are issued with similar proportions at three GP Centres. (n | trials(total) ~ diagnosis + PAK, binomial response distribution). N = 225. One diagnosis was randomly sampled from each case summary. Points denote the best fit of the model. Thick and thin error bars denote 67% and 95% credible interval, respectively.

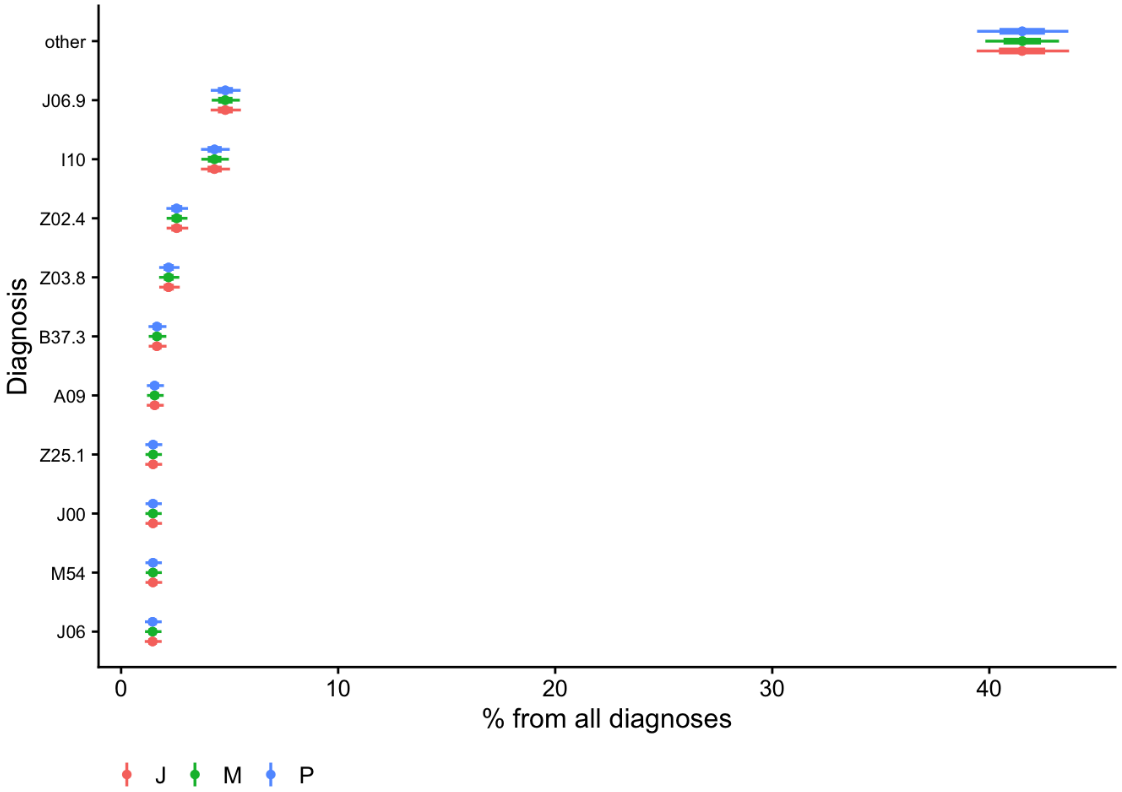


Figure 12 Top 10 common diagnoses versus unique “other. Simple logistic regression reveals that common diagnoses are issued with very similar proportions at three different PAK-s (n | trials(total) ~ diagnosis + PAK, binomial response distribution). N = 225. One diagnosis was randomly sampled from each case summary. Unique and fewer common diagnoses were assigned to class “other.” Points denote the model’s best fit. Thick and thin error bars denote 67% and 95% credible interval, respectively.

4.2.3. Case summaries proportion submitted to EHIS

The PAK sample includes 4,061 case summaries documents, from which 1,718 (42.3%) was submitted to EHIS during the study period. By analysing the mean proportion of the number of case summaries submitted to ET was 49% by GP Centres combined with a possible marginal increase during the study period (see Table 12).

Table 12 The proportion of case summaries submitted to EHIS by three GP Centres. Values are best-fit values from Bernoulli model fit.

PAK	2014	2015	2016	2017	2018	2019	Mean
P	51%	56%	58%	60%	52%	67%	57%
J	66%	63%	58%	61%	65%	63%	63%
M	24%	19%	26%	28%	31%	33%	27%
Total Mean	47%	46%	47%	50%	49%	55%	49%

The multi-level logistic model was used to understand the dynamics of submission status (submitted/non-submitted) in the proportion of case summaries submitted to EHIS. The findings were that the trend fluctuates yearly without a clear direction (see Figure 13).

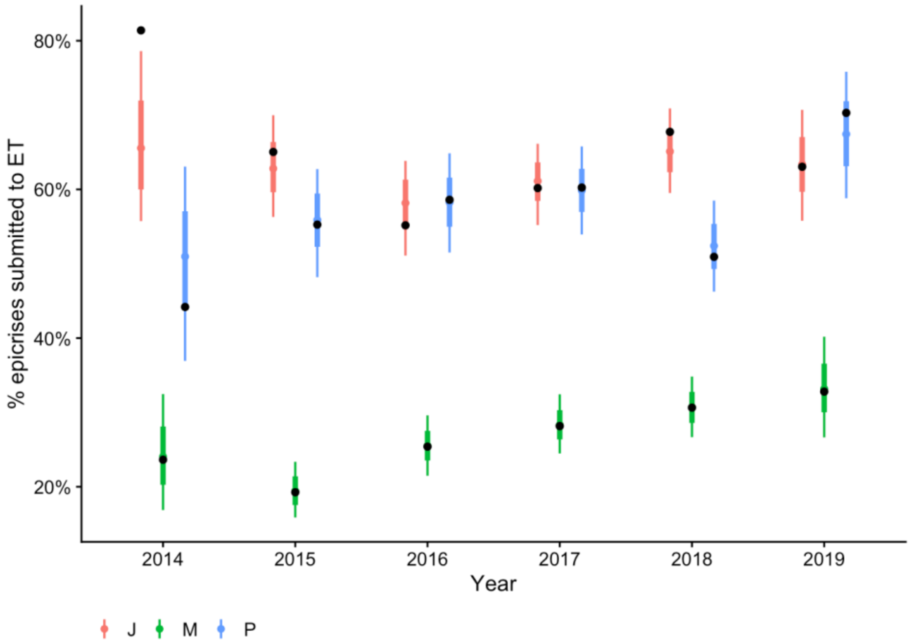


Figure 13 Documents per year, variable intercept for each PAK. The proportion of case summaries submitted to ET by three different PAK-s (submitted ~ asutus + year + (asutus | year), Bernoulli response distribution, N = 4,061). Black points denote the proportion calculated from original data. Coloured points denote the best fit of the model. Thick and thin error bars denote 67% and 95% credible interval, respectively.

The same data and model previously were also visualised as a line and ribbon graph (seeAppendix 6).

4.2.4. Submission of different diagnoses to EHIS

The PAK sample includes 647 unique ICD-10 diagnoses. The case summary submission model across diagnoses was created to understand the submission of different diagnoses to the ET database. The analysis shows a significant bias in diagnoses submission, e.g. *Hypothyroidism* 95% vs *Migraine with aura* 76% on average submitted to ET by the GP Centres (see Table 13/ Figure 14).

Table 13 The proportion of diagnoses submitted to EHIS compared to three GP Centres.

ICD-10	**Diagnosis Codes**	**P**	**J**	**M**	**Mean**
E03.9	Hypothyroidism	97%	98%	90%	95%
Z02.4	Encounter for examination for driving license	93%	93%	75%	87%
F41.2	Anxiety disorders	89%	90%	66%	81%
N63	Unspecified lump in breast	87%	89%	62%	80%
G43.1	Migraine with aura	85%	86%	57%	76%
I49	Cardiac arrhythmias	81%	83%	50%	71%
R10.4	Other and unspecified abdominal pain	80%	82%	49%	70%
J20	Acute bronchitis	77%	79%	45%	67%
N64.4	Mastodynia	77%	79%	45%	67%
F41.1	Generalized anxiety disorder	74%	76%	40%	63%
Z23	Encounter for immunization	73%	76%	40%	63%
M54.8	Other dorsalgia	73%	75%	39%	63%
Z03	Medical observation and evaluation for suspected diseases	73%	75%	39%	62%
J06	Acute upper respiratory infections of multiple sites	73%	75%	39%	62%
M79.6	Pain in limb, hand, foot, fingers and toes	72%	75%	39%	62%
Z03.8	Encounter for medical observation for suspected conditions ruled out	72%	75%	38%	62%
I10	Essential (primary) hypertension	72%	75%	38%	62%
Z71.8	Other specified counseling	70%	72%	36%	59%
K29	Acute haemorrhagic gastritis	70%	72%	36%	59%
M17	Osteoarthritis of knee	69%	72%	35%	58%
other	Other Diagnoses combined	59%	62%	26%	49%

