TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Heidi Urmet YVEM178207

MAPPING THE NICE EVIDENCE STANDARDS FRAMEWORK AND INTERNATIONAL EXPERIENCES INTO A GUIDELINE FOR THE ASSESSMENT OF MOBILE HEALTH SOLUTIONS IN ESTONIA

Master's thesis

Supervisor: Janek Metsallik
MSc

TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Heidi Urmet YVEM178207

NICE TÕENDUSPÕHISUSE STANDARDITE RAAMISTIKU JA RAHVUSVAHELISTE KOGEMUSTE KAARDISTAMINE TERVISETEEMALISTE MOBIILIRAKENDUSTE HINDAMISE JUHENDIKS

Master's thesis

Supervisor: Janek Metsallik
MSc

Author’s declaration of originality

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

Author: Heidi Urmet

23.05.2019

Abstract

Background: mHealth apps are medical apps that have the ability to impact the user's health. These apps have the possibility to improve patient-doctor communication, health management, education and make exchanging and sharing medical data much more easier. There are numerous apps in the app stores today, but it is difficult to choose between all of them.

Aim: The aim of this thesis is to collect the standards that have been used so far to evaluate mobile apps in and create a list of evaluation items that can be used to evaluate apps in Estonia.

Method: Overall a case study research method was followed where the collection method is multiple systematic literature reviews. To find app evaluation categories, a literature overview was carried out. Grey literature reviews are done to find guidelines created by organisations and national entities.

Results: The chosen app evaluation categories are usability, credibility, functionality, privacy & security, interoperability and transparency. For finding relevant evaluation standards for each of these categories, a total on 1764 literature search results were review, after applying inclusion criteria 93 were chosen.

Discussion: The standards, frameworks and guidelines found from literature reviews are analysed and a set of evaluation items is compiled that could be used to evaluate mHealth apps.

This thesis is written in English and is 77 pages long, including 8 chapters, 4 figures and 20 tables.

Annotatsioon

Taust: Terviseteemalistel rakendustel on võimalus edendada patsiendi-arsti suhtlemist, oma tervise juhtimist, tõsta informatsiooni teadlikkust ning muuta meditsiiniliste andmete vahetamine ja jagamine palju lihtsamaks. Tänapäeval on rakenduste kauplustes arvukalt rakendusi, kuid nende vahel on raske valida.

Eesmärk: Käesoleva töö eesmärk on koguda standardeid, mida on seni kasutatud mobiilirakenduste hindamiseks, ja koostada hindamisobjektide loend, mis oleks kasutatav Eestis.

Meetod: Üldiselt järgiti juhtumiuuringute uuringumeetodit, kus kogumise meetod on korduv süstemaatiline kirjanduse ülevaade. Rakendushindamise kategooriate leidmiseks viidi läbi kirjanduse ülevaade. Organisatsioonide ja riiklike üksuste poolt loodud juhiste leidmiseks tehakse halli kirjanduse ülevaated.

Tulemused: Valitud rakenduste hindamise kategooriad on kasutatavus, usaldusväärsus, funktsionaalsus, privaatsus ja turvalisus, koostöövõime ja läbipaistvus. Kõigi nende kategooriate jaoks asjakohaste hindamisstandardite leidmiseks vaadati kokku 1764 kirjanduse otsingutulemust, pärast valikukriteeriumide rakendamist kahanes valik 93le.

Arutelu: Analüüsitakse kirjanduse ülevaadetest leitud standardeid, raamistikke ja suuniseid ning koostatakse hindamisteobjektide kogum, mida saaks kasutada terviseteemaliste rakenduste hindamiseks.

Töö on kirjutaud inglise keeles ning on 77 lehekülge pikk, sealhulgas 8 peatükki, 4 joonist ning 20 tabelit.

List of abbreviations and terms

EC	The European Commission
EHIF	Estonian Health Insurance Fund
GDPR	General Data Protection Regulation
MARS	Mobile App Rating Scale
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
SUS	System Usability Scale
WHO	World Health Organisation

Table of contents

List of Figures

List of Tables

1 Introduction

Millions of people use different kind of applications daily. Among these might be different health applications. Some use them just for their own fun or interest whereas some use them to help deal with chronic health issues.

Health applications open up a lot of possibilities between app users and doctors and healthcare providers. These applications provide a means to promote healthy habits and lifestyles through a means that is used daily by most. Apps can help monitor one's self management, vitals, progress. They can even help motivate the users. But as there are so many apps available, it can be very difficult to make the choice between one app and the other and oftentimes the ratings the app stores provide is not sufficient to make an educated choice as. These apps can be called mHealth apps. mHealth is a field in the health sector that aims to provide health care services and interventions through mobile applications and mobile devices (e.g. smartphones, tablets) [1] and mHealth apps are apps that are created created for a specific health or care related purpose and can be downloaded to one's mobile device. [2]

While there are so many apps available for the users, there are no specific guides or frameworks on how these apps could be evaluated. In a survey by the World Health Organization (WHO) 10 countries (22% of the respondents) indicated that they have a national entity that is responsible for the quality, safety and reliability of mHealth apps. Of the respondents, 16 countries (36% of the total) reported having a national entity that provides incentives and guidance on the innovation, research and evaluation of such apps. Although the guidance provided is not offered consistently on the national level. Several respondents mentioned that there is a need for evaluations to maintain an assurance that the apps are of good quality. The report recommends national health authorities to develop evaluation methodologies that could be used to evaluate the usability, functionality and meaningfulness of mHealth solutions. [3]

In a workshop in Tallinn, Estonia on the 14th of March 2019 on the topic of "Evaluating digital health services – what is a suitable solution for Estonia?" the author of the thesis

took part of, an evidence standards framework by the National Institute for Health and Care Excellence was introduced by Estonian Health Insurance Fund (EHIF) and was used in the workshop as an evaluation tool. The framework is meant for helping assess if the solution under evaluation is clinically effective. That framework is aimed towards solutions that are commissioned in the UK health and care system and is less relevant for solutions that users can download directly – mobile apps. During personal meetings with EHIF on the topic, it was expressed that the topic is relevant and important to them as well even though only one app had approached them on the topic of being considered a health care service.

Taking into account that the WHO recommends the development of an evaluation framework for mHealth solutions and EHIF is also interested in this topic, the aim of this thesis is to create a prototype of a possible guideline for evaluating mHealth solutions in Estonia based on literature found from literature reviews. The research questions this thesis aims to answer are the following:

RQ1. How to evaluate health mobile apps?

RQ2. What kind of standards exist for evaluating health apps today?

The main part of the thesis is divided into four parts. The first part describes what kind of international guidelines and frameworks already exist on the topic. The second part describes the methodology of the thesis. The third part describes the results that were found and finally, the fourth part analyses the results and develops the evaluation items.

2 Background and world experience

Some countries, organisations and researchers have written down guidelines or frameworks on how health apps should be evaluated. The following chapter brings forth a selection of these to illustrate how the rest of the world has tackled this problem.

2.1 From literature

To find the relevant literature on the subject, three databases were used. These databases are PubMed, ScienceDirect and Wiley. From each database the same search criteria was used. The search was carried out in March 2019. The first combination of words used for the search was “health AND mobile AND app AND guideline” which resulted in 65 results in PubMed, 841 in ScienceDirect and 2810 in Wiley. After browsing some of the titles of the articles from this search, it was clear the search string had to be altered to be more specific. The new search string was “mhealth AND app AND (framework OR guideline) AND (evaluation OR assessment)” which resulted in 121 results in PubMed, 383 in ScienceDirect and 209 in Wiley (including duplicates). The titles and sometimes abstracts were read through to find the exact articles that are of interest. The inclusion criteria included the following points

- The title or abstract had to have the keywords app and framework/guideline in it
- The title or abstract had to indicate that a framework or guideline was created or followed to evaluate health apps

After checking for duplicates, this resulted in 17 articles from PubMed, 11 from ScienceDirect and 5 from Wiley. Next each article was looked at more closely and based on their content they were either excluded or included. Articles that created or followed a framework or guideline that only looked at one aspect of an app (for example usability) were excluded. After this, 5 articles were chosen which to conduct the literature review portion of this thesis.

Brown et al. [4] and used the Mobile App Rating Scale (MARS) and Coventry, Aberdeen & London-Refined (CALO-RE) taxonomy for assessing different pregnancy iPhone apps. The apps were evaluated based on their quality, inclusion of behavioural change techniques and nutrition information. The MARS tool was used to evaluate the quality and CALO-RE to evaluate the app content. The quality of the apps was assessed using the MARS tool created by Stoyanov et al. [5]. The tool is a rating scale with categories that relate to app classification, app qualities and the subjective app quality. The MARS tool concluded that the apps were in moderate quality but scored higher than average in functionality and engagement. Brown et al. found the apps did not provide much clarity whether the information was evidence based and if it were clear, the app quality would rise. They found that using these frameworks was a strength because they have been established and are a reliable tool for the task.

McMillan et al. [6] chose to adapt the NICE health behaviour change guideline to assess the quality of mobile apps. The suggestions from the NICE guideline were changed into questions one could answer about the app. They extracted 9 themes from the guideline for the assessment: purpose, planning and development, usability, initial assessment and tailoring, behaviour change technique, maintenance and relapse prevention, evaluation, documentation and data protection. The tool was tested on apps found in the NHS App library. The authors found that even though the tool was systematic it was also very time consuming.

Zelmer et al. [7] created a framework for assessing eMental health apps in Canada. The aim was to achieve quality and ethics standards in the mental health mobile-based services. To create the framework the stakeholders (e.g. app makers, mental health professionals, end-users) were recruited to help guide which criteria is suitable for the framework. As a result a framework with nine guiding principles and 15 supporting criteria was created. These nine principles are evidence based, gender responsive, culturally appropriate, user centred, risk based, internationally aligned, enabling innovation, transparent and fair, and based on ethical norms. It also ensures transparency and fairness of all providers.

The outcome of this study can not only be used by mobile application makers to achieve a higher level of user benefit and satisfaction, but also by policy makers aiming to regulate newly-emerged market.

2.2 Existing guidelines and frameworks by organizations

Different organizations in the world have created guidelines or frameworks for assessing mHealth or digital health (but can be applied to mHealth) solutions. Following are the descriptions of three solutions from the World Health Organization, ORCHA and NICE.

2.2.1 Word Health Organization

The World Health Organization (WHO) has published a classification for digital health interventions (DHI) in 2018. Digital health interventions here are different digital and mobile technologies that can be used in health care. The primary targets for this classification are public health audiences. The classification was created to support the communication between public health practitioners and technology-oriented audiences. WHO recommends to use this classification with the list of Health System Challenges (HSC). The HSC can be used to explain and determine the problem the solution then wants to solve. [8] The classification can be used on more than mHealth solutions, but in this thesis, it is looked at form only the mHealth perspective.

The DHIs are organized into four categories: interventions for clients, interventions for health care providers (the health workforce who deliver health services), interventions for health system or resource managers (administration and oversight of public health systems) and interventions for data services (data collection, management, use and exchange). [8]

Interventions for clients as the name suggests are solutions that are aimed towards the public and the caregivers of clients receiving health services. It has subcategories for communication (for example public health notifications, health education, alerts for preventive services, test results, mass messaging, peer learning and messaging), tracking (for example access to health records, personal health monitoring, self-tracking and self-care), reporting (for example reporting on the availability and quality of services, accountability monitoring and reporting and disease notifications by users), information sharing (for example making out of pocket payments for services and voucher programs) and financial transactions. [8]

Interventions for health care providers as the name suggests are aimed at the member of the health care workforce who deliver the services health care providers offer. It has

subcategories for communicating and authenticating clients, keeping health care records, decision support, telemedicine, coordinating referrals between institutions, planning activities and trainings, prescription management and laboratory and imaging management. [8]

2.2.2 ORCHA

ORCHA is an organisation in the UK which focuses on health app evaluation and advising. They aim to create tools that can help health care professionals choose applications that suit their patients the best. In addition, they offer help to app developers in the health field. [9]

To evaluate the applications they have developed a review guide that helps divide the apps into 23 app categories (based on topic) [10] and gives the app a final score and level. It consists of seven stages. The first stage filters the eligible apps from App Store and Google Play. The second stage classifies the apps into five levels based on their functionalities and their area of focus. The higher the level the app has the stricter the review process is as the apps that have the highest levels are apps that for example analyse the gathered data or contain advanced features. The third stage is all about documenting what kind of functionalities an app possesses to make the search process easier for the end user in the future. The fourth stage is where data and security, clinical assurance and user experience is evaluated by reviewers. They have a yes/no questionnaire to fill out for each application. Even if an application seems to be compliant with an aspect of data and security, but the information of this is not easily findable or not clear the app is presumed to not be compliant. The fifth stage calculates the score for the application. The scores are divided into three. Apps that get less than 45% are considered to be lacking in areas severely, apps that get 45%-65% are not considered bad, but the user should make sure they are certain of using the app. During the sixth stage the developer is notified of their score so they can improve their score if they please to do so. After a certain amount of time the review is published. The final seventh stage is about making sure the review is up to date. If a new app version is released, the review becomes void. [11]

They have used their extensive review process to rate numerous applications available in different app stores (mostly) in the UK and create an app finder for whoever wishes with the reviews available for anyone. [12]

2.2.3 NICE

In a very recent publication by the National Institute for Health and Care Excellence (NICE), a classification scheme for digital health technologies (DHT) was presented. DHTs are different kinds of products (apps, software, online platforms) used in health care with the aim of benefiting the user in some way. The framework was created to help bring down the barriers for commissioning DHTs in the UK. The aim of the framework is to provide a criteria against which different kinds of DHTs can be compared to. It is not meant for describing an evaluation process for a DHT or for the assessment of security. It is aimed at developers, researchers, invertors, evaluators and commissioners among others. [13]

The DHTs are classified into evidence tiers in this framework by their function which will later help separate them into evidence tiers based on the risk to the end users. Tier one is for solutions that offer systematic benefits but no direct user benefits. For example an electronic health record system. Tier two is for solutions that help understand what healthy living is and illnesses but will probably have unmeasurable health outcomes. For example an application that provides healthy recipes. Tier 3a is for solutions that prevent and manage diseases and the outcomes can more likely be measured. For example an application that allows the user to record data and then send it to a physician. Tier 3b is for solutions that can be used for treatment and diagnosis or active monitoring and has clearly measurable health impacts. Solutions in this tier might also qualify as a medical device and as such would have to adhere to the medical device laws instead. An example of a solution in tier 3b is an app that advices clinicians on what diagnosis to give a patient. The evidence tiers can be seen in Figure 1 and a longer description can be found in Appendix 1 – Evidence tiers. Source National Institute for Health and Care Excellence . [13]