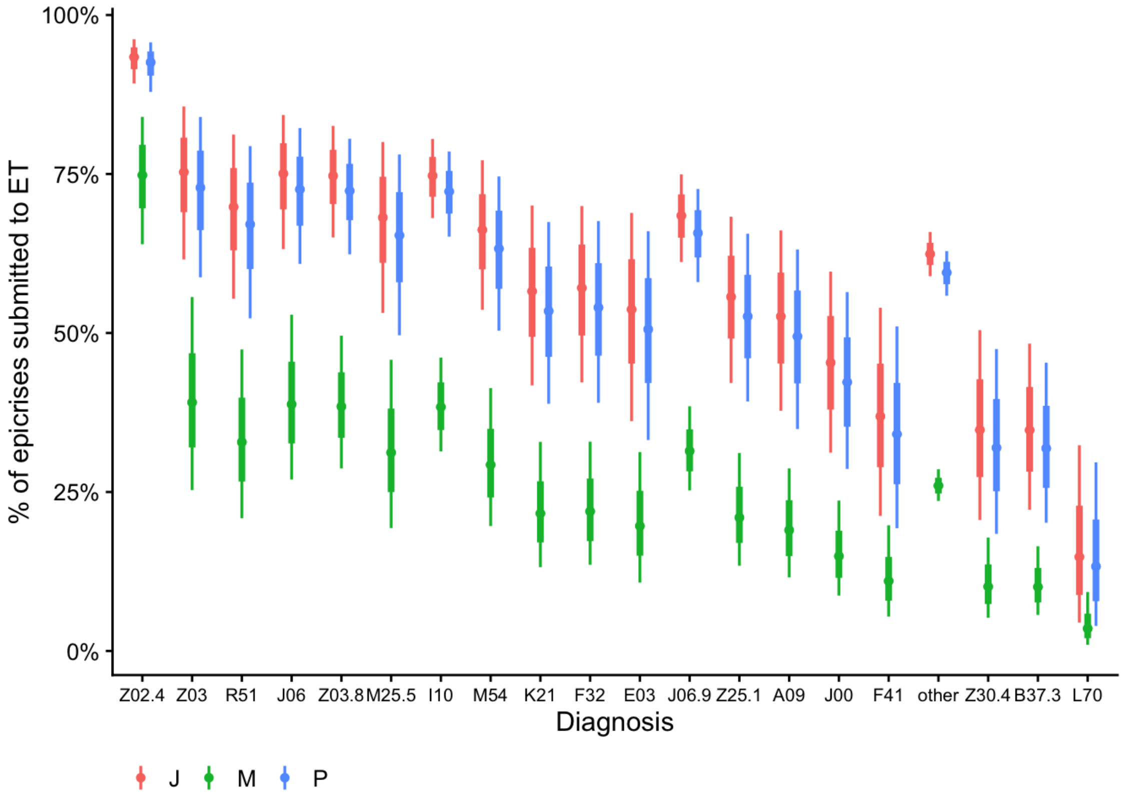


Figure 14 Proportion of diagnoses submitted to EHIS (submitted ~ asutus + (1 | diagnoos), Bernoulli response distribution. N = 4,061 Points denote model best fit.). The top twenty most frequent diagnoses to the proportion of diagnoses submitted to ET. Thick and thin lines denote 67% and 95% credible intervals, respectively.

The same data and model as previous were also visualised to include all diagnoses (see Appendix 5).

4.2.5. Time submission of case summaries to EHIS

The analysis of how quickly PAKs are submitting their case summaries to the ET database found significant differences in submission speed between GP Centres (see Table 14 Average time to submit case summaries to ET for GP Centres by years and days./ Figure 15). Whereas average submission time has significantly decreased during the period from 2014 to 2019, e.g. in Merekivi PAK OÜ, submission time has reduced from 889 to 73 days, in Jürgenson PAK OÜ from 196 to 16 days and in Pirita PAK OÜ from 226 to 19 days.

Table 14 Average time to submit case summaries to ET for GP Centres by years and days.

	2014	2015	2016	2017	2018	2019
PAK	Days	Days	Days	Days	Days	Days
P	226	137	83	50	31	19
J	196	119	72	44	27	16
M	889	539	327	198	120	73

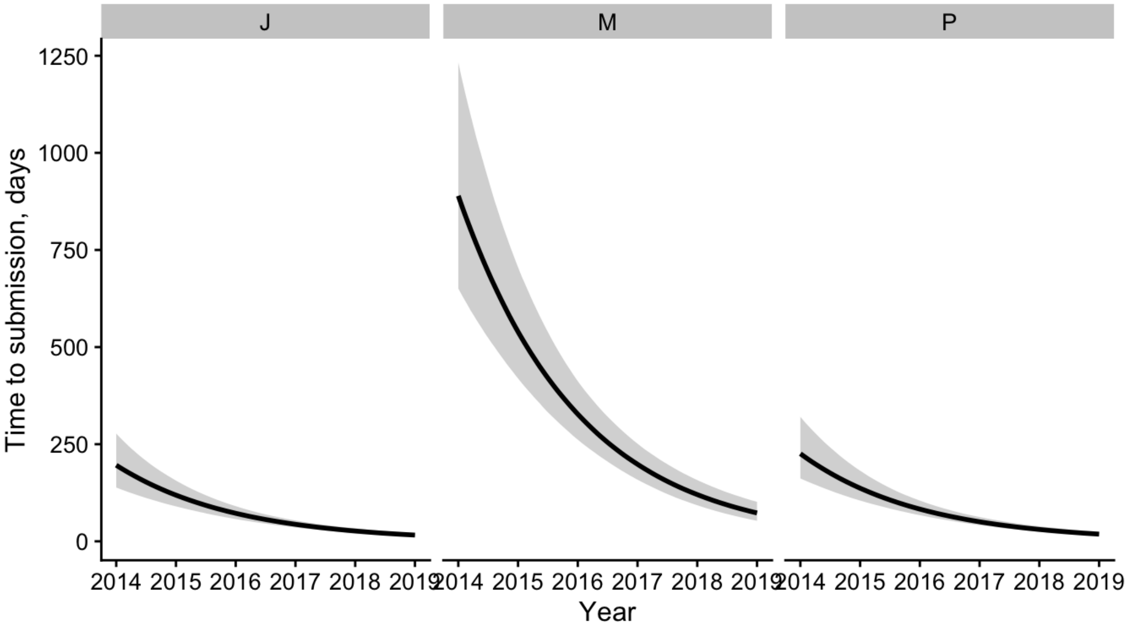


Figure 15 Time to submission in days of case summaries to ET (time ~ year + PAK) Weibull response distribution. N = 1,718. Lines denote model best fit. A grey area denotes a 95% credible region.

4.2.6. Time of diagnoses submission to EHIS

The analysis of how fast PAKs submit their different diagnoses to the ET database found significant differences in submission speed and bias between GP Centres. The analysis shows a significant bias in diagnoses submission speed, e.g. *Headache* 10 days (2019) vs *Cough* 15 days (2019) on average at GP Centres. The average submission speed of different diagnoses has significantly decreased from 2014 to 2019 (see Table 15). For visualisation Weibull response distribution graph has used in a submission time of diagnosis analysis. Twenty most common diagnoses shown in charts (see Figure 16).

Table 15 Top 20 diagnoses of time to submission of to EHIS in days.

			P		J		M		Mean	
	ICD-10		2014	2019	2014	2019	2014	2019	2014	2019
No	Diagnose	Diagnose	Days	Days	Days	Days	Days	Days	Days	Days
1	Z02.4	Examination for driving license	2	0	1	0	5	0	3	0
2	K21	Gastro-esophageal reflux disease	70	6	53	4	228	19	117	10
3	R51	Headache	76	6	58	5	249	20	128	10
4	K29	Acute haemorrhagic gastritis	100	8	76	6	326	27	167	14
5	N64.4	Mastodynia	102	8	77	6	331	27	170	14
6	Z71.8	Other specified counseling	109	9	83	7	355	29	182	15
7	Z24.1	Immunization against encephalitis	110	9	84	7	356	29	183	15
8	Z23.6	Immunization against diphtheria	110	9	83	7	359	29	184	15
9	R05	Cough	113	9	85	7	369	30	189	15
10	R42	Dizziness and Giddiness	116	10	89	7	380	31	195	16
11	H92.0	Otalgia	118	10	89	7	381	31	196	16
12	E78	Disorders of lipoprotein metabolism	119	10	90	7	386	31	198	16
13	L40.0	Psoriasis vulgaris	120	10	91	7	389	32	200	16
14	M54.5	Low back pain	120	10	91	7	390	32	200	16
15	N63	Unspecified lump in breast	121	10	92	7	396	32	203	16
16	H61.2	Impacted cerumen	122	10	93	8	396	32	204	17
17	F41.1	Generalized anxiety disorder	122	10	92	8	396	32	203	17
18	B02	Zoster	125	10	94	8	405	33	208	17
19	N30	Cystitis	125	10	96	8	410	33	210	17
20	H60	Otitis externa	129	10	97	8	418	34	214	17
33	other	Other Diagnoses combined	159	13	121	10	517	42	266	21