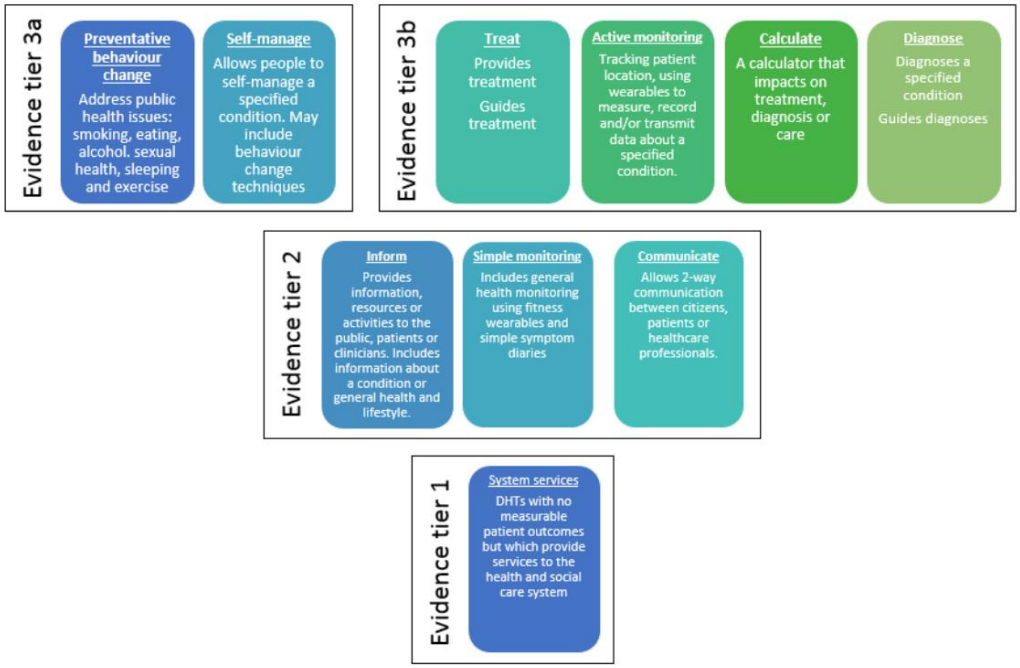


Figure 1. Evidence tiers by NICE framework. Source: National Institute for Health and Care Excellence *[13]*

The framework consists of three parts: evidence for effectiveness standards, evidence for economic impact standards and supporting resources. The first part – effectiveness standards – describes three evidence tiers in which the DHT could be and then details to which effectiveness standards a DHT in each category should adhere to. [13]

The second part – economic impact standards – are meant to create a streamlined process for evaluating and assessing the economic aspects of DHTs and is based on the current experience of the authors. The standards in this second part are divided into three components: key economic information (to create an economic model), appropriate economic analysis (preferably an analysis of the consequences of implementing the DHT from the perspective of a payer or commissioner) and economic analysis reporting standards. [13]

They have also specified which kind of evidence (minimum and best practice) should be able to be shown for each DHT to justify them being in a specific tier. For example to provide evidence for reliable information content, at minimum it must be possible to show that the information provided by the DHT is valid, accurate, up to date, reviewed and updated at defined intervals and sufficiently comprehensive.

2.3 Existing guidelines and frameworks by country

Different countries have created a guideline or framework for mHealth solutions. These are what one could call grey literature. Grey literature is literature that is produced on a governmental, business, academics or industry levels, but this literature is not controlled by commercial publishers. [14] The search for grey literature was carried out through Google and the keywords used for search were “app”, “health”, “evaluation”, “guideline”, and “framework”. In addition a few of the found literature even listed other countries and organizations that have app evaluation guidelines or frameworks.

Although there were a lot of countries with different kinds of guidelines, three of them were chosen to be described below. These three are France, the Netherlands and the USA.

2.3.1 France

In 2016 the French National Authority for Health (HAS; independent public scientific authority who aim to contribute to the regulation of the healthcare system) published a good practice guideline for health apps and smart devices. [15] In addition to being a guide to increase trust in health apps for consumers and developers, the guide is a baseline for evaluators to start their evaluation process on. This guideline is meant for apps and smart devices that are not medical devices but still can potentially change one’s health. The guideline is not meant as an assessment tool for reimbursement or professional recommendations. [16]

The guideline was put together with the help of literature review on the subject, a workgroup, review group and input from the stakeholders. They decided on five areas of assessment that are the most important. These are informing users, health content, technical content, security/reliability and usability/use. In addition to the categories there is a risk matrix where one should position their app based on the intended audience and main functionalities. This matrix then helps to determine which parts of the guideline are most important. Under each category of the five are questions the guideline user then has to answer. For each question it is then indicated whether an app of the determined risk level should be able to answer the question at hand. [16]

2.3.2 The Netherlands

The Royal Dutch Medical Association (KNMG) has also published a framework for assessing mobile apps. They call it the Medical App Checker. It is meant for assessing applications that qualify as a medical device, apps that track, store and share information and apps for communication. It is guided towards patients, physicians and caregivers and divided into three sections. The first section is about how one should/could search for medical apps. The second section is about assessing the quality of the app and the third focuses on evaluation the protection and data of personal information. They have also stated in their document that this guideline does not give the ultimate/last guarantee of the reliability or quality of the app. [17]

2.3.3 The USA

In the USA the Food and Drug Administration (FDA) approves medical apps. Medical apps are defined slightly differently though. In addition to it being software that is run on a smartphone or other mobile device, it is intended to be used as an accessory to a medical device or transform a mobile device into a regulated medical device. Only apps that fit this description will be regulated by the FDA. For apps that meet the definition of a device, but do not pose a high risk to the customers (for example an app that provides information about an illness) they do not have to be registered or approved by the FDA (the FDA exercises enforcement discretion). The aim of this guideline is to provide clarity and predictability to the manufacturers of these apps. [18]

The FDA has also published a guideline on how they intend to regulate the apps that meet their mobile app definition. It is meant for manufacturers, distributors and other entities. According to the guideline the app must fall under one classification and then adhere to the regulatory requirements that apply to that classification. There are three classes for devices that are classified based on the risk to patients and the necessary regulatory control level. The first class is for apps with low risk, the second class for apps with medium risk and finally the third class is for high risk apps. Most apps that fall under class one are the ones that the FDA exercises the enforcement discretion. [19] [20]

2.4 European level documentation and policies

In 2014, the European Commission published a green paper[1] about mobile health. The document discussed the potential mobile health and its potential market has and also what could be the issues that have to be tackled. The aim of it was to create a larger discussion amongst the stakeholders about the barriers and opportunities on mHealth in order to find the correct actions to further this field. The paper brought out multiple issues that need to be taken into consideration: data protection, big data, EU legislation, patient safety, transparency of information, role in healthcare systems, equal access, interoperability, reimbursement models, liability, research, innovation, international cooperation and access of web entrepreneurs. [21] In 2015 the summary report of the public consultation of the green paper was published that detailed the responses of the stakeholders. [22]

In 2014 the European Commission published a Staff Working Document which gives an overview of the existing legal framework of the time that can be applied to lifestyle and wellbeing apps. The document is aimed towards app developers. The document outlines the European Union safety and performance requirements (if the app does not fall under a medical device or an in-vitro diagnostic medical device, the app does not have any specific rules the app has to adhere to) and the app users' rights (the right to privacy and data protection, consumer's rights directive, eCommerce directive and rights set out in the Unfair Commercial Practices Directive). [23]

In 2016-2017 a workgroup was compiled to create a central guideline for the assessment of mHealth apps. The mandate of the group was "to develop guidelines for assessing the validity and reliability of the data that health apps collect and process". They workgroup chose five categories for that guideline. These are privacy, transparency, reliability, validity, interoperability. The conclusion of this work group was that a consensus between the members of the group was not reached and thus a guideline was impossible to compile. [24]

[1] a document of proposals and ideas on a certain subject that is published in order to provoke discussion https://www.collinsdictionary.com/dictionary/english/green-paper

A secondary draft of the guideline was published in 2016 with a more detailed guideline already in place. In this draft the existing legislation and standards were brought out. For the framework a set of criteria was chosen. This criteria contained nine aspects: stability, effectiveness, usability, transparency, security, safety, credibility, desirability and reliability. The process itself is divided into three activities: initial validation (does the app exists, can one download it etc), risk assessment and scrutiny (from medical and security aspect). [25]

Although these drafts suggest that work is being done, no evidence of a third of final draft has been found by the author.

3 Estonian example of an existing guideline

A physician looks at the data that is presented to them by laboratories and does not think twice before deciding on the next course of action for a patient. It is because these laboratories and medicines have been through a lengthy process of proving themselves and getting licenses and being accredited to ensure that people can trust them. This kind of trust process is missing for mobile apps. In the perfect world, a doctor would look at the data provided to them and prescribe the patient use an app for example.

Accreditation is the assessment and confirmation that a laboratory is in compliance with different standards and rules and the proficiency of the testing providers. It proves to the customers that they are trustworthy and know what they are doing. The assessment is regulated by specific standards for different fields. The process is completely voluntary and is done by the Estonian Accreditation Centre (EAC) which is recognized by the government and itself complies with the relevant international standards. [26] [27]

In Estonia there are 13 accredited medical laboratories. [28] To become an accredited laboratory the laboratory must go through the process and meet the requirements of quite a few standards. For a medical laboratory, it has to adhere to are the EVSEN ISO 15189:2012 (quality and competence).

Figure 2. Accreditation process in Estonia. Source Estonian Accreditation Centre [33]

The accreditation process is depicted in Figure 2. It is a nine-step process. A resource review means that the accreditation centre has to look over their capacities for making the review and if they done a similar review before. The initial assessment consists of

appointing a lead assessor to the application, creating the assessment programme, reviewing documentation, appointing an assessment team, carrying out the on-site assessment, reviewing corrective actions and closing nonconformities. After the assessment the decision is declared, and the certificate is issued for five years. After the decision is made, the lead assessor will put together a plan for future assessments taking into account how the initial assessment went. Once a year an on-site assessment will be carried out based on the programme created, corrective actions are reviewed, nonconformities closed, and the programme is updates if necessary. Before the certificate expires, a new review is necessary which means this process will start from the beginning. [29]

4 Method and methodology

When browsing through the described guidelines and frameworks from Chapter 2, it became apparent that a lot of them had followed a similar pattern when creating the guidelines. All the guidelines and frameworks that specified at least a little on the method mentioned reviewing literature and input from stakeholders. Some also added workgroups and surveys and consultations with experts to the list to make sure the end result is as relevant as possible. For creating the framework only one article described what kind of method was used – Zelmer et al. [7] used a modified Delphi method. If at first glance it seemed odd that a lot of those has approached the issue at hand in a similar way but had not mentioned explicitly what framework or guideline they were using, it became apparent that these methods loosely following different manuals or guidelines for development guidelines. For example, the "Manual for ESHRRE guideline development" by the European Society of Human Reproduction and Embryology (ESHRE) [30] aimed towards creating recommendations for improving health care delivery quality in human reproduction and embryology, the "The guidelines manual" by NICE for creating guidelines for clinical guidelines [31] or the "WHO Handbook for Guideline Development" by WHO for guidelines about clinical or public health problems or policy areas. Each of those has similar flow, but also at the same time each of them is different from each other. The guidelines roughly follow the following steps: planning (topic selection, determining scope, timeline, proposal etc), doing a systematic review to gather evidence, creating a steering group with which the first draft of the guideline is created, have the draft reviewed, edit the document as necessary and publish it. The thesis at hand very loosely follows the steps for creating a guideline as well, with the exception of creating a screening group to create the framework.

As the first research question for this thesis is a "how" question and the topic that is being researched is a relatively new and contemporary field and does not have anything of the equivalent in Estonia already, an exploratory case study approach was chosen.

The process is described by Robert K. Yin [32].The process itself is an iterative process and is illustrated in Figure 3.

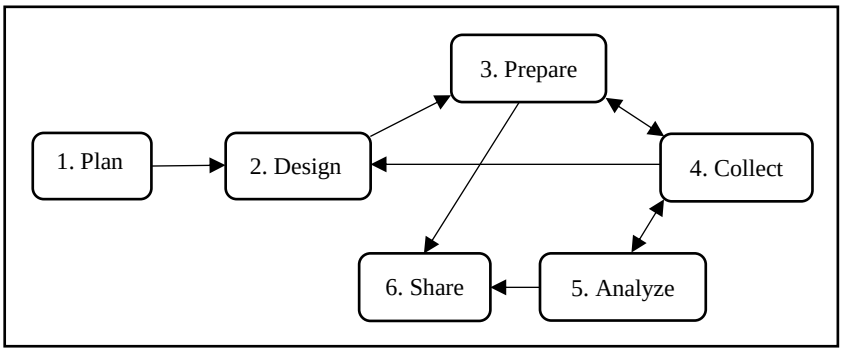


Figure 3. Case study process. Source Robert K Yin [33]

Following the process of a case study from Figure 3, the first step – planning - was carried out before writing this thesis, so the discussions and research that went into deciding on a specific topic. The second part – design – is what the current chapter aims to describe. The third step – prepare – is about looking at the experiences of the world and how others have evaluated apps as a whole before. Based on this, the app evaluation categories were chosen for the next steps. The next step – collect – is about collecting all the articles about specific app evaluation categories. The collection of data is further described in Chapter 4.1. The final and 6th step in the process is share which in the case of this thesis will be the submission of this thesis. The third, fourth and fifth step of the process were carried out twice in the course of this thesis. Once to determine app evaluation categories and secondly to find standards for the categories that were determined from the first review.

4.1 Choosing the evaluation categories

For choosing the categories, the literature overview done in Chapter 2 was used. In addition, the references of the selected studies were browsed to see if any relevant literature could be found from there as well that described the different categories an app can be evaluated in.

4.2 Systematic literature reviews of app evaluation topics

To understand and find the existing literature on evaluating all the six chosen categories, systematic literature reviews were carried out about each topic. A guideline proposed by Kitchenham [33] was used as a structure to carry out the literature reviews. The guideline consists of three big sections – planning the review, conducting the review and reporting the review. In the first section the need for a systematic review and review protocol are discussed. In the second section the review is carried out (database searches, quality assessment, data extraction, data synthesis) and in the final section the results are reported and discussed.

In this thesis the need for systematic reviews comes from the desire to understand how (health) mobile apps are evaluated in different categories and are there specific standards that can be used to evaluate them. These reviews then help answer RQ2.

The search process was a manual search done on different databases by the author of the thesis during May 17th 2019 and May 19th 2019. To search for relevant literature the databases PubMed Central, ScienceDirect (Elsevier), Wiley and IEEE were chosen.

In each database the timeframe from which to search articles from was from 2009 to today. The year 2009 was chosen because during an exploratory search into the topic indicated that there are very few papers on the topic from before that as mHealth is a relatively new field. Even for categories that are not health specific – e.g. privacy & security – the timeframe was chosen to make sure that the newest and up to date knowledge would be found. The category selection reasoning is described in Chapter 6.1.1.

As there were six different systematic reviews carried out, it means that there were six different sets of search strings, inclusion criteria and data extraction tables. For each search string there were common keywords which these were “mobile”, “app”, “evaluate”, “standard” and “assess”. Then depending on the topic, keywords were added to the list. In all the reviews the search was done based on the abstract or in some cases abstract and title and keywords. This was chosen because searching through all the possible fields resulted in too many results.

The specific search terms, exclusion criteria and data extraction for each category will be described in the next six subchapters.

If the systematic review produced too few results, an additional grey literature review was done with the same keywords as meant for the topic to understand if there were any standards or guidelines created by specific organisations or countries already that could be helpful or were missed. The procedure for the grey literature research is described in Chapter 4.2.7.

4.2.1 Usability

To find out how usability has been evaluated and which standards were used, the keywords "usability" and "user experience" were added to the list of keywords described before. For different databases the search strings were:

- mobile AND app AND (usability OR (user AND experience)) AND (evaluate OR standard OR assess)
- mobile[Abstract] AND app[Abstract] AND (usability[Abstract] OR (user[Abstract] AND experience[Abstract])) AND (evaluate[Abstract] OR standard[Abstract] OR assess[Abstract])

The inclusion of studies was carried out in two parts in this thesis. Firstly, based on the title and abstract alone the article was included if

- The full text was available
- It was in English
- The article was not a systematic review
- The title/abstract indicated that the usability of apps was evaluated or a framework for evaluating usability was created

In the second round of inclusion criteria, the article was evaluated based on its content and was included if

- The article states what method was chosen for evaluating usability
- The article described adequately the reasoning behind the method choice
- The article reports some results for evaluation or framework validation

For data extraction, the article title, source database and evaluation method was extracted into an Excel sheet for further analysis and documentation. The results for this systematic review can be found in Chapter **Error! Reference source not found.**.

4.2.2 Credibility

To find out the standards and methods used for evaluating the credibility of mHealth apps, the keywords “credible”, “evidence” and “quality” were added to the original list of keywords. This resulted in the following search strings for different databases:

- Mobile AND App AND health AND (credible OR evidence OR quality) AND (evaluate OR assess OR standard)
- Mobile[Abstract] AND App[Abstract] AND health[Abstract] AND (credible[Abstract] OR evidence[Abstract] OR quality[Abstract]) AND (evaluate[Abstract] OR assess[Abstract] OR standard[Abstract])

The inclusion of studies was carried out in two parts in this thesis. Firstly, based on the title and abstract alone the article was included if

- The full text was available
- It was in English
- The article was not a systematic review
- The title/abstract indicated that the credibility of apps was evaluated or a framework for evaluating usability was created

In the second round of inclusion criteria, the article was evaluated based on its content and was included if

- The credibility or evidence-baseness of apps was evaluated or reviewed
- The method used for the evaluation or review was described and justified
- Some results on the evaluation or review were present

For data extraction, the article title, source database and what the evaluation was based on was extracted into an Excel sheet for further analysis and documentation. The results for this systematic review can be found in Chapter 5.3.

4.2.3 Functionality

To find out the standards and methods used for evaluating the credibility of mHealth apps, the keywords “functionality”, “operability” and “performance” were added to the original list of keywords which meant that the following search strings were used in different databases:

- mobile AND app AND (functionality OR performance OR operability) AND (evaluate OR standard OR assess)
- mobile[Abstract] AND app[Abstract] AND (functionality[Abstract] OR performance[Abstract] OR operability[Abstract]) AND (evaluate[Abstract] OR standard[Abstract] OR assess[Abstract])

The inclusion of studies was carried out in two parts in this thesis. Firstly, based on the title and abstract alone the article was included if

- The full text was available
- It was in English
- The article was not a systematic review
- The title/abstract indicate that the functionality or operability or performance of apps was evaluated or discussed

In the second round of inclusion criteria, the article was evaluated based on its content and was included if

- The functionality or operability or performance of mobile apps are evaluated
- The method for evaluating said indicators is described
- The results of the evaluation are present

For data extraction, the title of the article, source database, evaluation method were marked down in an Excel file for further analysis and documentation. The results for this systematic review can be found in Chapter 5.4.

4.2.4 Privacy & security

For finding literature on privacy & security of mobile apps, the keywords “privacy”, “security” and “framework” were added. “Framework” was added because during the initial research on the topic and test runs in the databases, this extra keyword added

some results that were pertinent to the topic. In addition, the keyword “mobile” was switched with “health” because apps that deal with medical personal data might need different regulations. So thus the search strings in different databases were:

- health AND app AND (privacy OR security) AND (evaluate OR assess OR standard OR framework)
- health[Abstract] AND app[Abstract] AND (security[Abstract] OR privacy[Abstract]) AND (assess[Abstract] OR evaluate[Abstract] OR standard[Abstract] OR framework[Abstract])

The inclusion of studies was carried out in two parts in this thesis. Firstly, based on the title and abstract alone the article was included if

- The full text was available
- It was in English
- The article was not a systematic review
- The title/abstract indicate that the privacy & security of apps was evaluated or discussed or a standard presented

For the second round, the full contents of the article was looked at. An article was included if

- The privacy or security of mobile apps was evaluated
- The standard or guideline of framework against which the evaluation took place was described or the privacy or security of mobile apps was discussed in relation to a specific standard or regulation

For data extraction the title of the article, the source database and the regulation or guideline or standard against which mobile apps were evaluated, was written down for further analysis and documentation. The results for this systematic review can be found in Chapter 5.5.

4.2.5 Interoperability

For finding literature on interoperability of mobile apps, the keywords “interoperable”, “compatible” and “information exchange” were added. In one/two databases the first two were switched with “interoperab*” and “compatib*. The two have an Asterix to

them because it became apparent that more relevant literature could be found that way. So thus the search strings in different databases were:

- health AND app AND (interoperab* OR compatib* OR (information AND exchange)) AND (assess OR evaluate OR standard)
- health AND app AND (interoperability OR compatibility OR (information AND exchange)) AND (assess OR evaluate OR standard)
- health[Abstract] AND app[Abstract] AND (interoperability[Abstract] OR compatibility[Abstract]) AND (evaluate[Abstract] OR assess[Abstract] OR standard[Abstract])

The inclusion of studies was carried out in two parts in this thesis. Firstly, based on the title and abstract alone the article was included if

- The full text is available
- Text is in English
- The article is not a systematic literature review
- The title/abstract indicate that the interoperability of health themed apps was evaluated or discussed or a standard provided

For the second round, the full contents of the article was looked at. An article was included if

- A framework or standard for interoperability is used or discussed or created or the interoperability of an app is evaluated

For data extraction, the article title, source database, and method (or standard or regulation) used for data exchange is collected into and Excel sheet for further analysis and documentation. The results for this systematic review can be found in Chapter 5.6.

4.2.6 Transparency

When searching for transparency evaluations it was very difficult to find anything on the topic. Especially to find evaluations of health related apps. In this care the search keywords described before were not used at all and instead the keywords “app” and “transparency” were used to find as much literature on the topic as possible. Again since

some of the databases allowed for an asterix in the search terms, the following search strings were used:

- app[Abstract] AND transparen*[Abstract]
- app AND transparen*
- app AND transparency

Like all the previous reviews in this thesis, the articles were evaluated twice. Once based on the title and abstract only and a second time based on the content. An article was included based on the title and abstract if

- The full text is available
- Text is in English
- The article is not a systematic literature review
- The title/abstract indicate that the transparency of apps was evaluated or discussed

For the second round, the entire contents of the article was looked at. An article was included if

- It describes a standard or guideline or framework an app has to adhere to in terms of transparency
- If the app was evaluated in terms of transparency, the results had to be reported

For data extraction, the article title, source database and standard or regulation or method for transparency was extracted into an Excel sheet for further analysis and documentation. The results for this systematic review can be found in Chapter 5.7.

4.2.7 Grey literature research procedure

The aim of conducting an extra grey literature research for a topic is to help add more value to specific categories. The literature was search for through the Google search engine. Based on the topic the same keywords were used for the search as described in the previous chapters.

Only the first few pages of Google search results were looked through. The title which could be seen in the search results had to indicate that the page or document that is

linked there evaluates mobile apps or presents a standard that mobile apps could be evaluated against. Even news articles about app evaluation were looked through to see if they link to any frameworks or guidelines. The search results chose cannot be from any source, it had to be well understood what the source of the document is. Mostly documents by governments, European Commission of established organizations were selected.

5 Results

In this chapter, the results of the systematic literature reviews and grey literature reviews done are presented after carrying out the review as described in Chapter 4.2 and its subchapters.

5.1 Category results

The search process for evaluation categories and descriptions of the results are in Chapter 2. In addition to the search done for the background chapter, the references of the found studies were looked at to see if there was any literature that is relevant to the search and was missed. The evaluation categories from these sources is presented in Table 1.

Table 1. App evaluation categories across studies

Source	Evaluation categories
Good Practice Guidelines on Health Apps and Smart Devices (Mobile Health or mHealth) [16]	Informing users, health content, technical content, security/reliability and usability/use
ORCHA [11]	Data and Security, Clinical Assurance and User Experience
Report on international practice on digital apps [34]	Privacy/data protection, credible sources/evidence-based information, functionality, security/authentication, usability/user experience, effectiveness/impact, interoperability
Second draft of EU guidelines on assessment of the reliability of mobile health applications [25]	Usability & accessibility, desirability, credibility, transparency, reliability, technical stability, safety, effectiveness, and privacy & security
Mobile App Rating Scale [5]	Engagement, functionality, aesthetics, and information quality; and one subjective quality scale
Quality assessment of a sample of mobile app-based health behaviour change interventions using a tool based on the	Purpose, planning and development, usability, initial assessment and tailoring, behaviour change technique, maintenance and relapse prevention,

Source	Evaluation categories
National Institute of Health and Care Excellence behaviour change guidance [6]	evaluation, documentation and data protection
An Assessment Framework for e-Mental Health Apps in Canada: Results of a Modified Delphi Process [7]	Evidence based, gender responsive, culturally appropriate, user centred, risk based, internationally aligned, enabling innovation, transparent and fair, and based on ethical norms
Challenges in Assessing Mobile Health App Quality: A Systematic Review of Prevalent and Innovative Methods [35]	Scientific/clinical basis, functionality, usability, accountability, impact, and popularity

5.2 Usability

When following the search instructions from Chapter 4.2 and Chapter 4.2.1, the initial search resulted in 519 articles found across the four sources. After removing duplicates, the number of articles was reduced down to 514. After applying the first set of inclusion criteria described in Chapter 4.2.1, the amount of articles was decreased down to 57 and after applying the second set of inclusion criteria described in Chapter 4.2.1, the number of articles left was 41. The numbers for this process can also be seen in Table 2.

Table 2. Usability search results after applying inclusion criteria

	Initial search results	After removing duplicates	After applying first set of inclusion criteria	After applying second set of inclusion criteria
IEEE	212	211	12	8
ScienceDirect	134	134	11	8
PubMed	142	138	30	22
Wiley	31	31	4	3
Sum	519	514	57	41

Of the 41 articles that evaluated usability tasks as an evaluation method were used 25 times. Most of the times the tasks were derived from the functionalities and features the apps provide or even based on ISO standards.

Questionnaires and surveys were also very popular, they were used in some for or other 32 times. In some cases, the questionnaire was created within the study to match the

functionalities of the app specifically, but in other cases other existing questionnaire were used. The System Usability Scale (SUS) was used 14 times. Some of the other questionnaires (in their original form or modified) were Computer System Usability Questionnaire (CSUQ), Health Information Technology Usability Evaluation Scale (Health-ITUES), Post-Study System Usability Questionnaire (PSSUQ), and Usefulness, Satisfaction, and Ease of Use (USE) Questionnaire.

Interviews were used six times within those 41 studies. Oftentimes the interview went hand in hand with tasks. In 23 of the studies more than one method for evaluating was used. Both Gunter et al. [36] and Nugraha et al. [37] took the ISO 9241-11 standard basis for usability evaluation. In addition 5 studies used heuristic evaluation. These numbers can also be seen in Table 3.

Table 3. Count of usability evaluation methods used and the respective studies

Tasks	Survey	Questionnaire	System Usability Scale	Interview, think aloud	Heuristics
25	7	11	14	15	5
[38], [39], [40], [41], [42], [42], [43], [36], [44], [45], [46], [47], [37], [36], [45], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57]	[58], [40], [59], [60], [61], [61], [53]	[38], [62], [63], [64], [65], [48], [66], [50], [67], [54] , [68]	[38], [39], [42], [69], [70], [36], [63], [64], [71], [56], [57], [72], [73], [74]	[40], [41], [42], [43], [44], [46], [61], [61], [47], [50], [53], [54], [55], [56], [57]	[40], [41], [75], [65], [52]

For usability, a brief grey literature search was also done using the same keywords as for the systematic literature search. Of the search the results which were found to be useful and have a potential use in this thesis are described in Table 4.

Table 4. Usability grey literature search results

Source	**Description**
Human Interface Guideline [76]	A collection of recommendations for an interface by Apple. The aim of it is to help app developers design their apps with the highest quality, functionality and best user experience in mind.

Source	Description
Design for Android [77]	A set of guidelines to ensure that the Android app is of the highest quality – in visual and navigation patterns, but also in compatibility, performance and security.
Web Content Accessibility Guidelines (WCAG) 2.1 [78]	A collection of recommendations that aim to make Web content more accessible. Though the guideline was initially created for Web content, it has also bene updated to be possible to be used on mobile devices as well.

5.3 Credibility

The search as per the description in Chapters 4.2 and 4.2.2 resulted in 340 articles found across the four databases. Two duplicates were removed from the list, making the total number of articles 338. After applying the first set of inclusion criteria described in Chapter 4.2.2, only 30 articles were left. After applying the second set of criteria described in the same chapter, 12 articles were chosen. The numbers described here can be seen in Table 5.

Table 5. Credibility search results after applying inclusion criteria

	Initial search results	After removing duplicates	After applying first set of inclusion criteria	After applying second set of inclusion criteria
IEEE	32	32	1	0
ScienceDirect	72	70	5	2
PubMed	194	194	19	8
Wiley	42	42	5	2
Sum	340	338	30	12

In 4 ([79], [80], [81], [72]) of the 12 studies, apps were evaluated against national or international guidelines such as the American College of Sports Medicine Guidelines or the European Resuscitation Council Guidelines for Resuscitation 2015. These guidelines could only be used if the content of the app matched that guideline. For example, Xie et al. [82] evaluated Chinese apps about cardiovascular disease with three frameworks and used guidelines recommended by the National Library of Medicine of

the National Institutes of Health to do it. Grainer et al. [80] combined using MARS and the American College of Rheumatology and European League against Rheumatism (ACR and EULAR) guidelines for monitoring of RA disease activity.