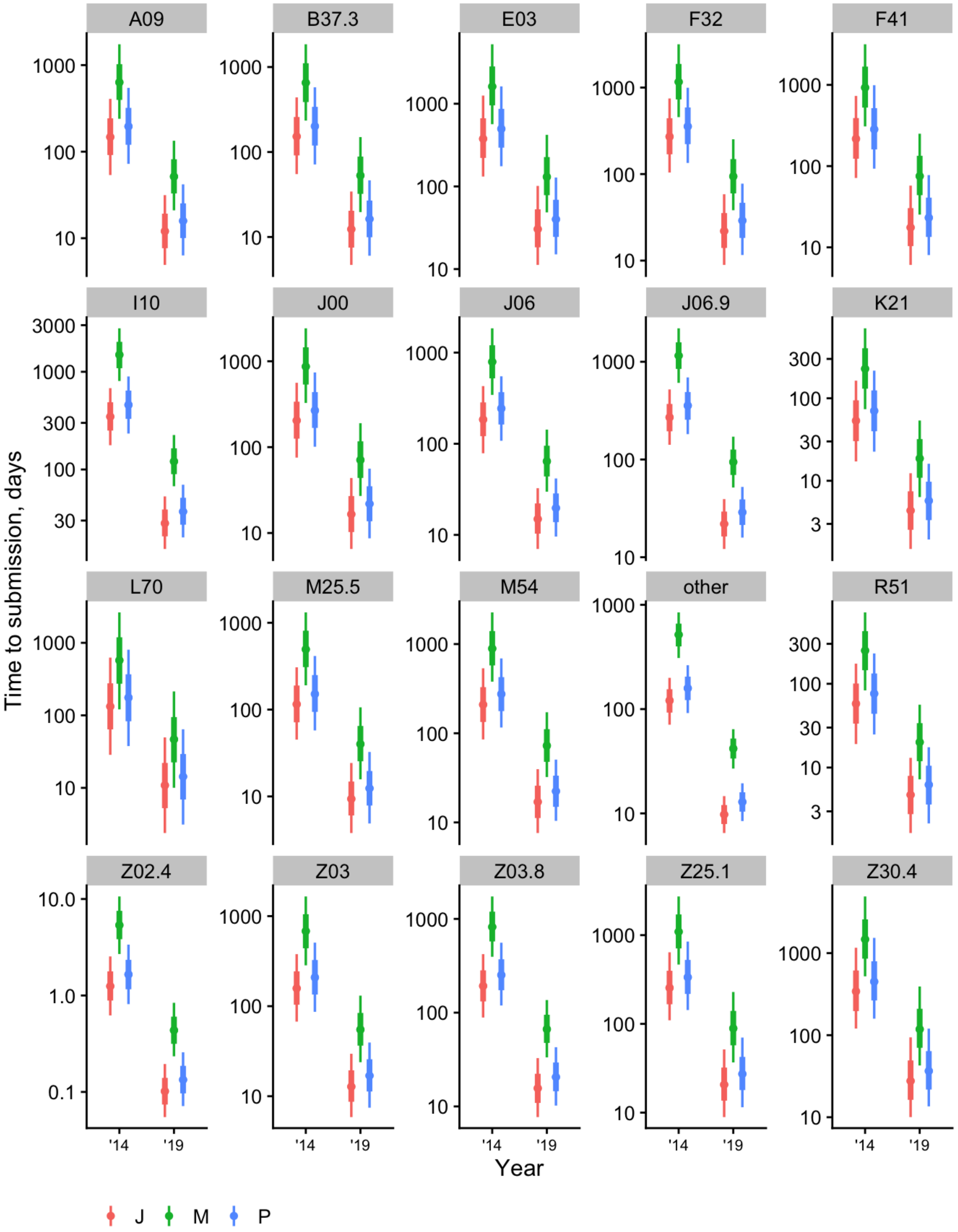


Figure 16 Time of submission diagnoses to ET (time ~ year + PAK + diagnosis + (1 | diagnosis) + (1 | PAK) + (1 | year), Weibull response distribution. N = 1,718. Facet label denotes diagnoses. Points denote model best fit. Thick and thin lines denote credible interval at 67% and 95% levels.

5. Discussion

5.1. Extracting and acquiring data

The data extraction processes described in the current study required querying, linking and merging EHRs datasets for research use in the secondary analysis. The method to acquire data from the Perearst2 (Query#1) EHR database was straight forward processing with query and functionality of extraction developed in the GP Centres information system software. However, to acquire data from the EHIS (Query#2) database with a direct query method with GP Centres software, Perearst2 was partly implemented (see Ch. 3.1.3). Although the export data functionality of Query#2 was previously agreed to be implemented by Medisoft in Perearst2 software, the developer was not able to complete it in time regarding this thesis. The only solution to retrieving the data from Query#2 was taking screenshots and extracting data from pictures. This process proved to be lengthy (the data acquisition process took ~7 months from July 2019 till January 2020) and required a large amount of manual labour.

For querying the EHIS data, the two alternative methods were also evaluated:

- Firstly, the EHRs data could have been extracted by TEHIK from the EHIS analysis module database with a simple direct query. However, this method lacks a standardised process to link patient data originating from a different organisation and connecting them into one database. Although, in theory, if both organisations have the same unique identifier for de-personalisation for merging the databases, then the data could be prepared for and handled in the secondary analysis.
- Secondly, the EHRs data could have been extracted in GP Centres with custom made software for direct query from the EHIS database. However, this method required overcoming specific obstacles and need various additional resources before the data analysis, e.g., finding a software developer, accessing the EHIS database through the MISP2 standard X-Road component that GP Centres use.

Another aspect in acquiring data with the current method was the limited sample size. Only three GP Centres agreed to be in the patient enrolment phase that preceded secondary analysis. It limited sources for data research in a secondary analysis. The broader selection of GP Centres would give better odds that the results observed are not just a chance result.

The empirical findings emerged from the short discussion above that combining data from multiple EHRs databases that required matching (linking) related records and safeguarding that data refers to the correct patient proved complicated for secondary analysis. In 2012, Great Britain successfully instituted the Clinical Practice Research Datalink service that promotes healthcare research and drives innovation through EHRs. The institution provides anonymised EHRs to researchers within academic, regulatory, and pharmaceutical organisations worldwide to support observational public health research and health services planning can be made query ready and de-identified [43]. In 2015, Zozus et al. described the process and highlighted the need to standardise the extraction process (methods) using EHR data for research [30]. Therefore, Estonia should consider implementing a data warehouse that would collect anonymised patient data from the network of GP centres.

5.2. Measuring the data quality

5.2.1. Perearst2 database composition

The measurements of the Perearst2 database composition for case summaries and diagnoses showed that the total number of case summaries entered by the three GP Centres in the Perearst2 database was 4,061 during the retrospective period of October 1st, 2014 till April 30th,2019. The total number of case summaries issued by GP Centres should remain equal during the study period. For the verification, the documents were fitted to the simple negative binomial model (see Ch. 4.1). The model considers the number of patients exposure in each GP Centres and reveals that the number of case summaries issued by healthcare workers has remained stable with a possible marginal increase during the study period (see Figure 10). It does seem to link to better documentation habits, increased illnesses of ageing (41.6 years), or affected by the enrolment of the study population.

Evaluation of different 647 diagnoses entered the Perearst2 database revealed that common diagnoses were entered with similar frequency at three different GP Centres (see Table 11). The most frequent diagnoses were J06.9 – *Acute upper respiratory infection* by 4,8% [4.14-5.5; 95% credible interval] and I10 – *Essential Hypertension* by 4.3%

[3.69-5.5; 95% credible interval] (see Table 10/ Figure 12). Overall, these findings agree with findings in 2018 reported by Finley found similar diagnostic consistency in primary care physician disease coding [44].

5.2.2. Data completeness

The current thesis first research question was about EHIS data completeness - whether the GP Centres case summaries are satisfactorily presented in the EHIS database? The overall submission of case summaries to the EHIS database by GP Centres shows that the Perarst2 sample includes 4,061 case summaries documents, from which 2,343 (57.7%) were not submitted, and 1,718 (42.3%) were submitted to EHIS. The case summaries submission of 42.3% is higher than the previous study conducted by TAI in 2017, resulting in 22% [4]. However, when analysing yearly trends in the submission of case summaries, it is feasible to observe transmitting volatility and difference between GP Centres. In 2019, there were significant differences in database completeness between GP Centres with the case summaries submissions of Merekivi PAK OÜ 32%, Jürgenson PAK OÜ 75% and Pirita PAK OÜ 69% (see Table 12/ Figure 17).