McMillan et al. [6], evaluated apps against different behavioural change techniques – they created a general list of topics for different apps. Brown et al. [4] used both MARS and CALO-RE (a taxonomy of behaviour change techniques) to evaluate the credibility of apps. Santo et al. [84] and Bakker et al. [85] both used only MARS to evaluate the credibility of apps that aim to improve medical adherence and to evaluate a low moods and anxiety management app respectively.

De Korte et al. [86] and Crane et al. [87] both searched for the behavioural change techniques in use in mHealth apps. De Korte et al. found that there is still limited presence of these techniques in mHealth apps for mental and physical health.

Richardson et al. [88] assessed the quality of apps that educate parents of neonatal intensive care unit patients and used the Trash it or Trust[2] it tool to asses it's credibility. They found that five out of the 18 apps they evaluated were deemed trustworthy for educating the parents.

For credibility a grey literature research was done, but nothing that added extra insights into the topic was found.

5.4 Functionality

For functionality the search was carried out in the databases mentioned in Chapter 4.2 and the search strings from Chapter 4.2.3. The initial search resulted in 574 articles found. Only two were removed as they were duplicates. After applying the first inclusion criteria described in Chapter 4.2.3, 33 articles were left. After applying the second inclusion criteria also described in Chapter 4.2.3, 15 articles remained. The course of the number of articles decreasing can also be seen from Table 6.

[2] http://www.trustortrash.org

Table 6. Functionality search results after applying inclusion criteria

	Initial search results	After removing duplicates	After applying first set of inclusion criteria	After applying second set of inclusion criteria
IEEE	318	317	5	1
ScienceDirect	122	122	8	3
PubMed	93	92	16	9
Wiley	41	41	4	2
Sum	574	572	33	15

For functionality only the MARS evaluation scale was used for assessing functionality was 9 times ([4], [89], [80], [91], [92], [93], [94], [95], [96]). Although MARS is intended for evaluating the entire app, it has a section dedicated to functionality. Anderson et al. [97] synthesised peer reviewed usability evaluation checklists and a study of user experiences into a checklist they could use to evaluate the app.

Fernandes et al. [98] created a framework for the assessment of the performance of mobile apps. The framework takes the information about the app's expected operations, creates a program based on it and the operations are executed, collecting the information about performance. This framework though is aimed more towards developers.

In the evaluation criteria created by Coulon et al. [99] from existing tools and literature the functionality is assessed as well. The functionality subpoint assesses ease of use, reliability and performance, appearance and design. DiFilippo et al. [100] created an app quality evaluation (AQEL) tool which among other things also assesses the technical functionality of an app, but this app is only aimed towards evaluating nutrition apps. Brown et al. [101] assessed the fit of using the Health IT Usability Evaluation Model (Health-ITUEM) to evaluate mobile apps. One of the concepts in the model is also performance speed and from the study it was one of the most frequent used codes that was gathered during their assessment.

Idri et al. [102] used the ISO/IEC 25010 standard for evaluating apps. The standard covers functional suitability, reliability, performance efficiency, operability, security,

compatibility, maintainability and transferability. A framework developed based on this standard was used.

For functionality a grey literature search was carried out. Three sources of interest were found and are described in Table 7.

Table 7. Grey literature on functionality

Source	Description
Xcertia mHealth App Guidelines [103]	Guidelines on how to evaluate mHealth apps. These guidelines are not available yet, they are in the phase of collecting feedback.
Digital Assessment Questionnaire [104]	A list of questions created by experts for assessing apps and tools for the NHS Apps Library.
European Commission second draft on mHealth evaluation guidelines [25]	Presented a list of questions to evaluate if the app is technically stable.

5.5 Privacy & security

For privacy & security the search was carried out in the databases mentioned in Chapter 4.2 and the search strings from Chapter 4.2.4. The initial search resulted in 126 articles found. Only two were removed as they were duplicates. After applying the first inclusion criteria described in Chapter 4.2.4, 29 articles were left. After applying the second inclusion criteria also described in Chapter 4.2.4, 11 articles remained. The course of the number of articles decreasing can also be seen from Table 8.

Table 8. Privacy & security search results after applying inclusion criteria

	Initial search results	**After removing duplicates**	**After applying first set of inclusion criteria**	**After applying second set of inclusion criteria**
IEEE	48	48	8	3
ScienceDirect	33	33	9	4
PubMed	36	34	11	4
Wiley	9	9	1	0
Sum	126	124	29	11

Not every article from the chosen 11, evaluated the privacy and/or security of mobile apps. There was one conference paper by Ferreira and Muchagata [105] which brings out the key changes GDPR brought with it, the impact of it and described a use case of how GDPR is used correctly within an app. Hutton et al. [106] and used GDPR for evaluating privacy of the apps. Huckvale et al. [107] evaluated the apps against the 1998 Data Protection Act. This same study used a man-in-the-middle attack approach also to find any security issues. In another study by Parker et al. [108] GDPR and the Australian Privacy Principle were used to assess privacy policies of apps.

HIPAA was used in a few studies ([109], [110] and [111]) for evaluating specifically health apps and their privacy. HIPAA is an American act that states how personally identifiable information should be protected from fraud and theft.

Three chosen studies used different sources for evaluating the privacy of apps. Hussain et al. [112] created a mHealth Apps Security Framework (MASF) that for example analyses the installation of the app, and the different policies. The authors found the framework to be very effective against different attacks. O'Loughlin et al. [113] reviewed the privacy policies of different mHealth apps by combining the Enlight Evaluation tool by Baumel et al. and the App Evaluation Model by the American Psychiatric Association. Lastly, Robillard et al. [114] also reviewed and analysed mHealth but created the coding system for analysis based on the first 10% of literature they found.

As it was already known to the author of the thesis that there are specific Estonian legislation that pertains to the field, a grey literature research was carried out as well using the same keywords described in Chapter 4.2.4.

Table 9. Grey literature on privacy & security

Source	Description
Three-level IT Baseline Security System ISKE [115]	Information security standard developed for Estonia which is compulsory for state and local government organisations. Can be used to assign a security class and then describes what standards the system should adhere to.
European Commission second draft on mHealth evaluation guidelines [25]	In the guideline privacy and security are measured against the EU Code of Conduct on mHealth App Privacy. This code of conduct is

Source	Description
	not relevant anymore though but a list of questions about security and privacy are still described.

5.6 Interoperability

As with the previous subchapters, the search for interoperability related studies was carried out following the procedure described in Chapters 4.2.5 and 4.2. The initial search resulted in 66 articles across four databases. Four were removed from the list on the account of being duplicates, making the total number 64. After applying the first inclusion criteria 12 articles remained and after applying the second set 10 articles remained. These numbers across the phases are also depicted in Table 10.

Table 10. Interoperability search results after applying inclusion criteria

	Initial search results	After removing duplicates	After applying first set of inclusion criteria	After applying second set of inclusion criteria
IEEE	24	23	6	5
ScienceDirect	10	9	2	1
PubMed	14	14	4	4
Wiley	18	18	0	0
Sum	66	64	12	10

Of the final 10 studies, six created new apps using different interoperability standards though not all of the apps are available for consumption. The other studies created an middleware information model, reviewed different interoperability standards, designed a system for better reporting or improved mHealth apps. Table 11 illustrates how many of the chosen studies use which standards.

From the results the most popular standard that was used when designing mHealth apps was Fast Healthcare Interoperability Resources (FHIR) which is a standards framework created by HL7 and among other uses can be used with mHealth apps [116]. FHIR incorporates different standards for data exchange such as LOINC, SNOMED-CT, ICD-

9 and ICD-10. [117] From the 10 studies, seven studies described using FHIR. Of those six, four studies used Substitutable Medical Applications and Reusable Technologies (SMART) on FHIR which is a an improvement on the existing FHIR. For example authentication was added to further improve the FHIR standard. [118]

Rossi et al. [119] used the CDA-2 standard as the data exchange standard in an app supporting for supporting homecare transcranial Direct Current Stimulation therapy. Plastiras et al. [120] describe the development of an information model used the HL7 CDA standard. Pfiffner et al. [121] incorporated the FHIR standard with the ResearchKit (a framework that allows to create apps for medical research) that is available through Apple.

Adamko et al. [122] gives an overview of existing issues on interoperability and describes standards. They recommend to use SNOMED-CT or HL7 because of their wide acceptance.

Table 11. Interoperability standards used in chosen studies

	SMART on FHIR	**FHIR**	**HL7-CDA or CDA-2**	**Other**
Studies	[118], [123], [124], [125]	[121], [126], [127]	[119], [120]	[122]
Count	4	3	2	1

As with privacy & security, it was known beforehand that some research had been done for this in Estonia so the grey literature research was carried out. The results of the research can be found in Table 12.

Table 12. Interoperability grey literature

Source	**Description**
Report of legal and technical alternatives for integrating with health information systems [113]	A report analysing the legal and technical alternatives for integrating with the health information system in Estonia. Main outcome: mapped the existing and wished for components on which to build an interoperable service; to accommodate IT companies' wishes, no regulatory changes are needed.

5.7 Transparency

After carrying out the search described in Chapters 4.2. and 4.2.6, there were a total of 139 articles found across the four sources. Firstly, the articles were searched for duplicates. This resulted in one article being removed, keeping the total number of articles at 138. After applying the first set of inclusion criteria described in Chapter 5.1.6 only nine articles were left and after applying the second set on inclusion criteria, only four articles remained which were chosen for this thesis. These numbers are also depicted in Table 13.

Table 13. Transparency search results after applying inclusion criteria

	Initial search results	**After removing duplicates**	**After applying first set of inclusion criteria**	**After applying second set of inclusion criteria**
IEEE	90	90	2	2
ScienceDirect	13	12	2	1
PubMed	27	27	5	1
Wiley	9	9	0	0
Sum	139	138	9	4

Coulon et al. [99] identified, evaluated and presented evidence-based apps about effective stress management. A part of this evaluation was also transparency evaluation. To evaluate transparency, they used clinical standards that every intervention should be evidence based, transparent in its purpose, development and content and user friendly. Regarding transparency, the app had to provide information about the developers, whether medical personnel was involved and contact information, state that it is not a replacement for a physician, address confidentiality and privacy, provide references and justification, relevant financial information, and advertising policies. As a result of the evaluation, the authors found that two thirds of the selected apps addressed at least half of the criteria for transparency, only 12% of the apps addressed all the criteria.

In an article by Grundy et al. [128] on the data sharing practiced of health mobile apps, it was found that there is little transparency around third party data sharing and overall

shortcomings on providing privacy assurances. On a positive side, GDPR has forced more transparency on data sharing among some apps.

Although Fahy et al. [129] discusses the data privacy and transparency of mobile apps, the topic is generic enough that it can be relevant to health apps as well. They describe how App Store and Google Play can significantly influence the field from their side. These two ecosystems have different layers where transparency rules or guidelines are enforced. As a result both Apple and Google encourage the developers to provide a privacy policy and in it be transparent about the collection and use of personal data.

Muchagata and Ferreira [105] looked into how GDPR fits with mHealth. In relation to transparency, the privacy policies and terms and conditions have to be easily readable and understandable and most importantly is has to show that the company is transparent with how the data is processed by who and where and for what purpose.

A grey literature search was also conducted to see if something was available on the topic from other sources and to find if there were any Estonia specific literature on this. The search terms were “app”, “transparency” and “standard” (in Estonian and English) and the search was carried out on Google but no new relevant literature was found that could be of use.

6 Discussion and creation of evaluation items

The guideline could benefit multiple parties involved with health mobile apps. Developers can get an indication into which kind of direction they need to guide their development process and also be aware of standards beforehand.

For the app end users the guideline can be too overwhelming and difficult to understand. The evaluation process will take time and if the user just wants to download an app, then the faster and easier solution for them is to just check if the app has been evaluated to a specific standard and what are its shortcomings and strengths. For the end users a simple and fast tool should be necessary, something more the lines of MARS because it has less questions and might not need to dig too deep into the app.

For a commissioner this guideline is probably not as helpful. They are more interested in knowing if the app should be considered a health care service, so usability or even functionality may not be of the utmost importance. The guideline in this thesis is more to understand the quality of these health apps.

6.1 RQ1 – How to evaluate health mobile apps?

In general there were very few sources found where an health app was evaluated as a whole. In most cases, apps specific to one topic or only one aspect of apps was evaluated. There were surprisingly few app evaluation frameworks/tools found from literature. From literature the most popular seems to be using MARS which has been cited more than 250 [130] times since its publication in 2015.

As it can be seen from the different sources found in the background and systematic reviews, apps can be evaluated in many ways. Some use surveys, questionnaires, tasks or even automated tooling. In the sources found, the approach to evaluating an app was to take on a list of specific criteria (e.g. in the form of questions, statements, or just topics) and evaluate the app against those. A similar approach was chosen for this thesis.

As not all apps are equal in their functionalities and risks, it is important to somehow categorize it before evaluating it. In the guideline by France [16], the app is categorized with a risk matrix based on the intended users and the app functionalities. In [NICE] the app will be categorized based on its functionalities and risk into four evidence tiers. The evidence tiers were briefly described in Chapter 2.2.3 and a longer explanation can be found in Appendix 1. Though this is the approach chosen in this thesis, not all evaluate apps like this. For example, in MARS there is a set number of questions for evaluation and they hold true to any kind of app.

The framework to in most cases was a list of questions. For example, the MARS framework has topics and numerous questions underneath each. Similarly, the French guideline has questions about each section which were mandatory or not based on the risk level of the application. The second draft for the European Commission also had mostly yes or no questions about the different topics. One of the proposed scoring systems was based on the risk level of the app (low, medium or high) which then indicated if it was necessary for the question to be answered positively. As this approach was seen in a few places, it was decided a similar scoring approach would be used. Instead of the risk levels to indicate whether a question is mandatory or not, the evidence tiers from the NICE guideline are used (Tiers 1, 2, 3a and 3b). For each evaluation item it is indicated whether an app of a specific tier has to fulfil it. The possible requirement levels for each tier are mandatory (marked as M in the following tables), nice-to-have (marked as NTH in the following tables) or not applicable (marked as NA in the following tables). An example of how each evaluation item will look like can be seen in Table 2.

Table 14. Example framework evaluation item

	Item	Tier 1	Tier 2	Tier 3a	Tier 3b
0	The app is blue.	Not applicable	Nice to have	Mandatory	Mandatory

6.1.1 Choosing evaluation categories

To determine which sections the guideline should contain the different guidelines and sources that were found and reported in Chapter 5.1 were taken as a reference. Many of them evaluated apps in categories that overlapped in name or even definition. Some

chose less categories and being more vague (like ORCHA with only three categories) whereas some very specific and had multiple categories (e.g. European Commission framework with nine categories).

After reading the definitions for the different categories proposed from the different sources, the following categories were chosen for the guideline being described in this thesis: usability, credibility, functionality, privacy & security, interoperability, and transparency. This list does not contain all the possible categories that were described in Table 1, but hopefully is the minimum acceptable criteria that could cover all facets of the app that makes it a quality app. In addition to these categories, the basic app information and choosing of the evidence tier section should also be a part of the guideline.

Usability combines two categories in itself: the user experience (UX) and the user interface (UI). Usability refers to the ease of use of an app. The UI in combination with the user should make for a good UX. The ISO 9241-11 standard defines usability as the extent which a system or product can be used in order to fulfil the specified goals efficiently, effectively and comfort. [131] It is important to evaluate the usability of apps in order to make a good app and make sure that it has users and is thus successful. Usability is especially important if the app is targeted towards users with limiting disabilities.

As the apps that are under evaluation are health apps the **credibility** of the health information it offers is very important. There apps can impact the health of the user and it could have devastating affects if a drug calculator app would calculate the dosages wrong. To make sure the medical information the app advertises or uses are evidence based the credibility of the app has to be evaluated. Of all the categories chosen, this is one of the most important.

The **functionality** category is about evaluating the app's functionality and its performance. The performance of the app can be very important when the app is used for real time monitoring for example. This category is probably the least to have to do with health as it aims to understand if the app itself is capable of providing the user a smooth experience under different circumstances and that the app runs without any issues.

The **privacy & security** category is to evaluate if the app has taken all the necessary steps for ensuring data privacy and security. If the app sends or collects personal (medical) data, it needs to comply with different laws or regulations. For apps that do not require no user input or do not communicate with any outside source, this section might not be as important.

Interoperability is the capability of sending health related data to other entities e.g. physician's system, electronic health record. It makes it possible for the health data to be available for more people and in different environments. This evaluation category does not apply to all apps, only the ones who wish to exchange data. To offer a wide range of integration possibilities to apps and to incentivise health care providers to integrate apps into their systems, it is necessary to know what evaluation criteria should the app answer to.

Transparency is about evaluating who built the app, who are the stakeholders, who is it funded by, who are the beneficiaries and so on. A transparent app (the development, testing etc) only helps build more trust in the app. An important part of transparency is also succinct explanations about data collection and retention and sharing of said data. If all this information is available, the app provider is capable of building trust with the app users.

6.2 RQ2 – What kind of standards exist for evaluating health apps today?

The standards that were found for the different categories varied. For some (e.g. credibility) there were no specific unified standard to use, for some only guidelines or frameworks that others have developed.

It is important to understand what kind of standards exist in order to build a guideline that only expects the best from the apps. It is also important to know that even though different standards are presented here for evaluating apps, it does not mean that the app or the developers do not have to comply to any other laws or regulations that exist. The following chapters discuss each evaluation category based on their results.

6.2.1 Usability

In usability evaluations was the biggest mix of different methods that wre used. Although there is an ISO standard for usability (ISO 9241), only one of the articles found used that to evaluate the app. No other standards on this topic were found. As this category can change quite fast (with new technologies, new designs or even new interfaces) it can be quite difficult to create a standard that could apply to health apps and stand the test of time. In addition, an

Of all of the categories evaluating usability was the most common.

As was found from the literature review, quite a few usability evaluations used the SUS questionnaire for evaluating the usability. In a project report for Ministry of Economic Affairs and Communications in 2014 [132], the SUS was brought out as an example of what to use to measure usability. Because of its frequency in the found literature and mentioning in the ministry document, SUS was chosen to be the evaluation method for usability in this thesis as well. As SUS was created for evaluating systems, the word "system" was replaced in each evaluation item with "app" to make it more easily readable. Each item still retained its intention and clarity. The questions for usability can be seen in Table 15.

Table 15. Usability evaluation items

	Item	**Strongly disagree**				**Strongly agree**
1	I think I would like to use this app frequently.	1	2	3	4	5
2	I found this app unnecessarily complex.	1	2	3	4	5
3	I thought this app was easy to use.	1	2	3	4	5
4	I think that I would need assistance to be able to use this app.	1	2	3	4	5
5	I found the various functions in this app were well integrated.	1	2	3	4	5
6	I thought there was too much inconsistency in this app.	1	2	3	4	5

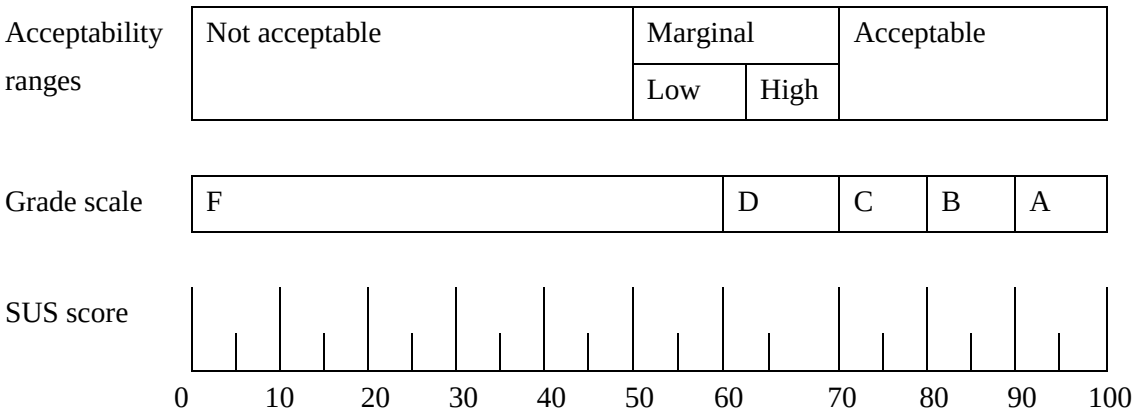
	Item	Strongly disagree				Strongly agree
7	I would imagine that most people would learn to use this app very quickly.	1	2	3	4	5
8	I found this app very cumbersome/awkward to use.	1	2	3	4	5
9	I felt very confident using this app.	1	2	3	4	5
10	I needed to learn a lot of things before I could get going with this app.	1	2	3	4	5

6.2.1.1 Usability scoring

As the final score of the scale is in the range of 0 – 100, some calculations need to be made to get this from the one to five scales that are in the table. For items 1, 3, 5, 7 and 9 (so every odd item), 1 should be subtracted from the score. So, for example, if item 1 has a score of 3, the score for this item will be 2. For items 2,4,6,8 and 10 the score given for the item has to be subtracted from 5. For example, if item 2 has a score of 2, the calculation should return 3 as the value. After this has been done, the sum has to be multiplied with 2.5 in order to have it in the range of 0-100. This makes the score now comparable. [133]

In order to understand if the score that an app has, it should be known what score is good and which is bad. As one would expect, lower scores are not as good and higher scores are better. Bangor et al. [134] and Bangor et al. [135] put the SUS scores next to an adjective rating scale and a grade scale. The scales can be seen in Figure 4. As a result, it is easier now for people who use the scale give the apps a score in a format which is known to most people. For this thesis it was decided to use the acceptability ranges Bangor et al. [135] developed. The scale consists of three ranges – not acceptable (score of 0-50), marginal (51-70) and acceptable (70-100). Marginal also divides into two – low (51-62.26) and high (62.27 – 70). Apps that fall into the low marginal category just barely have a passable score, apps that have a high marginal score have a passable score, almost fully acceptable. So, apps that want to pass this evaluation, they have to achieve a score the falls under marginal or acceptable ranges of the acceptability range.

Figure 4. SUS score grade scale and acceptability range. Sources Bangor et al. [39] and Bangor et al. [40]



6.2.2 Credibility

For apps in different medical and health categories the credibility standards have to be different. From the review results it became apparent that very specific guidelines (like clinical guidelines or behavioural change techniques) can be used for evaluation if the app is specifically geared towards a specific medical goal. A third of the studies used MARS to evaluate the health information content. The information section of the scale asks seven questions about the app. The most popular option that was chosen for evaluating apps was using different medical guidelines provided by national organizations. For example Modave et al [81] used the American College of Sports Medicine fitness principles and guidelines for evaluating the credibility of the app whereas Xie et al. [82] used guidelines recommended by National Library of Medicine of the National Institutes of Health to evaluate the app.

Because of different kinds of health apps having to adhere to different clinical standards and guidelines it is almost impossible to create or find a comprehensive evaluation method or tool or questions that can be applied to every single app. Even so, many of the studies had to translate the clinical standards and guidelines into something that can be checked from the apps. Evaluations meant for specific groups of apps can be more rigorous and detail oriented in the group. Because of this ambiguity and the fact that the NICE framework was already use in this thesis, it was decided to not use any specific clinical guideline to formulate the credibility questions, but the NICE evidence effectiveness framework instead.

For each evidence tier the framework described what kind of evidence it should offer to indicate its evidence base. Each of those was taken and changed into an evaluation item for the guideline for this thesis. For example, one of the evidence categories in the framework is “Credibility with UK health and social care professionals” with the minimum evidence standard it requires. In this case it was that medical personnel had been involved in the makings of the technology or that a medical professionals have indicated their approval of the solution. This prompted the evaluation item “Indicate that relevant medical personnel have been involved in the development process or approved it.” which can also be seen in Table 16. All of the evaluation items about credibility for the framework can be found in Table 16.

Table 16. Credibility evaluation items

	Item	Tier 1	Tier 2	Tier 3a	Tier 3b
11	Indicate that relevant medical personnel have been involved in the development process or approved it.	M	M	M	M
12	Evidence that the app is relevant in the context of Estonia.	M	M	M	M
13	If relevant, evidence that the data recorded or manipulated in the data is accurate, reproducible and relevant.	M	M	M	M
14	If applicable, evidence that the data is transmitted does not change when transmitting.	M	M	M	M
15	The information provided by the app is relevant.	NTH	M	M	M
16	The information provided by the app is correct.	NTH	M	M	M
17	The information provided by the app is up to date.	NTH	M	M	M
18	The information provided by the app is updated at certain time intervals.	NTH	M	M	M
19	The information provided by the app is comprehensive.	NTH	M	M	M
20	Data on the usage of the app is available to decision makers.	NA	M	M	M
21	Data on user satisfaction or outcomes is	NA	M	M	M

	Item	Tier 1	Tier 2	Tier 3a	Tier 3b
	collected and available to decision makers.				
22	Studies of the effectiveness of the app are available.	NA	NTH	M	M
23	Evidence that the techniques used in the app are based on published and recognized sources	NA	NA	M	M
24	Evidence that the techniques are appropriate with the target audience.	NA	NA	M	M

6.2.3 Functionality

For functionality, it was presumed that more specific standards would emerge as this topic can be evaluated regardless of the category of the app and could be generic enough that some standards have emerged. In the literature review the most used method for evaluating functionality was the MARS tool which has which covers the functionality section in four questions. But when looking at the Digital Assessment Questionnaire by NHS, it covers functionality in a technical stability section with more questions.

One study used the ISO standard ISO/IEC 25010 to evaluate the functionality of the app as the app is a software product and that is what the standard is aimed towards. The standard has separate sections for functionality and performance, that could be applicable to mobile apps, partly if not fully. No Estonian specific standard was found for this topic. To create the evaluation items for functionality, the ISO standard was consulted, app operability guideline from Xcertia [103], the second draft by the European Commission and the Digital Assessment Questionnaire developed by NICE to assess products available in the NHS App Library [104]. For example, one of the Xcertia operability guideline items is "*The app downloads and installs on the target device(s) and target operating system(s) as confirmed by user notification*." [103], and one of the questions in the EC draft was "Does the app install and uninstall properly?" [25] which were then used as inspiration to create the evaluation item "The app can be downloaded and installed without issues.". The full list of items can be found in Table 17.

Table 17. Functionality evaluation items

	Item	Tier 1	Tier 2	Tier 3a	Tier 3b
25	The app can be downloaded and installed without issues.	M	M	M	M
26	The app appropriately lets the user know when something has gone wrong (e.g. network requests, wrong input).	M	M	M	M
27	If applicable, the app connects to the internet without any issues.	M	M	M	M
28	If applicable the app performs without issues without an internet connection.	M	M	M	M
29	If applicable, the app connects to secondary devices without issue.	M	M	M	M
30	If applicable, the app connects to other applications without issue.	M	M	M	M
31	The app is updated regularly	M	M	M	M
32	The app does all it advertises and aims to do	M	M	M	M
33	The performance of the app does not deteriorate over time and use.	M	M	M	M
34	If applicable, the calculations the app does are correct and reproducible.	M	M	M	M

6.2.4 Privacy & security

When evaluating privacy, the most common route is to evaluate the privacy policies of the apps. One study explicitly states that there is no specific regulation on the topic in China and thus evaluated security based on self-reported security measures. [136]

As the literature review revealed, the General Data Protection Regulation (GDPR) is the most important legislation for developers that want to deal with collecting data from users. A European Commission workgroup tried to write down a privacy code of conduct for health apps but failed. They came to the conclusion that with the

introduction of GDPR, an app should adhere to the guidelines of GDPR and as the proposed code of conduct it failed to do so, so GDPR is currently the ruling word. [137] [138] But they still brought out the most important aspects app development should take into account: user consent; purpose limitation and data minimization; privacy by design and by default; data subject rights and information requirements; data retention; security measures; advertising in mHealth apps; use of personal data for secondary purposes; disclosing data to third parties for processing operations; data transfers; personal data breach and data gathered from children.

In Estonia, for data protection, there are a few standards and good practice guidelines. These are ISKE, Information Technology Infrastructure Library and ISO standards with the serial number 2700x. But as the AKI mentions on their website, there is no universal data protection standard that would fit every need. [139] As ISKE is already in use in the health information system, it was chosen to be one of the sources from which to create evaluation items from. Based on what the app intends to do with personal health data, different ISKE levels apply to them. [115]

If the app intends to send or collect user data (be it medical or not), the app has to compliant with GDPR. To be compliant with GDPR, multiple checklists and guidelines have been made by various sources. For example, a website on GDPR, backed by the European commission supplies a high-level checklist[3]. Some have even created a checklist for app developers to make sure they do all that is possible to comply with GDPR.[4][5][6] For complying with GDPR, evaluation items from the aforementioned sources were compiled and added to the list. All evaluation items for security and privacy can be seen in Table 18.