Based on the results of this study, Merekivi PAK OÜ was unable to reach this goal set by family doctors' quality-indicator system. In 2020, the EHIF/ Estonian Society of Family Doctors in the family doctor quality system indicated objective is set at 50% for case summaries submission to EHIS in a year [45]. Jürgenson PAK OÜ and Pirita PAK OÜ reached the goal set in 2020 sufficiently. In the author's opinion, the probable reason behind this is the lack of automatisation in completed case summaries submission in information software Perearst2 and general attitudes of healthcare workers towards the EHIF and EHIS processes.

The proportion of diagnoses in documents submitted to EHIS by three GP Centres reveals significant bias, e.g. *Hypothyroidism* was sent in an average of 95% compared to *Migraine with aura* in an average of 76% (see Table 13/ Figure 14). The result of prejudice in sent indications is interesting as it shows the subjectivity of healthcare workers when deciding over pathways of care processes. Although it is challenging to explain diagnoses submission results within the context of data quality, the significant bias between submitting one diagnosis more than another by healthcare worker can be attributable to attitude towards diseases and is an issue for future research to explore.

5.2.3. Data timeliness

The current thesis second research question was about how fast case summaries were submitted to EHIS by GP Centres during the observed period. In the evaluation of year-to-year transmission delay in days, there is a significant reduction in submission time. A positive submission trend is an essential finding in understanding the EHIS database's timeliness from year to year. The trend is similar for all GP Centres (see Table 15/ Figure 15). However, the result in 2019 still misses the legal compliance. The statutes of EHIS require sending of the finalised case summaries by the following working day after case completion [46]. Merekivi PAK OÜ submitted the documents on average in 73 days, Jürgenson PAK OÜ in 16 days, and Pirita PAK OÜ in 19 days. The study findings of the year-to-year trend of improved submission to the EHIS database resulted in improved timeliness of data quality. However, submitting the closed case summaries by healthcare workers takes longer than permitted by law. It appears to be a case of the lack of software automatisation in submission to EHIS. Also, healthcare workers are more focused to fulfil the requirements of reimbursement processes rather than health data sharing. The reasons for delayed document submission could be a topic of further research.

The time of submission of diagnoses to EHIS shows a progressive reduction across three GP Centres. It is important to note that the timestamp of diagnosis is the same as closed case summaries in GP Centers Perearst2 software. The result shows that there is a submission bias by diagnoses. For example, an *Examination for a driving license* (Z02/ ICD-10 Code) is sent on the same day compared to *Cystitis* (N30/ ICD-10 Code), which was sent in an average of 17. It is possible to observe significant delay differences in sending diagnose between the GP Centres. It is challenging to explain diagnoses sending delays within the context of data quality, but the significant bias between submitting one diagnosis before another can be attributable to a different attitude towards diseases and is an issue for future research to explore.

5.3. Conclusions

The study aimed to compare the EHIS and Perearst2 data quality. The author hypothesised that lack of submission and the untimely availability of the case summaries in the EHIS database would result in incorrect clinical decisions for DSS algorithms. The limited amount in case summaries sent to the EHIS restricts the data available. The timebound trend of inclination towards important diagnoses submitted sets a limit to the accuracy of DSS algorithms.

The findings in completeness indicated that 57.7% of case summaries from GP Centres were not submitted to the EHIS database, although sending trend line of documents improved year-to-year. The results about timeliness showed in 2019 that GP Centres delay sending the case summaries to the EHIS. Although, the trend line of sending delays improved in a year-to-year comparison. Overall completeness and timeliness findings in the EHIS correlate with 2015 results reported by TAI [4].

The study results support the hypothesis that the submission of the case summaries and their timely availability in the EHIS is low. The DSS algorithms highly probably result in incorrect clinical decisions due to the unavailability of data. Ross and Metsallik reached a similar conclusion. Their results suggested that the data quality in the databases affects the reliability of DSS results in real-time healthcare settings [14]. Linking and combining Perarst2 and EHIS databases would help to reduce data errors and increase the trustworthiness of DSS. Also, in 2017 Wagner *et al.* affirmed that the information systems where healthcare data is available to the right people at the right place and at the right time with efficient data flow are vital in establishing secondary data use in decision support software [2].

However, the year-to-year trendline in data completeness and timeliness shows improvement during the trial period of October 1st,2014, till April 30th,2019. In the future, it is crucial to monitor the trend of data quality. The GP Centres and TEHIK must improve their information systems by introducing better error checking, upgrade to structured data entry, and use automatisation for accuracy and consistency. The same was concluded by Davoudi *et al.* and TAI [3] [4] [12].

5.4.Limitations

The study population was relatively small due to the enrolment policy. The study followed a previous study, where the HIV indicator condition disease algorithm of

Diagnostic Match DSS triggered the inclusion of subjects. The data extraction method from the EHIS database resulted in unusual formats, e.g. screenshots that made the data extraction susceptible for errors (see Ch. 3.1.3). The data available from GP Centres was too limited to make conclusive statistical analysis on overall regional or primary healthcare level on data quality. The selected GP Centres were relatively skilled to work with their EMR. It may introduce a positive bias in working with electronic information, which exceeds the average level of the primary health care system.

6. Summary

The thesis aimed to test whether it is possible to utilise Diagnostic Match DSS on the EHIS data. The study linked the data from two sources to identify the data completeness and timeliness.

The study provides evidence that the number of case summaries sent to the EHIS database is limited. It restricts the data available. The study also found that there is a trend of inclination towards submitting only documents with favourable diagnoses. Limited availability of diagnosis data hinders the accuracy of the DSS algorithms. Therefore, utilising Diagnostic Match DSS only on the EHIS data is not sufficient. It can result in partial or delayed effects. The implementation of the DSS should combine both EHIS and Perearst2 data to achieve decent data completeness and timeliness.

In addition, the findings of the study provide additional information about combining data from multiple databases. The matching (linking) related records and safeguarding that data refers to the correct patient proved complicated. Therefore, Estonia should consider a specialised data warehouse that actively collects patient data from a healthcare network for future research.

6.1.Future research

Further research needs to study the EHR data quality because of the limitations identified in the thesis. In addition, future research needs to identify the reasons for biases, trends, and variation of the submission in case summaries and diagnoses at the individual and organisational level.

Acknowledgement

The author would like to thank the supervisors and reviewers for the help and constructive feedback—also the massive support of GP Centres to provide the data and resources.

References

[1] M. A. Makary, “Medical error—the third leading cause of death in the US,” *BMJ,* vol. i, no. 2139, p. 353, 2016.

[2] J. P. G. a. F. W. L. Karen A. Wager, Health Care Information Systems : A Practical Approach for Health Care Management 4th edition, South Carolina: John Wiley & Sons, Incorporated, 2017.

[3] D. J. G. B. J. T. K. L. O. S. R. K. W. A. Davoudi S, “Data Quality Management Model (2015 Update),” AHIMA, 2015.

[4] L. P. Eva Anderson, “Family doctor’s offices’ outpatient consultations data in the e-health system, 2015,” National Institute for Health Development and Department Health Statistics, Tallinn, 2017.

[5] Health Board, “HIV Statistika 2019,” Health Board, 29 04 2020. [Online]. Available: https://www.terviseamet.ee/sites/default/files/Nakkushaigused/Haigestumine/nakkush_statistika/hiv_statistika_2019_maakond.pdf. [Accessed 29 04 2020].

[6] G. Kikas, *Enhancing HIV Indicator Disease-Guided Testing Strategy Implementation in Estonia by Using HIV Clinical Decision Support System - a Pilot Study in Primary Care,* Tallinn, 2019.

[7] Eesti Haigekassa, “The EHIF brings a system of support for clinical decisions to family doctors,” 09 05 2019. [Online]. Available: https://www.haigekassa.ee/uudised/haigekassa-toob-perearstideni-kliiniliste-otsuste-tugisusteemi. [Accessed 02 05 2020].

[8] L. Bai, “A data quality framework, method and tools for managing data quality in a health care setting: an action case study,” *Journal of Decision Systems,* vol. 1, no. 27, pp. 144-154, 2018.

[9] N. G. Weiskopf, “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research,” *J Am Med Inform Assoc,* vol. 20, no. 144–151, 2013.

[10] Kitty S. Chan, “Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature,” *Medical Care Research and Review,* vol. 67, no. 5, p. 503–527, 2010.

[11] K. Bruce Bayley, “Challenges in Using Electronic Health Record Data for CER,” *Medical Care,* vol. 51, no. 8, p. S80–S86, 2013.

[12] M. I. Mare Ruuge, “Statsionaarsete ja päevaravi epikriiside saatmise aeg tervise infosüsteemi 2015. aastal,” Tervise Arengu Instituut, Tallinn, 2016.

[13] P. Ross, “Feasibility study for the development of digital decision support systems for personalised medicine,” Ministry of Social Affairs, Tallinn, 2015.

[14] P. Ross, “Eelanalüüs personaalmeditsiini otsustustoe hanke ettevalmistamiseks,” Tallinna Tehnikaülikool, Tallinn, 2017.

[15] J. Metsallik, *First usage of healthcare information systems in Estonian,* Tallinn: CEUR-WS.org, 2018.

[16] M. Tiik, *Access Rights and Organizational Management in Implementation of Estonian Electronic Health Record System,* Tallinn: TALLINN UNIVERSITY OF TECHNOLOGY, 2012.

[17] P. Kruus, “DEVELOPING AN EVALUATION FRAMEWORK FOR THE COUNTRY-WIDE ELECTRONIC PRESCRIBING SYSTEM IN ESTONIA,” TALLINN UNIVERSITY OF TECHNOLOGY, Tallinn, 2013.

[18] PRAXIS, “Assessing the Economic Impact/Net Benefi ts of the Estonian Electronic Health Record System,” PRAXIS, Tallinn, 2010.

[19] STACC, “Description of the current status and future needs of the Information Architecture and Data Management solutions for the national personalised medicine pilot project,” STACC, Tartu, 2015.

[20] Riigikogu, “Tervishoiuteenuste korraldamise seadus,” 17 05 2020. [Online]. Available: https://www.riigiteataja.ee/akt/110032011009?leiaKehtiv. [Accessed 07 02 2021].

[21] Sotsiaalminister, “Tervise infosüsteemi edastatavate dokumentide andmekoosseisud ning nende esitamise tingimused ja kord,” Riigi Teataja information system, Tallinn, 2019.

[22] P. Dr. Ross, “Feasibility study for the development of digital decision support systems for personalised medicine,” Ministry of Social Affairs, Tallinn, 2015.

[23] Ministry of Social Affairs, “Estonian eHealth Strategic Development Plan 2020,” ry of Social Affairs, Tallinn, 2015.

[24] Medisoft, “PEREARST2/ ERIARST,” Medisoft, 17 10 2017. [Online]. Available: https://medisoft.ee/tooted/perearst2-eriarst/. [Accessed 15 04 2019].

[25] E. Vanker, “EVALUATION OF ESTONIAN PRIMARY HEALTHCARE ELECTRONIC HEALTH RECORDS USABILITY,” TALLINN UNIVERSITY OF TECHNOLOGY, Tallinn, 2014.

[26] Sotsiaalministeerium, “HIV-nakkuse testimise ja HIV-positiivsete isikute ravile suunamise tegevusjuhis,” Sotsiaalministeerium, Tallinn, 2012.

[27] HIV-Europa, “HIV Indicator Conditions: Guidance for Implementing HIV Testing in Adults in Health Care Settings,” HIV Europa,, Copenhagen, 2016.

[28] ECDC, “HIV testing in Europe,” European Centre for Disease Prevention and Control, Stockholm, 2016.

[29] B. L. Ryan, “Methods to describe referral patterns in a Canadian primary care electronic medical record database: modelling multi-level count data,” *J Innov Health Inform,* p. 311–316, 24 4 2017.

[30] P. Meredith Nahm Zozus, "Acquiring and Using Electronic Health Record Data," *NIH Collaboratory Electronic Health Records,* vol. 1, no. 1, pp. 1-25, 2015.

[31] Gerli Kriiska, “Tervise infosüsteemi standard: Dokumendi väljavõtte päring (diagnoosid) Digilugu,” TEHIK, TALLINN, 2007.

[32] R Core Team, “R: A Language and Environment for Statistical Computing,” R Core Team, 10 10 2020. [Online]. Available: https://www.r-project.org/. [Accessed 11 11 2020].

[33] H. Wickham, “Welcome to the Tidyverse,” *Journal of Open Source Software,,* vol. 43, no. 1686, p. 4, 2019.

[34] J. B. Hadley Wickham, “readxl: Read Excel Files,” 13 03 2019. [Online]. Available: https://cran.r-project.org/web/packages/readxl/index.html. [Accessed 11 11 2020].

[35] Stan Development Team, “Stan Modeling Language Users Guide and Reference Manual,” 11 11 2020. [Online]. Available: https://mc-stan.org/users/documentation/. [Accessed 11 11 2020].

[36] P.-C. Bürkner, “brms: An R Package for Bayesian Multilevel Models Using Stan,” vol. 80, no. 1, pp. 1-23, 2017.

[37] K. M, “tidybayes: Tidy Data and Geoms for Bayesian Models,” R package version 2.3.0, 1 11 2020. [Online]. Available: http://mjskay.github.io/tidybayes/. [Accessed 11 11 2020].

[38] H. Wickham, “Modelling Functions that Work with the Pipe,” 19 05 2020. [Online]. Available: https://modelr.tidyverse.org, https://github.com/tidyverse/modelr. [Accessed 11 11 2020].

[39] H. Wickham, “Elegant Graphics for Data Analysis,” 01 01 2016. [Online]. Available: https://ggplot2.tidyverse.org.. [Accessed 11 11 2020].

[40] C. O. Wilke, “Cowplot: Streamlined Plot Theme and Plot Annotations for ’Ggplot2,” 30 12 2020. [Online]. Available: https://cran.r-project.org/web/packages/cowplot/index.html. [Accessed 07 02 2021].

[41] J. Hester, “glue: Interpreted String Literals,” 27 08 2020. [Online]. Available: https://cran.r-project.org/web/packages/glue/index.html. [Accessed 11 11 2020].

[42] K. Müller, “here: A Simpler Way to Find Your Files,” 13 12 2020. [Online]. Available: https://github.com/r-lib/here. [Accessed 13 12 2020].

[43] S. Padmanabhan, “Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications,” *European Journal of Epidemiology,* vol. 4, no. 34, p. 91–99, 2019.

[44] M. Caitlin R. Finley, “What are the most common conditions in primary care?,” *Canadian Family Physician,* vol. 64, no. 1, pp. 832-841, 2018.

[45] ESFD/ EHSF, “Family doctor quality system 2020,” ESFD/ EHSF, 01 01 2021. [Online]. Available: https://www.perearstiselts.ee/images/2021AASTA_AUDITEERIMINE.pdf. [Accessed 21 03 2021].

[46] Sotsiaalminister, “Tervishoiuteenuse osutamise dokumenteerimise tingimused ja kord,” 26 11 2020. [Online]. Available: https://www.riigiteataja.ee/akt/122062016040?leiaKehtiv. [Accessed 07 02 2021].

Appendixes

Appendix 1 Non-exclusive Licence for Publication and Reproduction of Graduation Thesis

I, Kristjan Krass (date of birth: 17/10/1975.), grant Tallinn University of Technology a free licence (non-exclusive licence) for my thesis DATA QUALITY IN HEALTH INFORMATION SYSTEMS – COMPLETENESS AND TIMELINESS supervised by Janek Metsallik to be reproduced for the purposes of preservation and electronic publication, incl. to be entered in the digital collection of TUT library until the expiry of the term of copyright; published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of TTÜ library until the expiry of the term of copyright.

I am aware that the author also retains the rights specified in clause 1.

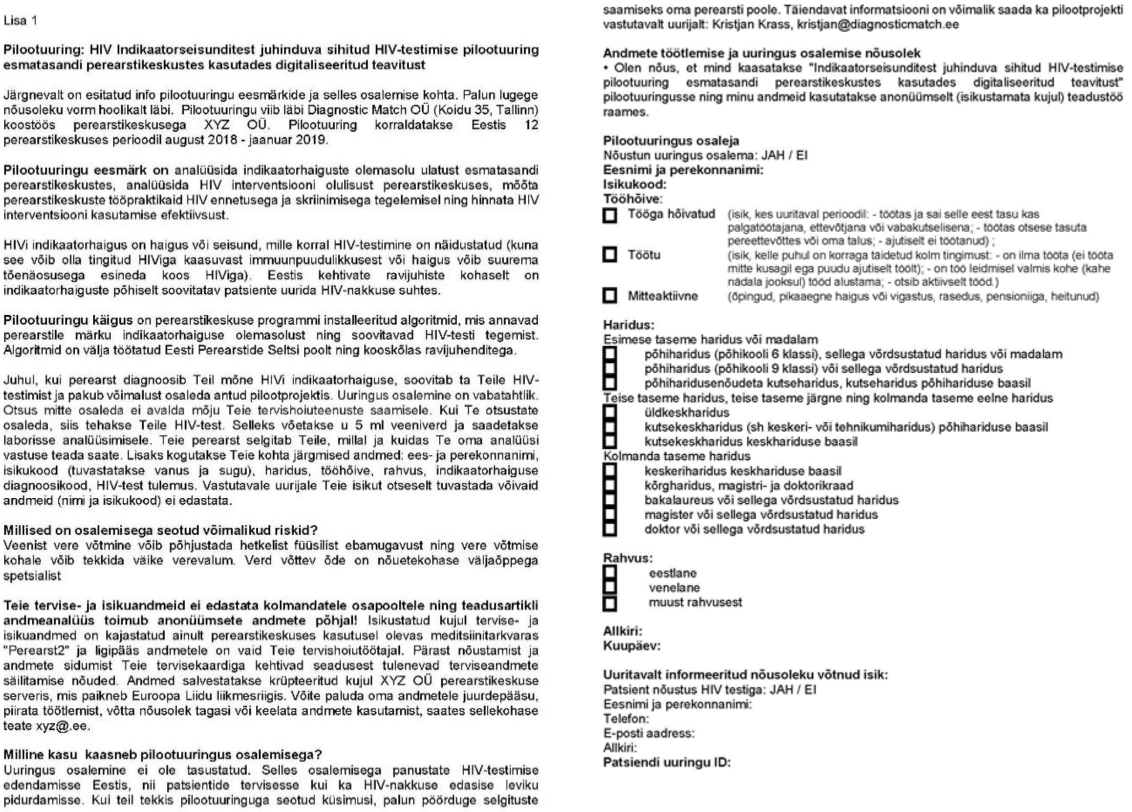
I confirm that granting the non-exclusive licence does not infringe third persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