Table 18. Privacy and security evaluation items

	Item	Tier 1	Tier 2	Tier 3a	Tier 3b
35	The app only collects and stores data that is necessary for its purpose.	NA	M	M	M
36	If applicable, is the app compliant with GDPR.	NA	M	M	M
37	If applicable, the third-party apps the app	M	M	M	M

[3] https://gdpr.eu/checklist/
[4] https://fueled.com/blog/gdpr-for-app-developers/
[5] https://www.mobiloud.com/blog/gdpr-compliant-mobile-app/
[6] https://medium.com/intuz/gdpr-for-mobile-app-owners-ac3228a3d2b7

	Item	Tier 1	Tier 2	Tier 3a	Tier 3b
	connects to are also GDPR compliant.				
38	The personal data that is sent is encrypted.	NA	M	M	M
39	If applicable, the app is compliant with ISKE S2	NA	NA	M	M
40	If applicable, the app is compliant with ISKE T3	NA	NA	M	M
41	If applicable, private medical data is only accessible after logging into app.	NA	NTH	M	M
42	If applicable, the data exchange is sufficiently encrypted.	M	M	M	M

6.2.5 Interoperability

From the literature review, this category was the strongest that offered different standards that could be used. HL7, FHIR and all the other standards that were used are internationally recognized and in use. In the more recent studies FHIR and SMART on FHIR has been used indicating that these might be becoming more popular.

In a report analysing the possible legal and technical alternatives for integrating health applications and the health information system, it states that even though multiple existing solutions have many hints and suggestions on what should and could be done for successful and easy integrations, no single solution matched the Estonian context fully. The report concluded that there is no need for regulatory change, most problems now arise from economical, organizational, and technical obstacles. The report also found that information about integrating is right now scattered and buried in different documents and guidelines. There is currently no standard for data exchange that is available for everybody and creating that might take years and years. [115]

As in Estonia the data exchange between health care providers and the health information system happens over X-road and uses the HL7 standard, mobile apps that want to send or read health data from HIS, should strive towards using HL7 or FHIR for this. This is of course dependent on the health information systems being capable of sending and receiving information to apps as well. But as the report stated, it could be necessary that new data exchange standard could be necessary to create and implement and that could take years. Although if an app wants to communicate with another health

care provider or software that offers a different data exchange format and standards, those can be taken into account as well.

As data exchange may not be a part of every app, not all apps have to evaluate the interoperability of the app. Taking all of this into account, the questions in Table 19 were put together to cover the interoperability section of the framework.

Table 19. Interoperability evaluation items

	Item	Tier 1	Tier 2	Tier 3a	Tier 3b
43	The app states if data is exchanged with another app.	NA	M	M	M
44	If applicable the app is capable of sending or uploading information to the health information system.	NA	M	M	M
45	If applicable, the standards which are used for sending data are stated.	NA	M	M	M
46	If applicable, the data exporting formats are stated.	NA	NTH	M	M
47	If applicable, the app states if it shows data from other sources (e.g. EHR).	NA	M	M	M
48	If applicable, the standards in which the app stores data are stated.	NA	M	M	M
49	If applicable, evidence that the data does not change during exchange is available.	NA	M	M	M

6.2.6 Transparency

The literature review showed that there is very little literature on evaluating the transparency of health or even mobile apps and the literature that was found on the standards, has to do with GDPR.

Article 12 of GDPR explains how one must comply to it and what it entails. In addition, the “Guidelines on transparency under Regulation 2016/679” [140] provides a guideline for transparency. The guideline takes parts of the GDPR and explains each part of the sentence in detail.

When it comes to transparency of mobile apps no standards were found for it and no Estonian specific frameworks or guidelines were found. In literature when transparency

is referenced, oftentimes it goes hand in hand with data privacy. For this thesis, not only the transparency of data privacy is chosen, but also the transparency of the makings and people involved in the app. Even when looking at the seconds draft for the EC guideline, some of the transparency questions deal with data privacy. To create the evaluation items for transparency, GDPR, the EC guideline and the literature review results were used. GDPR was the basis for questions about the privacy policy and a few questions were derived from the EC guideline and the results from the literature review. The evaluation items can be found in Table 20.

Table 20. Transparency evaluation items

	Item	Tier 1	Tier 2	Tier 3a	Tier 3b
50	The app has information or a way to get into contact with the developers or app company.	M	M	M	M
51	The app lists all partners and contributors that took part in the development process.	M	M	M	M
52	The app lists all stakeholders and beneficiaries	M	M	M	M
53	A privacy policy is available	M	M	M	M
54	The app specifically asks for consent to collect data.	NA	M	M	M
55	If applicable, the app states who holds the personal data collected in the app.	NA	M	M	M
56	The privacy policy states which kind of data is collected	M	M	M	M
57	The privacy policy states how the collected data is used	M	M	M	M
58	The privacy policy is no more than two clicks away	NTH	NTH	NTH	NTH
59	Evidence is available that the intended users were a part of the development or test phase of the app.	NA	NTH	M	M
60	Information about who the app communicates to and with is easily readable and understandable.	M	M	M	M
61	The medium for presenting the privacy policy suits the intended users.	M	M	M	M
62	The privacy policy states how it intends to use collected information.	NA	M	M	M
63	The app makes clear the it is not a replacement	M	M	M	M

	Item	Tier 1	Tier 2	Tier 3a	Tier 3b
	for a medical professional				
64	The app states which are the potential health risks of using it.	M	M	M	M
65	The app makes available previous evaluations and assessments.	NTH	NTH	NTH	NTH
66	The intended users were a part of the design/development/testing phase of the app.	M	M	M	M
67	If applicable, descriptions on users and admins and their access levels is available.	NTH	M	M	M

6.3 Limitations

Because of the time and scope of the thesis, it is possible that some articles or studies could have been missed during the search process. As the systematic literature review was carried out by one person, it means the decision whether to include an article or not were not consulted with anybody and so an article that could fit the criteria may have been discarded instead. Mistakes when extracting data from the selected studies could have happen also because only one person carried out the review.

As for the content of the thesis, a point of limitations is not including the opinions and experiences of experts of the field. In addition, the field is very small in Estonia.

7 Conclusion

Mobile apps are becoming more and more popular nowadays. The app stores are flooded with different kinds of apps. Among these are also health apps. These health apps range from calorie counting and wellbeing apps to drug calculation and chronic disease management apps. These apps have the potential to impact the user's health but also improve the patient doctor communication and data exchange for the better. Doctors could monitor or be in contact with the patient more easily and faster.

In order to choose the correct app, the app has to be evaluated or approved by someone to make sure that the app being used does not too any harm. In different countries, there are already app evaluation guidelines or frameworks. Even some organizations have taken it upon themselves to create these and evaluate apps to give the potential users a good overview of what kind of apps they should trust or not. So far nothing like this has been done in Estonia or even the European level.

In a workshop in Tallinn, Estonia on the 14th of March 2019 on the topic of "Evaluating digital health services – what is a suitable solution for Estonia?" the author of the thesis took part of, an evidence standards framework by the National Institute for Health and Care Excellence was introduced by EHIF and was used in the workshop as an evaluation tool. From a personal interview with EHIF, they expressed interest in such a guideline, the framework was taken as the basis of the guideline for Estonia.

Before such a guideline could be implemented in Estonia, first the research into the parts that the guideline consists of have to be done. In order to understand which categories an app should be evaluated on, a literature overview is done and based on the results six evaluation categories were chosen. These categories are usability, credibility, functionality, privacy & security, interoperability and transparency. These categories can help create a comprehensive overview of the app. To understand how each of these categories should be and have been evaluated, systematic reviews were done for each of the categories to map out the standards that exist. Some standards specific to Estonia were also taken into account when putting together evaluation items.

As a result of all the literature reviews, the popular choices for evaluating apps were described and brought out. Based on these a set of evaluation items were developed – an example of an evaluation guideline. The guideline is currently based on only literature, so it still needs validation and testing, but it is a step into understanding what will be necessary and what will need more development in Estonia.

7.1 Future work and recommendations

As the field is only getting bigger and stronger on the international level, it is inevitable that it grows stronger in Estonia too. This means that Estonia's take on this subject should be studied as well (be it from the developer, consumer or health care provider aspect). For future research the proposed guideline should validated and tested with the experts of the field and its potential users.

8 References

[1] J. Milward, “What Is… Mobile Health?,” February 2015. [Online]. Available: https://www.addiction-ssa.org/commentary/what-is...mobile-health. [Accessed 1 April 2019].

[2] Consard Limited, “Open stakeholder meeting on mHealth assessment guidelines: Presentations and survey,” 2016. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/open-stakeholder-meeting-mhealth-assessment-guidelines-presentations-and-survey. [Accessed 30 March 2019].

[3] World Health Organization, “From innovation to implementation – eHealth in the WHO European Region (2016),” 2016. [Online]. Available: http://www.euro.who.int/en/publications/abstracts/from-innovation-to-implementation-ehealth-in-the-who-european-region-2016. [Accessed 31 March 2019].

[4] H. M. Brown, T. Buche, C. E. Collins and M. E. Rollo, “A review of pregnancy iPhone apps assessing their quality, inclusion of behaviour change techniques, and nutrition information,” *Maternal & Child Nutrition,* vol. e12768, 2018.

[5] S. R. Stoyanov, L. Hides, D. J. Kavanagh, O. Zelenko, D. Tjondronegoro and M. Mani, “Mobile App Rating Scale: A New Tool for Assessing the Quality of Health Mobile Apps,” *JMIR Mhealth Uhealth,* vol. 3, no. 1, p. e27, 2015.

[6] B. McMillan, E. Hickey, M. G. Patel and C. Mitchell, “Quality assessment of a sample of mobile app-based health behavior change interventions using a tool based on the National Institute of Health and Care Excellence behavior change guidance,” *Patient Education and Counseling,* vol. 99, no. 3, pp. 429-435 , 2016.

[7] J. Zelmer, K. van Hoof, M. Notarianni, T. van Mierlo, M. Schellenberg and C. Tannenbaum, “An Assessment Framework for e-Mental Health Apps in Canada: Results of a Modified Delphi Process,” *JMIR Mhealth Uhealth,* vol. 6, no. 7, p. e10016, 2018.

[8] World Health Organization, “Classification of digital health interventions v1.0,” [Online]. Available: https://www.who.int/reproductivehealth/publications/mhealth/classification-digital-health-interventions/en/. [Accessed 29 March 2019].

[9] ORCHA, “ORCHA - About us,” [Online]. Available:

https://www.orcha.co.uk/about-us/. [Accessed 25 March 2019].

[10] ORCHA, “App Categories,” [Online]. Available: https://www.orcha.co.uk/about-us/app-categories/. [Accessed 25 March 2019].

[11] ORCHA, “The ORCHA review,” [Online]. Available: https://www.orcha.co.uk/our-solution/the-orcha-review/#0. [Accessed 25 March 2019].

[12] ORCHA, “App Finder,” [Online]. Available: https://appfinder.orcha.co.uk. [Accessed 25 March 2019].

[13] NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE, “Evidence standards framework for digital health technologies,” March 2019. [Online]. Available: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies. [Accessed 26 March 2019].

[14] Grey Literature Report, “What is Grey Literature?,” [Online]. Available: http://www.greylit.org/about. [Accessed 21 May 2019].

[15] French National Authority for Health, “HAS Profile,” [Online]. Available: https://www.has-sante.fr/portail/jcms/c_415964/en/has-profile?portal=c_2567632. [Accessed 26 March 2019].

[16] French National Authority for Health , “Good Practice Guidelines on Health Apps and Smart Devices (Mobile Health or mHealth),” October 2016. [Online]. Available: https://www.has-sante.fr/portail/upload/docs/application/pdf/2017-03/dir1/good_practice_guidelines_on_health_apps_and_smart_devices_mobile_health_or_mhealth.pdf. [Accessed 26 March 2019].

[17] The Royal Dutch Medical Association, “Medical App Checker: a Guide to assessing Mobile Medical Apps,” January 2016. [Online]. Available: https://www.knmg.nl/actualiteit-opinie/nieuws/nieuwsbericht/medical-app-checker-a-guide-to-assessing-mobile-medical-apps.htm. [Accessed 31 March 2019].

[18] U.S. Food and Drug Administration, 2018. [Online]. Available: https://www.fda.gov/medicaldevices/digitalhealth/mobilemedicalapplications/default.htm. [Accessed 1 April 2019].

[19] U.S. Department of Health and Human Services Food and Drug Administration, “Guidance for Industry and Food and Drug Administration Staff,” 9 February 2015. [Online]. Available: https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM263366.pdf. [Accessed 1 April 2019].

[20] R. Onodera and S. Sengoku, “Innovation process of mHealth: An overview of FDA-approved mobile medical applications,” *International Journal of Medical*

Informatics, vol. 118, pp. 65-71, 2018.

[21] The European Commission, “Green Paper on mobile health ("mHealth"),” 2014. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/green-paper-mobile-health-mhealth. [Accessed 31 March 2019].

[22] European Commission, “Summary Report on the Public Consultation on the Green Paper on Mobile Health,” 12 January 2015. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/summary-report-public-consultation-green-paper-mobile-health. [Accessed 1 April 2019].

[23] The European Commission, “Commission Staff Working Document on the existing EU legal framework applicable to lifestyle and wellbeing apps,” 10 April 2014. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/commission-staff-working-document-existing-eu-legal-framework-applicable-lifestyle-and. [Accessed 1 April 2019].

[24] European Commission, “Report of the Working Group on mHealth assessment guidelines,” 9 June 2017. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/report-working-group-mhealth-assessment-guidelines. [Accessed 1 April 2019].

[25] The European Commission, “mHealth at eHealth Week 2016,” 2016. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/mhealth-ehealth-week-2016. [Accessed 30 March 2019].

[26] Eesti Akrediteerimiskeskus, “What is accreditation and what can be accredited?,” [Online]. Available: http://www.eak.ee/index_eng.php?pageId=23. [Accessed 30 March 2019].

[27] Eesti Akrediteerimiskeskus, “Who is carrying out the accreditation?,” [Online]. Available: http://www.eak.ee/index_eng.php?pageId=25. [Accessed 30 March 2019].

[28] Eesti Akrediteerimiskeskus, “Akrediteeritud meditsiinilaboris,” [Online]. Available: http://www.eak.ee/?pageId=280. [Accessed 30 March 2019].

[29] Eesti Akrediteerimiskeskus, “EAK guidance documentation for accreditation. Procedure for accreditation assessment,” [Online]. Available: http://www.eak.ee/index_eng.php?pageId=19. [Accessed 30 March 2019].

[30] European Society of Human Reproduction and Embryology, “Guideline development process,” [Online]. Available: https://www.eshre.eu/Guidelines-and-Legal/Guidelines/Guideline-development-process. [Accessed 20 May 2019].

[31] National Institute for Health and Care Excellence, “The guidelines manual,” November 2012. [Online]. Available: https://www.nice.org.uk/process/pmg6/chapter/introduction. [Accessed 8 May

2019].