(signature)

April 30th,2021

(date)

Appendix 2. Informed Consent Form (ICF). Tallinn Medical Research Ethics Committee (decision No. 2324, 17.05.2018)



Lisa 1

Pilootuuring: HIV Indikaatorseisunditest juhinduva sihitud HIV-testimise pilootuuring esmatasandi perearstikeskustes kasutades digitaliseeritud teavitust

Järgnevalt on esitatud info pilootuuringu eesmärkide ja selles osalemise kohta. Palun lugege nõusoleku vorm hoolikalt läbi. Pilootuuringu viib läbi Diagnostic Match OÜ (Koidu 35, Tallinn) koostöös perearstikeskusega XYZ OÜ. Pilootuuring korraldatakse Eestis 12 perearstikeskuses perioodil august 2018 - jaanuar 2019.

Pilootuuringu eesmärk on analüüsida indikaatorhaiguste olemasolu ulatust esmatasandi perearstikeskustes, analüüsida HIV interventsiooni olulisust perearstikeskuses, mõõta perearstikeskuste tööpraktikaid HIV ennetusega ja skriinimisega tegelemisel ning hinnata HIV interventsiooni kasutamise efektiivsust.

HIVi indikaatorhaigus on haigus või seisund, mille korral HIV-testimine on näidustatud (kuna see võib olla tingitud HIViga kaasuvast immuunpuudulikkusest või haigus võib suurema tõenäosusega esineda koos HIViga). Eestis kehtivate ravijuhiste kohaselt on indikaatorhaiguste põhiselt soovitatav patsiente uurida HIV-nakkuse suhtes.

Pilootuuringu käigus on perearstikeskuse programmi installeeritud algoritmid, mis annavad perearstile märku indikaatorhaiguse olemasolust ning soovitavad HIV-testi tegemist. Algoritmid on välja töötatud Eesti Perearstide Seltsi poolt ning kooskõlas ravijuhenditega.

Juhul, kui perearst diagnoosib Teil mõne HIVi indikaatorhaiguse, soovitab ta Teile HIV-testimist ja pakub võimalust osaleda antud pilootprojektis. Uuringus osalemine on vabatahtlik. Otsus mitte osaleda ei avalda mõju Teie tervishoiuteenuste saamisele. Kui Te otsustate osaleda, siis tehakse Teile HIV-test. Selleks võetakse u 5 ml veeniverd ja saadetakse laborisse analüüsimisele. Teie perearst selgitab Teile, millal ja kuidas Te oma analüüsi vastuse teada saate. Lisaks kogutakse Teie kohta järgmised andmed: ees- ja perekonnanimi, isikukood (tuvastatakse vanus ja sugu), haridus, tööhõive, rahvus, indikaatorhaiguse diagnoosikood, HIV-test tulemus. Vastutavale uurijale Teie isikut otseselt tuvastada võivaid andmeid (nimi ja isikukood) ei edastata.

Millised on osalemisega seotud võimalikud riskid?
Veenist vere võtmine võib põhjustada hetkelist füüsilist ebamugavust ning vere võtmise kohale võib tekkida väike verevalum. Verd võttev õde on nõuetekohase väljaõppega spetsialist

Teie tervise- ja isikuandmeid ei edastata kolmandatele osapooltele ning teadusartikli andmeanalüüs toimub anonüümsete andmete põhjal! Isikustatud kujul tervise- ja isikuandmed on kajastatud ainult perearstikeskuses kasutusel olevas meditsiinitarkvaras "Perearst2" ja ligipääs andmetele on vaid Teie tervishoiutöötajal. Pärast nõustamist ja andmete sidumist Teie tervisekaardiga kehtivad seadusest tulenevad terviseandmete säilitamise nõuded. Andmed salvestatakse krüpteeritud kujul XYZ OÜ perearstikeskuse serveris, mis paikneb Euroopa Liidu liikmesriigis. Võite paluda oma andmetele juurdepääsu, piirata töötlemist, võtta nõusolek tagasi või keelata andmete kasutamist, saates sellekohase teate xyz@.ee.

Milline kasu kaasneb pilootuuringus osalemisega?
Uuringus osalemine ei ole tasustatud. Selles osalemisega panustate HIV-testimise edendamisse Eestis, nii patsientide tervisesse kui ka HIV-nakkuse edasise leviku pidurdamisse. Kui teil tekkis pilootuuringuga seotud küsimusi, palun pöörduge selgituste saamiseks oma perearsti poole. Täiendavat informatsiooni on võimalik saada ka pilootprojekti vastutavalt uurijalt: Kristjan Krass, kristjan@diagnosticmatch.ee

Andmete töötlemise ja uuringus osalemise nõusolek
• Olen nõus, et mind kaasatakse "Indikaatorseisunditest juhinduva sihitud HIV-testimise pilootuuring esmatasandi perearstikeskustes kasutades digitaliseeritud teavitust" pilootuuringusse ning minu andmeid kasutatakse anonüümselt (isikustamata kujul) teadustöö raames.

Pilootuuringus osaleja
Nõustun uuringus osalema: JAH / EI
Eesnimi ja perekonnanimi:
Isikukood:
Tööhõive:
□ Tööga hõivatud (isik, kes uuritaval perioodil: - töötas ja sai selle eest tasu kas palgatöötajana, ettevõtjana või vabakutselisena; - töötas otsese tasuta pereettevõttes või oma talus; - ajutiselt ei töötanud) ;
□ Töötu (isik, kelle puhul on korraga täidetud kolm tingimust: - on ilma tööta (ei tööta mitte kusagil ega puudu ajutiselt töölt); - on töö leidmisel valmis kohe (kahe nädala jooksul) tööd alustama; - otsib aktiivselt tööd.)
□ Mitteaktiivne (õpingud, pikaaegne haigus või vigastus, rasedus, pensioniiga, heitunud)

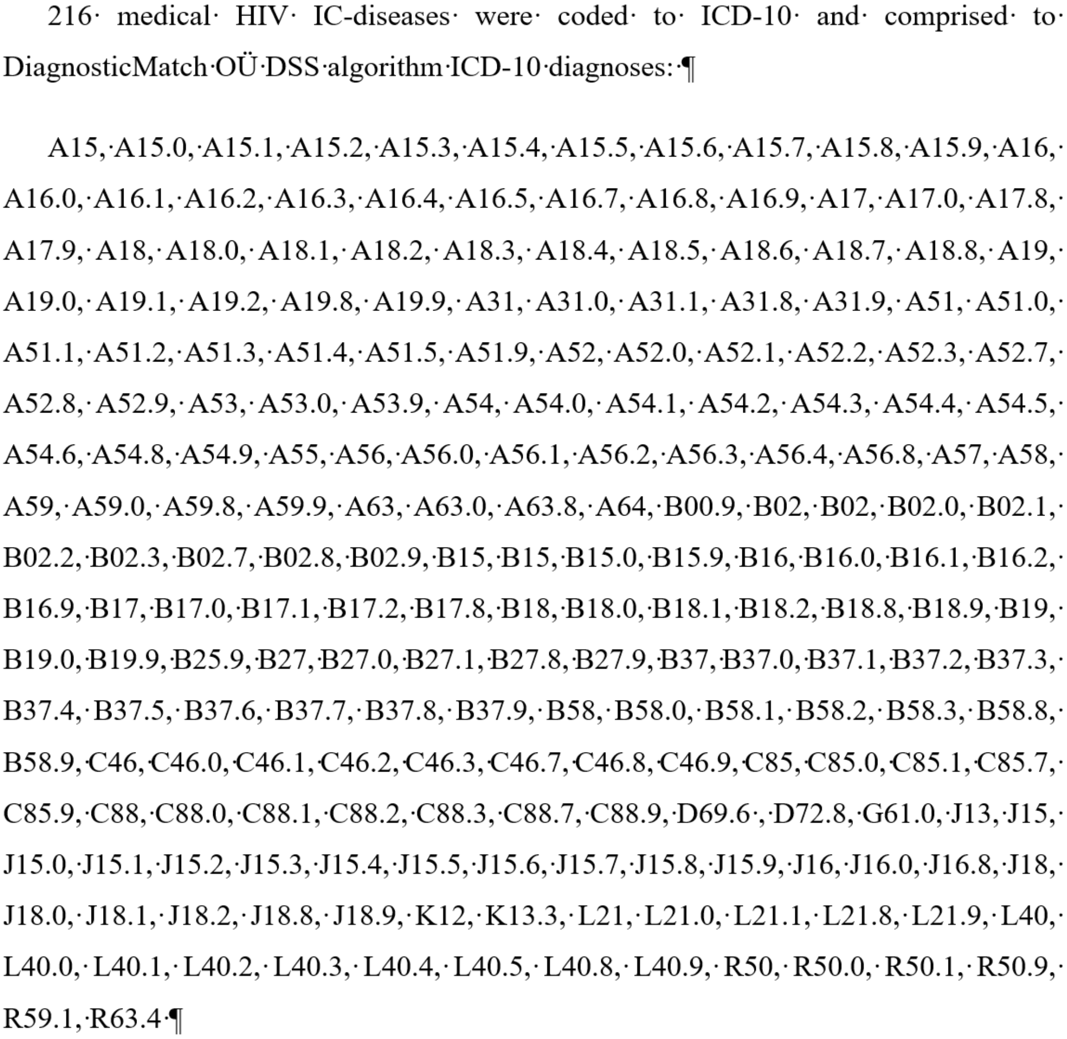
Haridus:
Esimese taseme haridus või madalam
□ põhiharidus (põhikooli 6 klassi), sellega võrdsustatud haridus või madalam
□ põhiharidus (põhikooli 9 klassi) või sellega võrdsustatud haridus
□ põhihariduseneõudeta kutseharidus, kutseharidus põhihariduse baasil
Teise taseme haridus, teise taseme järgne ning kolmanda taseme eelne haridus
□ üldkeskharidus
□ kutsekeskharidus (sh keskeri- või tehnikumiharidus) põhihariduse baasil
□ kutsekeskharidus keskhariduse baasil
Kolmanda taseme haridus
□ keskeriharidus keskhariduse baasil
□ kõrgharidus, magistri- ja doktorikraad
□ bakalaureus või sellega võrdsustatud haridus
□ magister või sellega võrdsustatud haridus
□ doktor või sellega võrdsustatud haridus

Rahvus:
□ eestlane
□ venelane
□ muust rahvusest

Allkiri:
Kuupäev:

Uuritavalt informeeritud nõusoleku võtnud isik:
Patsient nõustus HIV testiga: JAH / EI
Eesnimi ja perekonnanimi:
Telefon:
E-posti aadress:
Allkiri:
Patsiendi uuringu ID:

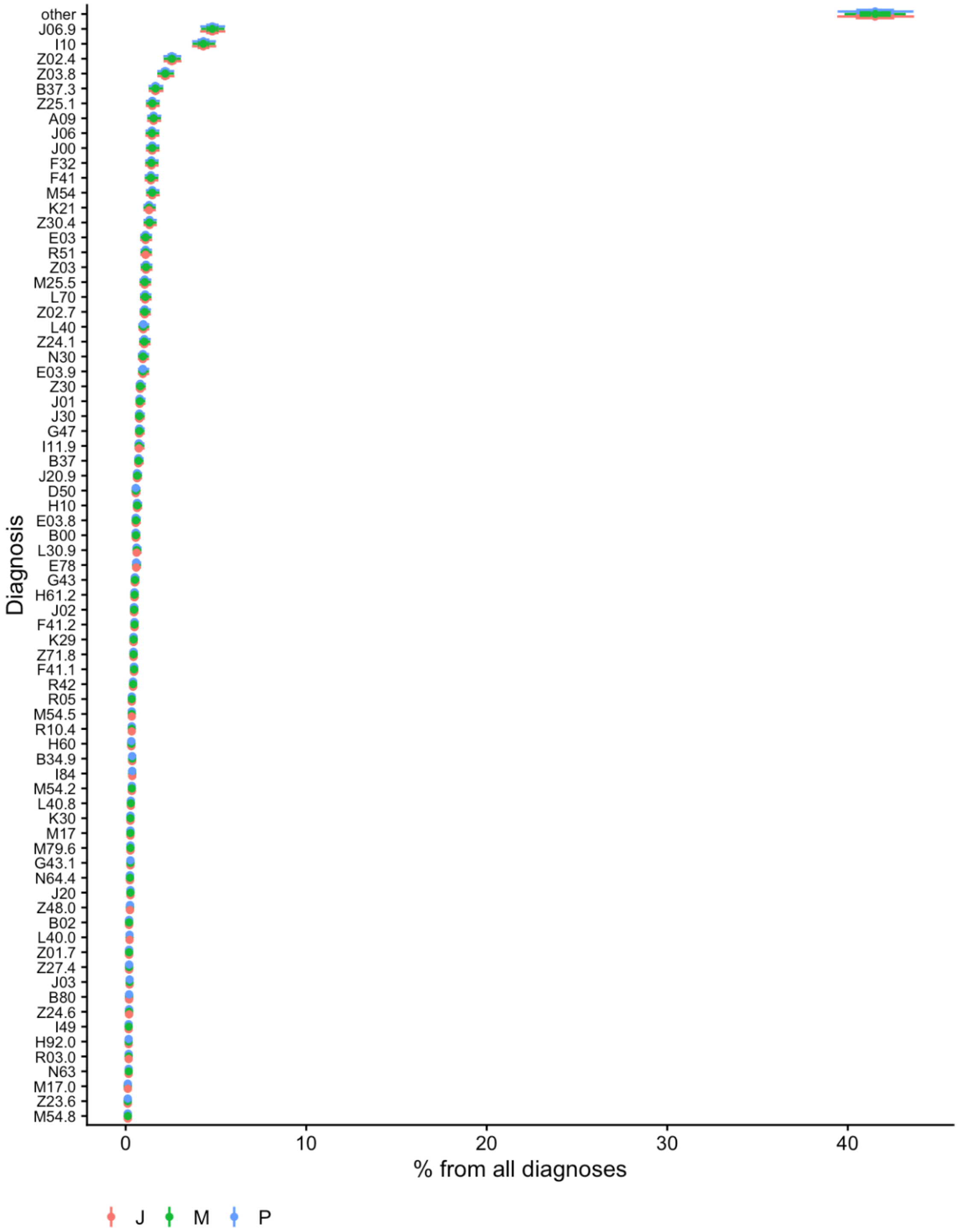
Appendix 3. 216 medical HIV IC-diseases were coded to ICD-10 and comprised of Diagnostic Match OÜ DSS algorithm ICD-10 diagnoses:



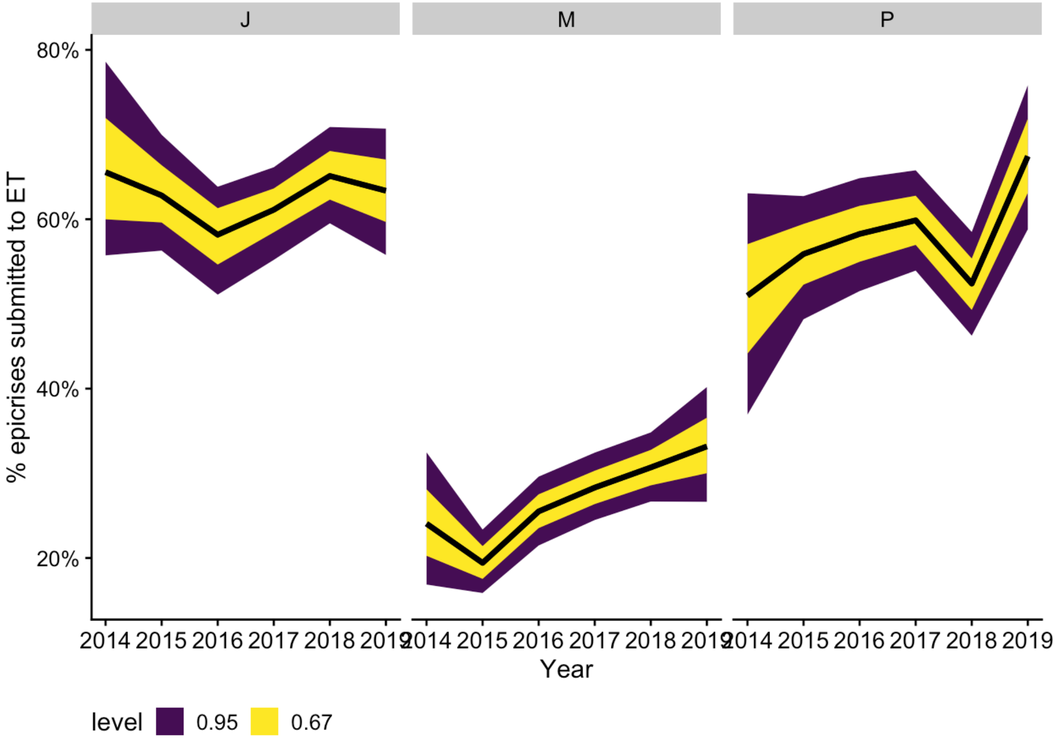
216· medical· HIV· IC-diseases· were· coded· to· ICD-10· and· comprised· to· DiagnosticMatch·OÜ·DSS·algorithm·ICD-10·diagnoses:·¶