[32] R. K. Yin, Case Study Research and Applications, London: Sage Publications, Inc, 2018.

[33] B. Kitchenham, “Procedures for Performing Systematic Reviews,” Computer Science Department, Keele University, 2004.

[34] S. L. Nielsen and S. Rimpiläinen, “REPORT ON INTERNATIONAL PRACTICE ON DIGITAL APPS,” Strathprints, 2018.

[35] Q. H. Grundy, Z. W. BPharm (Hon) and L. A. Bero, “Challenges in Assessing Mobile Health App Quality: A Systematic Review of Prevalent and Innovative Methods,” *American Journal of Preventive Medicine,* vol. 51, no. 6, pp. 1051-1059, 2016.

[36] R. Gunter, S. Fernandes-Taylor, A. Mahnke, L. Awoyinka, C. Schroeder, J. Wiseman, S. Sullivan, K. Bennett, C. Greenberg and K. C. Kent, “Evaluating Patient Usability of an Image-Based Mobile Health Platform for Postoperative Wound Monitoring”.

[37] A. P. Nugraha, D. H. Syaifullah and M. A. Puspasari, “Usability Evaluation of Main Function on Three Mobile Banking Application,” in *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Bangkok, 2018.

[38] C. Reynoldson, C. Stones, M. Allsop, P. Gardner, M. I. Bennett, S. J. Closs, R. Jones and P. Knapp, “Assessing the Quality and Usability of Smartphone Apps for Pain Self-Management,” *Pain Medicine,* vol. 15, no. 6, pp. 898-909, 2014.

[39] C. Zaror, G. Espinoza-Espinoza, C. Atala-Acevedo, P. Muñoz-Millán, Y. Li, K. Clarke, J. Onetto, J. Díaz, K. Hallet, D. Manton and R. Mariño, “Validation and usability of a mobile phone application for epidemiological surveillance of traumatic dental injuries,” *Dental Traumatology,* vol. 35, no. 1, pp. 33-40, 2019.

[40] H. Cho, . P.-Y. Yen, D. Dowding, J. A. Merrill and R. Schnall, “A multi-level usability evaluation of mobile health applications: A case study,” *Journal of Biomedical Informatics,* vol. 86, pp. 79-89, 2018.

[41] M. Beauchemin, M. Gradilla, D. Baik, H. Cho and R. Schnall, “A Multi-step Usability Evaluation of a Self-Management App to Support Medication Adherence in Persons Living with HIV,” *International Journal of Medical Informatics,* vol. 122, pp. 37-44, 2019.

[42] M. Broekhuis, L. van Velsen and H. Hermens, “Assessing usability of eHealth technology: A comparison of usability benchmarking instruments,” *International Journal of Medical Informatics,* vol. 128, pp. 24-31, 2019.

[43] M. Masood and M. Thigambaram, “The Usability of Mobile Applications for Pre-schoolers,” *Procedia - Social and Behavioral Sciences,* vol. 197, pp. 1818-1826, 2015.

[44] B. Ploderer, R. Brown, L. Si Da Seng, P. A. Lazzarini and J. J. van Netten, “Promoting Self-Care of Diabetic Foot Ulcers Through a Mobile Phone App: User-Centered Design and Evaluation,” *JMIR Diabetes,* vol. 3, no. 4, p. e10105, 2018.

[45] A. Zmily, Y. Mowafi and E. Mashal, “Study of the Usability of Spaced Retrieval Exercise Using Mobile Devices for Alzheimer’s Disease Rehabilitation,” *JMIR Mhealth Uhealth,* vol. 2, no. 3, p. e31, 2014.

[46] U. Sarkar, G. I. Gourley, C. R. Lyles, L. Tieu, C. Clarity, L. Newmark, K. Singh and D. W. Bates, “Usability of Commercially Available Mobile Applications for Diverse Patients,” *Journal of general internal medicine,* vol. 31, no. 12, p. 1417–1426, 2016.

[47] J. Mirkovic, D. R. Kaufman and C. M. Ruland, “Supporting Cancer Patients in Illness Management: Usability Evaluation of a Mobile App,” *JMIR Mhealth Uhealth,* vol. 2, no. 3, p. e33, 2014.

[48] A. S. Hashim and V. J. Xing Lee, “Usability of ShopCart Among Customers at Shopping Malls,” in *2018 IEEE Conference on e-Learning, e-Management and e-Services (IC3e),*, Langkawi Island, 2018.

[49] N. A. Ismail, N. A. Jaafar, M. Ghazali, N. A. A. Majid, N. Z. I. Zamri and Y. N. Gunasegaran, “Usability study of kitchen app with multimodal interaction among beginner & intermediate cookers,” in *2017 6th ICT International Student Project Conference (ICT-ISPC),* Skudai, 2017.

[50] S. Noori, E. I. Mansor, N. Ibrahim and J. Hinds, “Promoting awareness of depression with a mobile application: A usability study and evaluation,” in *2016 4th International Conference on User Science and Engineering (i-USEr),* Melaka, 2016.

[51] B. Cassidy, G. Stringer and M. H. Yap, “Mobile Framework for Cognitive Assessment: Trail Making Test and Reaction Time Test,” in *2014 IEEE International Conference on Computer and Information Technology,* Xi'an, 2014.

[52] S. Stonbraker, H. Cho, G. Hermosi, A. Pichon and R. Schnall, “Usability Testing of a mHealth App to Support Self-Management of HIV-Associated Non-AIDS Related Symptoms,” *Studies in health technology and informatics,* vol. 250, pp. 10-110, 2018.

[53] M. Price, T. Sawyer, M. Harris and C. Skalka, “Usability Evaluation of a Mobile Monitoring System to Assess Symptoms After a Traumatic Injury: A Mixed-Methods Study,” *JMIR Mental Health,* vol. 3, no. 1, p. e3, 2016.

[54] J. Lim, G. Cloete, D. T. Dunsmuir, B. A. Payne, C. Scheffer, P. von Dadelszen, G. A. Dumont and J. M. Ansermino, "Usability and Feasibility of PIERS on the Move: An mHealth App for Pre-Eclampsia Triage," *JMIR Mhealth Uhealth,* vol. 3, no. 2, p. e37, 2015.

[55] L. L. Samson, L. Pape-Haugaard, M. C. Meltzer, M. Fuchs, H. C. Schønheyder and O. Hejlesen, "Design of a Tablet Computer App for Facilitation of a Molecular Blood Culture Test in Clinical Microbiology and Preliminary Usability Evaluation," *JMIR Mhealth Uhealth,* vol. 4, no. 1, p. e20, 2016.

[56] F. Ehrler, T. Weinhold, J. Joe, C. Lovis and K. Blondon, "A Mobile App (BEDSide Mobility) to Support Nurses' Tasks at the Patient's Bedside: Usability Study," *JMIR Mhealth Uhealth,* vol. 6, no. 3, p. e57, 2018.

[57] C. H. Teo, C. J. Ng, S. K. Lo, C. D. Lim and A. White, "A Mobile Web App to Improve Health Screening Uptake in Men (ScreenMen): Utility and Usability Evaluation Study," *JMIR Mhealth Uhealth,* vol. 7, no. 4, p. e10216, 2019.

[58] L. E. Crosby, R. E. Ware, A. Goldstein, A. Walton, N. E. Joffe, C. Vogel and M. T. Britto, "Development and evaluation of iManage: A self-management app co-designed by adolescents with sickle cell disease," *Pediatric Blood & Cancer,* vol. 64, no. 1, pp. 139-145, 2016.

[59] F. A. Le Marne, S. Butler, E. Beavis, D. Gill and A. M. E. Bye, "EpApp: Development and evaluation of a smartphone/tablet app for adolescents with epilepsy," *Journal of Clinical Neuroscience,* vol. 50, pp. 214-220, 2018.

[60] M. McMullan, "Evaluation of a medication calculation mobile app using a cognitive load instructional design," *International Journal of Medical Informatics,* vol. 118, pp. 72-77, 2018.

[61] J. Hirschey, S. Bane, M. Mansour, J. Sperber, S. Agboola, J. Kvedar and K. Jethwani, "Evaluating the Usability and Usefulness of a Mobile App for Atrial Fibrillation Using Qualitative Methods: Exploratory Pilot Study," *JMIR human factors,* vol. 5, no. 1, p. e13, 2018.

[62] R. D. Stoll, A. A. Pina, K. Gary and A. Amresh, "Usability of a Smartphone Application to Support the Prevention and Early Intervention of Anxiety in Youth," *Cognitive and Behavioral Practice,* vol. 24, no. 4, pp. 393-404, 2017.

[63] L. Zhou, J. Bao, I. M. A. Setiawan, A. Saptono and B. Parmanto, "The mHealth App Usability Questionnaire (MAUQ): Development and Validation Study," *JMIR Mhealth Uhealth,* vol. 7, no. 4, p. e11500, 2019.

[64] T. Stütz, G. Emsenhuber, D. Huber, M. Domhardt, M. Tiefengrabner, G. J. Oostingh, U. Fötschl, N. Matis and S. Ginzinger, "Mobile Phone–Supported Physiotherapy for Frozen Shoulder: Feasibility Assessment Based on a Usability Study," *JMIR rehabilitation and assistive technologies,* vol. 4, no. 2, p. e6, 2017.

[65] H. Cho, D. Powell, A. Pichon, J. Thai, J. Bruce, L. M. Kuhns, R. Garofalo and R. Schnall, “A Mobile Health Intervention for HIV Prevention Among Racially and Ethnically Diverse Young Men: Usability Evaluation,” *JMIR Mhealth Uhealth,* vol. 6, no. 9, p. e11450, 2018.

[66] R. Xu, L. Liu and J. Panneerselvam, “User Experience Evaluation of Chinese Travel App Software,” in *2014 IEEE International Conference on Computer and Information Technology*, Xi'an, 2014.

[67] F. Nayebi, J. Desharnais and A. Abran, “An Expert-Based Framework for Evaluating iOS Application Usability,” in *2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement*, Ankara, 2013.

[68] R. Schnall, H. Cho and J. Liu, “Health Information Technology Usability Evaluation Scale (Health-ITUES) for Usability Assessment of Mobile Health Technology: Validation Study,” *JMIR Mhealth Uhealth,* vol. 6, no. 1, p. e4, 2018.

[69] D. C. Mattson, “Usability assessment of a mobile app for art therapy,” *The Arts in Psychotherapy,* vol. 43, pp. 1-6, 2015.

[70] P. Bondaronek, G. Alkhaldi, A. Slee, F. L. Hamilton and E. Murray, “Quality of Publicly Available Physical Activity Apps: Review and Content Analysis,” *JMIR Mhealth Uhealth,* vol. 6, no. 3, p. e53, 2018.

[71] S. Mirzaei, B. Wilkinson and M. Wyra, “Usability Testing of VLASTA: A Vocabulary Learning and Strategy Teaching App,” in *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Wollongong, 2018.

[72] B. Metelmann, C. Metelmann, L. Schuffert, K. Hahnenkamp and P. Brinkrolf, “Medical Correctness and User Friendliness of Available Apps for Cardiopulmonary Resuscitation: Systematic Search Combined With Guideline Adherence and Usability Evaluation,” *JMIR Mhealth Uhealth,* vol. 6, no. 11, p. e190, 2018.

[73] P. S. Sullivan, R. Driggers, J. D. Stekler, A. Siegler, T. Goldenberg, S. J. McDougal, J. Caucutt, J. Jones and R. Stephenson, “Usability and Acceptability of a Mobile Comprehensive HIV Prevention App for Men Who Have Sex With Men: A Pilot Study,” *JMIR Mhealth Uhealth,* vol. 5, no. 3, p. e26, 2017.

[74] M. Hassandra, T. Lintunen, M. S. Hagger, R. Heikkinen, M. Vanhala and T. Kettunen, “An mHealth App for Supporting Quitters to Manage Cigarette Cravings With Short Bouts of Physical Activity: A Randomized Pilot Feasibility and Acceptability Study,” *JMIR Mhealth Uhealth,* vol. 5, no. 5, p. e74, 2017.

[75] B. Tiffany, P. Blasi, S. L. Catz and J. B. McClure, “Mobile Apps for Oral Health Promotion: Content Review and Heuristic Usability Analysis,” *JMIR Mhealth*

Uhealth, vol. 6, no. 9, p. e11432, 2018.

[76] Apple Inc, "Human Interface Guidelines," [Online]. Available: https://developer.apple.com/design/human-interface-guidelines/. [Accessed 18 May 2019].

[77] Google, "Design for Android," [Online]. Available: https://developer.android.com/design. [Accessed 17 May 2019].

[78] World Wide Web Consortium, "Web Content Accessibility Guidelines (WCAG) 2.1," 2018. [Online]. Available: https://www.w3.org/TR/WCAG21/. [Accessed 30 April 2019].

[79] L. Thornton, C. Quinn, L. Birrell, A. Guillaumier, B. Shaw, E. Forbes, M. Deady and F. Kay-Lambkin, "Free smoking cessation mobile apps available in Australia: a quality review and content analysis," *Australian and New Zealand Journal of Public Health,* vol. 41, pp. 625-630, 2017.

[80] R. Grainger, H. Townsley, B. White, T. Langlotz and W. J. Taylor, "Apps for People With Rheumatoid Arthritis to Monitor Their Disease Activity: A Review of Apps for Best Practice and Quality," *JMIR Mhealth Uhealth,* vol. 5, no. 2, p. e7, 2017.

[81] F. Modave, J. Bian, T. Leavitt, J. Bromwell, C. Harris and H. Vincent, "Low Quality of Free Coaching Apps With Respect to the American College of Sports Medicine Guidelines: A Review of Current Mobile Apps," *JMIR Mhealth Uhealth,* vol. 3, no. 3, p. e77, 2015.

[82] B. Xie, Z. Su, W. Zhang and R. Cai, "Chinese Cardiovascular Disease Mobile Apps' Information Types, Information Quality, and Interactive Functions for Self-Management: Systematic Review," *JMIR Mhealth Uhealth,* vol. 5, no. 12, p. e195, 2017.

[83] K. Santo, S. S. Richtering, J. Chalmers, A. Thiagalingam, C. K. Chow and J. Redfern, "Mobile Phone Apps to Improve Medication Adherence: A Systematic Stepwise Process to Identify High-Quality Apps," *JMIR Mhealth Uhealth,* vol. 4, no. 4, p. e132, 2016.

[84] D. Bakker, N. Kazantzis, D. Rickwood and N. Rick, "Development and Pilot Evaluation of Smartphone-Delivered Cognitive Behavior Therapy Strategies for Mood- and Anxiety-Related Problems: MoodMission," *Cognitive and Behavioral Practice,* vol. 25, no. 4, pp. 496-514 , 2018.