A15,·A15.0,·A15.1,·A15.2,·A15.3,·A15.4,·A15.5,·A15.6,·A15.7,·A15.8,·A15.9,·A16,· A16.0,·A16.1,·A16.2,·A16.3,·A16.4,·A16.5,·A16.7,·A16.8,·A16.9,·A17,·A17.0,·A17.8,· A17.9,·A18,·A18.0,·A18.1,·A18.2,·A18.3,·A18.4,·A18.5,·A18.6,·A18.7,·A18.8,·A19,· A19.0,·A19.1,·A19.2,·A19.8,·A19.9,·A31,·A31.0,·A31.1,·A31.8,·A31.9,·A51,·A51.0,· A51.1,·A51.2,·A51.3,·A51.4,·A51.5,·A51.9,·A52,·A52.0,·A52.1,·A52.2,·A52.3,·A52.7,· A52.8,·A52.9,·A53,·A53.0,·A53.9,·A54,·A54.0,·A54.1,·A54.2,·A54.3,·A54.4,·A54.5,· A54.6,·A54.8,·A54.9,·A55,·A56,·A56.0,·A56.1,·A56.2,·A56.3,·A56.4,·A56.8,·A57,·A58,· A59,·A59.0,·A59.8,·A59.9,·A63,·A63.0,·A63.8,·A64,·B00.9,·B02,·B02,·B02.0,·B02.1,· B02.2,·B02.3,·B02.7,·B02.8,·B02.9,·B15,·B15,·B15.0,·B15.9,·B16,·B16.0,·B16.1,·B16.2,· B16.9,·B17,·B17.0,·B17.1,·B17.2,·B17.8,·B18,·B18.0,·B18.1,·B18.2,·B18.8,·B18.9,·B19,· B19.0,·B19.9,·B25.9,·B27,·B27.0,·B27.1,·B27.8,·B27.9,·B37,·B37.0,·B37.1,·B37.2,·B37.3,· B37.4,·B37.5,·B37.6,·B37.7,·B37.8,·B37.9,·B58,·B58.0,·B58.1,·B58.2,·B58.3,·B58.8,· B58.9,·C46,·C46.0,·C46.1,·C46.2,·C46.3,·C46.7,·C46.8,·C46.9,·C85,·C85.0,·C85.1,·C85.7,· C85.9,·C88,·C88.0,·C88.1,·C88.2,·C88.3,·C88.7,·C88.9,·D69.6·,·D72.8,·G61.0,·J13,·J15,· J15.0,·J15.1,·J15.2,·J15.3,·J15.4,·J15.5,·J15.6,·J15.7,·J15.8,·J15.9,·J16,·J16.0,·J16.8,·J18,· J18.0,·J18.1,·J18.2,·J18.8,·J18.9,·K12,·K13.3,·L21,·L21.0,·L21.1,·L21.8,·L21.9,·L40,· L40.0,·L40.1,·L40.2,·L40.3,·L40.4,·L40.5,·L40.8,·L40.9,·R50,·R50.0,·R50.1,·R50.9,· R59.1,·R63.4·¶

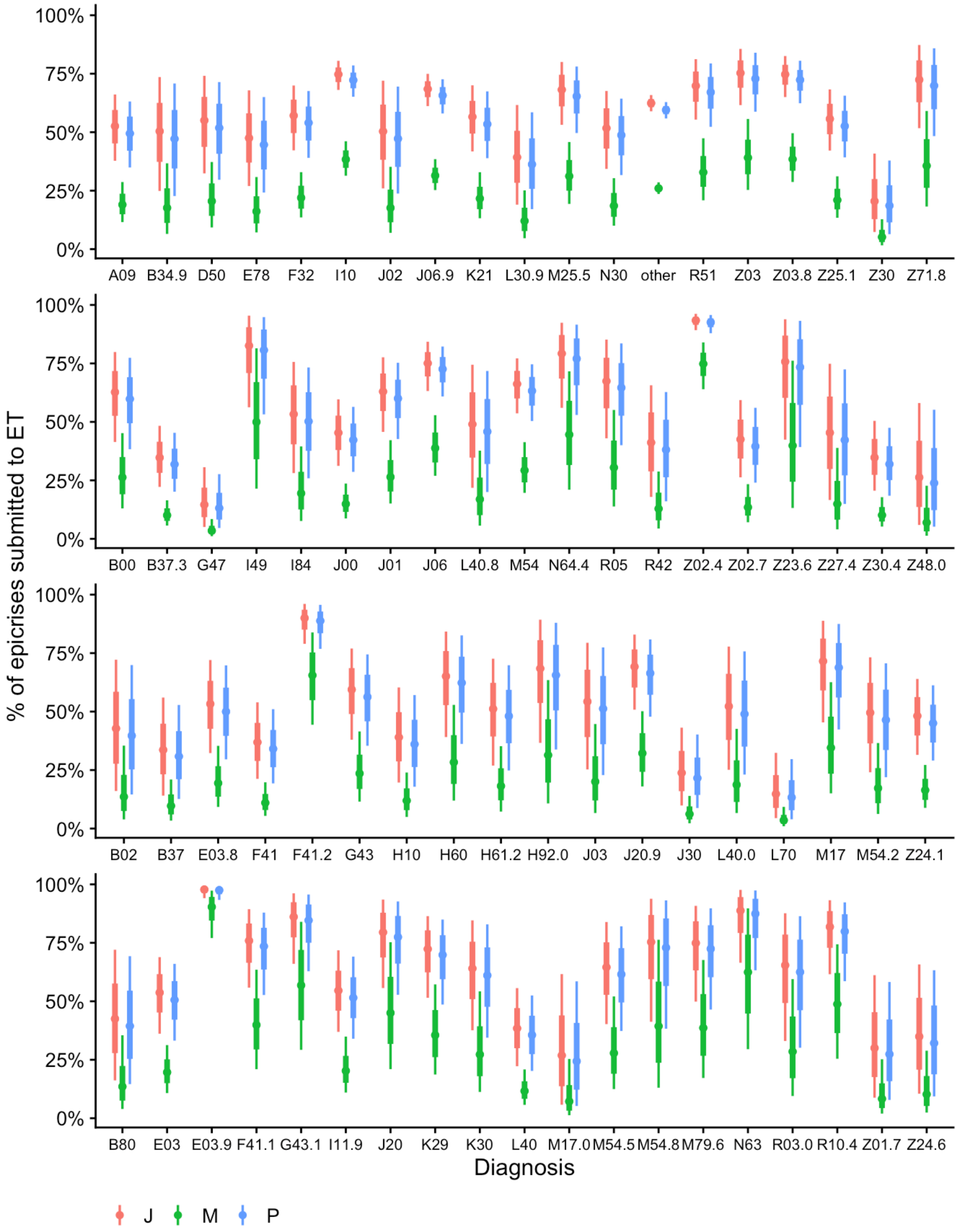
Appendix 4. Common diagnoses with “other”. Simple logistic regression reveals that common diagnoses are issued with similar frequency at three different PAK-s (n | trials(total) ~ diagnosis + PAK, binomial response distribution). N = 225. One diagnosis was randomly sampled from each case summary. Unique and fewer common diagnoses were assigned to class “other.” Points denote the model’s best fit. Thick and thin error bars denote 67% and 95% credible interval, respectively.



Appendix 5. The proportion of case summaries submitted to the ET database by three different PAK-s. A multi-level logistic model for the proportion of case summaries submitted to ET database where both intercepts and slope can vary by year (submitted ~ asutus + year + (asutus | year), Bernoulli response distribution), N = 4,061. Black lines denote the best fit of the linear model. Coloured regions denote credible interval at 0.95 and 0.67 level.



Appendix 6. The proportion of diagnoses submitted to ET. A logistic model with variable intercept for diagnoses reveals bias in diagnoses submitted to ET (submitted ~ asutus + (1 | diagnoos), Bernoulli response distribution). N = 4,061. Points denote model best fit. Thick and thin lines denote 67% and 95% credible intervals, respectively.


% of epicrises submitted to ET
A09 B34.9 D50 E78 F32 I10 J02 J06.9 K21 L30.9 M25.5 N30 other R51 Z03 Z03.8 Z25.1 Z30 Z71.8
B00 B37.3 G47 I49 I84 J00 J01 J06 L40.8 M54 N64.4 R05 R42 Z02.4 Z02.7 Z23.6 Z27.4 Z30.4 Z48.0
B02 B37 E03.8 F41 F41.2 G43 H10 H60 H61.2 H92.0 J03 J20.9 J30 L40.0 L70 M17 M54.2 Z24.1
B80 E03 E03.9 F41.1 G43.1 I11.9 J20 K29 K30 L40 M17.0 M54.5 M54.8 M79.6 N63 R03.0 R10.4 Z01.7 Z24.6
Diagnosis
J
M
P