[85] E. de Korte, N. Wiezer, M. B. Roozeboom, P. Vink and W. Kraaij, "Behavior Change Techniques in mHealth Apps for the Mental and Physical Health of Employees: Systematic Assessment," *JMIR Mhealth Uhealth,* vol. 6, no. 10, p. e167, 2018.

[86] D. Crane, C. Garnett, J. Brown, R. West and S. Michie, “Behavior Change Techniques in Popular Alcohol Reduction Apps: Content Analysis,” *Journal of medical Internet research,* vol. 17, no. 5, p. e118, 2015.

[87] B. Richardson, J. Dol, K. Rutledge, J. Monaghan, A. Orovec, K. Howie, T. Boates, M. Smit and M. Campbell-Yeo, “Evaluation of Mobile Apps Targeted to Parents of Infants in the Neonatal Intensive Care Unit: Systematic App Review,” *JMIR Mhealth Uhealth,* vol. 7, no. 4, p. e11620, 2019.

[88] C. Escoffery, R. McGee, J. Bidwell, C. Sims, E. K. Thropp, C. Frazier and E. D. Mynatt, “A review of mobile apps for epilepsy self-management,” *Epilepsy & Behavior,* vol. 81, pp. 62-69 , 2018.

[89] M. Bardus, A. Ali, F. Demachkieh and G. Hamadeh, “Assessing the Quality of Mobile Phone Apps for Weight Management: User-Centered Study With Employees From a Lebanese University,” *JMIR Mhealth Uhealth,* vol. 7, no. 1, p. e9836, 2019.

[90] B. Y. Kim, A. Sharafoddini, N. Tran, E. Y. Wen and J. Lee, “Consumer Mobile Apps for Potential Drug-Drug Interaction Check: Systematic Review and Content Analysis Using the Mobile App Rating Scale (MARS),” *JMIR Mhealth Uhealth,* vol. 6, no. 3, p. e74, 2018.

[91] A. Salazar, H. de Sola, I. Failde and J. A. Moral-Munoz, “Measuring the Quality of Mobile Apps for the Management of Pain: Systematic Search and Evaluation Using the Mobile App Rating Scale,” *JMIR Mhealth Uhealth,* vol. 6, no. 10, p. e10718, 2018.

[92] R. M. Masterson Creber, M. S. Maurer, M. Reading, G. Hiraldo, K. T. Hickey and S. Iribarren, “Review and Analysis of Existing Mobile Phone Apps to Support Heart Failure Symptom Monitoring and Self-Care Management Using the Mobile Application Rating Scale (MARS),” *JMIR Mhealth Uhealth,* vol. 4, no. 2, p. e74, 2016.

[93] M. Mani, D. J. Kavanagh, L. Hides and S. R. Stoyanov, “Review and Evaluation of Mindfulness-Based iPhone Apps,” *JMIR Mhealth Uhealth,* vol. 3, no. 3, p. e82, 2015.

[94] Y. K. Choi, G. Demiris, S.-Y. Lin, S. J. Iribarren, C. A. Landis, H. J. Thompson, S. M. McCurry, M. M. Heitkemper and T. M. Ward, “Smartphone Applications to Support Sleep Self-Management: Review and Evaluation,” *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine,* vol. 14, no. 10, p. 1783–1790, 2018.

[95] K. Anderson, O. Burford and L. Emmerton, “App Chronic Disease Checklist: Protocol to Evaluate Mobile Apps for Chronic Disease Self-Management,” *JMIR research protocols,* vol. 5, no. 4, p. e204, 2016.

[96] T. S. Fernandes, Á. F. Moreira and É. Cota, "EPE-Mobile—A framework for early performance estimation of mobile applications," *Software Practice and Experience,* vol. 48, p. 85– 104, 2017.

[97] S. M. Coulon, C. M. Monroe and D. S. West, "A Systematic, Multi-domain Review of Mobile Smartphone Apps for Evidence- Based Stress Management," *American Journal of Preventive Medicine,* vol. 51, no. 1, pp. 95-105, 2016.

[98] K. N. DiFilippo, W. Huang and K. M. Chapman-Novakofski, "A New Tool for Nutrition App Quality Evaluation (AQEL): Development, Validation, and Reliability Testing," *JMIR Mhealth Uhealth,* vol. 5, no. 10, p. e163, 2017.

[99] W. Brown, P.-Y. Yen, M. Rojas and R. Schnall, "Assessment of the Health IT Usability Evaluation Model (Health-ITUEM) for evaluating mobile health (mHealth) technology," *Journal of Biomedical Informatics,* vol. 46, no. 6, pp. 1080-1087, 2013.

[100] A. Idri, M. Bachiri, J. L. Fernández-Aleman and A. Toval, "ISO/IEC 25010 Based Evaluation of Free Mobile Personal Health Records for Pregnancy Monitoring," in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC),* Turin, 2017.

[101] Xcertia, "App Operability Guidelines & Survey," 2018. [Online]. Available: https://xcertia.org/app-operability-survey/ . [Accessed 10 May 2019].

[102] National Institute for Health and Care Excellence, "Digital Assessment Questions – Beta," August 2018. [Online]. Available: https://developer.nhs.uk/digital-tools/daq/. [Accessed 11 May 2019].

[103] J. Muchagata and A. Ferreira, "Translating GDPR into the mHealth Practice," in *2018 International Carnahan Conference on Security Technology (ICCST),* 2018.

[104] L. Hutton, B. Price, R. Kelly, C. McCormick, A. Bandara, T. Hatzakis, M. Meadows and B. Nuseibeh, "Assessing the Privacy of mHealth Apps for Self-Tracking: Heuristic Evaluation Approach," *JMIR Mhealth Uhealth,* vol. 6, no. 10, p. e185, 2018.

[105] K. Huckvale, J. T. Prieto, M. Tilney, P.-J. Benghozi and J. Car, "Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment," *BMC Medicine,* vol. 13, no. 1, 2015.

[106] L. Parker, V. Halter, T. Karliychuk and Q. Grundy, "How private is your mental health app data? An empirical study of mental health app privacy policies and practices," *International Journal of Law and Psychiatry,* vol. 64, pp. 198-204, 2019.

[107] S. Safavi and Z. Shukur, "Conceptual Privacy Framework for Health Information on Wearable Device," *PloS one,* vol. 9, no. 12, p. e114306, 2014.

[108] B. C. Zapata, A. H. Niñirola, J. L. Fernández-Alemán and A. Toval, "Assessing the privacy policies in mobile personal health records," in *014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, 2014.

[109] A. M. Khattak, F. Iqbal, P. C. K. Hung, J.-S. Sun, . G.-P. Pan and J.-J. Lin, "Privacy Requirements for mobile e-Service in the Health Authority - Abu Dhabi (HAAD)," in *2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)*, Barcelona, 2016.

[110] M. Hussain, A. Al-Haiqi, A. Zaidan, B. Zaidan, M. Kiah, S. Iqbal, S. Iqbal and M. Abdulnabi, "A security framework for mHealth apps on Android platform," *Computers & Security*, vol. 75, pp. 191-217, 2018.

[111] K. O'Loughlin, M. Neary, E. C. Adkins and S. M. Schueller, "Reviewing the data security and privacy policies of mobile apps for depression," *Internet Interventions*, vol. 15, pp. 110-115, 2019.

[112] J. M. Robillard, T. L. Feng, A. B. Sporn, J.-A. Lai, C. Lo, M. Ta and R. Nadler, "Availability, readability, and content of privacy policies and terms of agreements of mental health apps," *Internet Interventions*, no. 19, p. 100243, September.

[113] Republic of Estonia Government Office, Ministry of Social Affairs, "Uuring „Kolmandate osapoolte E-tervise rakenduste valideerimise korralduse mõjuanalüüs ja uuring tervise infosüsteemiga liidestumiseks"," Tallinn, 2017.

[114] Health Level Seven International, "Version History since DSTU #1," 27 December 2018. [Online]. Available: https://www.hl7.org/fhir/summary.html. [Accessed 22 May 2019].

[115] Health Level Seven International, "Mappings to Other Standards," 27 December 2018. [Online]. Available: https://www.hl7.org/fhir/mappings.html. [Accessed 22 May 2019].

[116] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane and R. B. Ramoni, "SMART on FHIR: a standards-based, interoperable apps platform for electronic health records," *Journal of the American Medical Informatics Association : JAMIA*, vol. 23, no. 5, pp. 899-908, 2016.

[117] E. Rossi, P. Fontelo, M. J. Ackerman, G. Pozzi and S. Marceglia, "A Prototype of Mobile app/EHR Communication through Standards for Home Treatment of Transcranial Direct Current Stimulation," in *2015 International Conference on Healthcare Informatics*, Dallas, 2015.

[118] P. Plastiras and D. O'Sullivan, "Exchanging personal health data with electronic health records: A standardized information model for patient generated health data and observations of daily living," *International Journal of Medical Informatics*, vol. 120, pp. 116-125, 2018.

[119] P. B. Pfiffner, I. Pinyol, M. D. Natter and K. D. Mandl, “C3-PRO: Connecting ResearchKit to the Health System Using i2b2 and FHIR,” *PLoS One,* vol. 11, no. 3, p. e0152722, 2016.

[120] A. Adamkó, A. Garai and I. Péntek, “Review and requirement specification of telemedicine systems interoperability with common data exchange hub,” in *2016 IEEE 14th International Symposium on Intelligent Systems and Informatics (SISY),* Subotica, 2016.

[121] S. Sinha, M. Jensen, S. Mullin and P. L. Elkin , “Safe Opioid Prescription: A SMART on FHIR Approach to Clinical Decision Support,” *Online journal of public health informatics,* vol. 9, no. 2, p. e193, 2017.

[122] J. L. Warner, M. J. Rioth, K. D. Mandl, J. C. Mandel, D. A. Kreda, I. S. Kohane, D. Carbone, R. Oreto, L. Wang, S. Zhu, H. Yao and G. Alterovitz, “SMART precision cancer medicine: a FHIR-based app to provide genomic information at the point of care,” *Journal of the American Medical Informatics Association : JAMIA,* vol. 23, no. 4, p. 701–710, 2016.

[123] R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun and M. D. Wang, “Intelligent Mortality Reporting With FHIR,” *IEEE Journal of Biomedical and Health Informatics,* vol. 22, no. 5, pp. 1583-1588, 2018.

[124] Y. Aliakbarpoor, S. Comai and G. Pozzi, “Designing a HL7 compatible personal health record for mobile devices,” in *2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI),* Modena, 2017.

[125] Y. K. R. Sánchez, S. A. Demurjian and M. S. Baihan, “Achieving RBAC on RESTful APIs for Mobile Apps Using FHIR,” in *2017 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud),* San Fransisco, 2017.

[126] Q. Grundy, K. Chiu, F. Held, A. Continella, L. Bero and R. Holz, “Data sharing practices of medicines related apps and the mobile ecosystem: traffic, content, and network analysis,” *BMJ,* vol. 364, p. l920, 2019.

[127] R. Fahy, J. Van Hoboken and N. Van Eijk, “Data Privacy, Transparency and the Data-Driven Transformation of Games to Services,” in *2018 IEEE Games, Entertainment, Media Conference (GEM),* Galway, 2018.

[128] JMIR Publications, “Works citing "Mobile App Rating Scale: A New Tool for Assessing the Quality of Health Mobile Apps",” [Online]. Available: https://mhealth.jmir.org/2015/1/e27/citations . [Accessed 21 May 2019].

[129] International Organization for Standardization, “ISO 9241-11:2018(en) Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts,” March 2018. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en. [Accessed 21 May

2019].

[130] Trinidad Consulting OÜ, “Kasutatavuse mõõdikute süsteem avaliku sektori tarkvarasüsteemidele,” [Online]. Available: https://www.mkm.ee/sites/default/files/kasutatavuse_moodikute_susteem_final_november_2014.pdf. [Accessed 19 May 2019].

[131] J. Sauro, “MEASURING USABILITY WITH THE SYSTEM USABILITY SCALE (SUS),” 2 February 2011. [Online]. Available: https://measuringu.com/sus/. [Accessed 14 May 2019].

[132] A. Bangor, P. Kortum and J. Miller, “Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale,” *Journal of Usability Studies,* vol. 4, no. 3, pp. 114-123 , 2009.

[133] A. Bangor, P. T. Kortum and J. T. Miller, “An Empirical Evaluation of the System Usability Scale,” *Intl. Journal of Human–Computer Interaction,* vol. 24, no. 6, pp. 574-594, 2008.

[134] J. Hsu, D. Liu, Y. M. Yu, H. T. Zhao, Z. R. Chen, J. Li and W. Chen, “The Top Chinese Mobile Health Apps: A Systematic Investigation,” *Journal of medical Internet research,* vol. 18, no. 8, p. e222, 2016.

[135] European Commission, “Code of Conduct on privacy for mHealth apps has been finalised,” 2016. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/code-conduct-privacy-mhealth-apps-has-been-finalised. [Accessed 30 March 2019].

[136] European Commission, “Privacy Code of Conduct on mobile health apps,” 2018. [Online]. Available: https://ec.europa.eu/digital-single-market/en/privacy-code-conduct-mobile-health-apps%20. [Accessed 1 April 2019].

[137] Andmekaitse inspektsioon, “Andmeturve,” 18 February 2014. [Online]. Available: https://www.aki.ee/et/avalik-teave/andmeturve. [Accessed 15 May 2019].

[138] European Commission, “Guidelines on Transparency under Regulation 2016/679 (wp260rev.01),” 22 August 2018. [Online]. Available: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227. [Accessed 21 May 2019].

Appendix 1 – Evidence tiers. Source National Institute for Health and Care Excellence [13]

Tier nr	Description	Examples
Tier 1	Apps with potential system benefits, no direct user benefits.	Electronic health record systems, ward management systems.
Tier 2	Apps that just provide information without collecting it	Lifestyle apps, encyclopaedia style apps about certain diseases
	Apps that collect simple information to create a health diary of sorts without sharing it with any other person or app.	Health diaries, mood tracker apps, apps that connect to a fitness wearable to count steps for example
	Apps that are able to communicate with a second party (other users, health care physicians etc.). The app does not provide clinical advice but given the opportunity for physicians to give medical advice through the app. The app does not provide clinical content.	Messaging apps, communication apps
Tier 3a	Apps that intend to change the health-related behaviour of the user. Recommended or prescribed by a professional	Apps that are used for weight loss
	Apps that aim to help the user self-manage a condition. Can also be connected to a healthcare professional who can then see this information	App that records health-related data and may be able to send this data to healthcare professionals.
Tier 3b	Provides treatment or guides treatment decisions	Apps that provide steps or guides on how to treat a specific disease. May be aimed towards clinicians instead of patients.
	Apps that are capable of automatically recording health related information and send this to health care professionals without being prompted by the user with the aim of remote monitoring.	