Final version available at ScienceDirect : http://dx.doi.org/10.1016/j.jbi.2017.09.002

Combining Multivariate Statistics and the Think-Aloud Protocol to Assess Human-Computer Interaction Barriers in Symptom Checkers

Luis Marco-Ruiz[a, b], Erlend Bønes[a], Estela de la Asunción[a], Elia Gabarrón[a, b], Juan Carlos Aviles-Solis[c], Eunji Lee[d], Vicente Traver[e], Keiichi Sato[f, g], Johan G. Bellika[a, b]

[a] Norwegian Centre for E-health Research, University Hospital of North Norway, P.O. Box 35, N-9038 Tromsø, Norway

[b] Department of Clinical Medicine, Faculty of Health Sciences, UIT The Arctic University of Norway, 9037 Tromsø, Norway

[c] Department of Community Medicine, Faculty of Health Sciences, UIT The Arctic University of Norway, 9037 Tromsø, Norway

[d] SINTEF, Forskningsveien 1, 0373 Oslo, Norway

[e] Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

[f] Institute of Design, Illinois Institute of Technology, 565 West Adams Street, Chicago, IL 60661, US

[g] Department of Computer Science, UIT The Arctic University of Norway, 9037 Tromsø, Norway

Abstract

Symptom checkers are software tools that allow users to submit a set of symptoms and receive advice related to them in the form of a diagnosis list, health information or triage. The heterogeneity of their potential users and the number of different components in their user interfaces can make testing with end-users unaffordable. We designed and executed a two-phase method to test the respiratory diseases module of the symptom checker Erdusyk. Phase I consisted of an online test with a large sample of users (n = 53). In Phase I, users evaluated the system remotely and completed a questionnaire based on the Technology Acceptance Model. Principal Component Analysis was used to correlate each section of the interface with the questionnaire responses, thus identifying which areas of the user interface presented significant contributions to the technology acceptance. In the second phase, the think-aloud procedure was executed with a small number of samples (n = 15), focusing on the areas with significant contributions to analyze the reasons for such contributions. Our method was used effectively to optimize the testing of symptom checker user interfaces. The method allowed kept the cost of testing at reasonable levels by restricting the use of the think-aloud procedure while still assuring a high amount of coverage. The main barriers detected in Erdusyk were related to problems understanding time repetition patterns, the selection of levels in scales to record intensities, navigation, the quantification of some symptom attributes, and the characteristics of the symptoms.

1. Introduction

Consumer-oriented Clinical Decision Support Systems (CDSSs) are software systems that aim to help information consumers making informed decisions about their health [1]. With shared decision making on the agendas of many health organizations [2–4] and an increasing number of patients who are willing to be involved in their own health decisions [5], consumer-oriented CDSSs can be an effective tool to enable patient empowerment, thus allowing patients to become active participants in decisions about their healthcare and, at the same time, allowing them to make sensible use of healthcare resources. Among the different types of existing consumer-oriented CDSSs [1], symptom checkers allow patients to register a set of symptoms and receive a list of possible diagnoses or advice about what actions might be appropriate to perform (self-triage) [6]. The first symptom checkers were static websites or CD-based applications [7], and they were not widely deployed by health trusts. However, with an increasing pressure on primary care, and studies showing that up to 50% of the visits to a general practitioner's (GP) office were unnecessary [8][9] and up to 70% were minor health incidents [10], consumer CDSSs, and particularly symptom checkers, have caught the attention of health organizations. Nowadays, several health organizations have started using symptom checkers to develop broad diagnostic and self-triage systems to guide each patient to the most appropriate action[11–16]. For example, the symptom checkers offered by the Mayo Clinic [14] and WebMD [16] provide information about the possible diseases linked to the symptoms reported by the patient. The British NHSDirect provides a more self-triage oriented service that combines a web application for patients to report symptoms with a call center where nurses provide advice. The appropriate use of symptom checkers can have a significant impact both on patient health and health organizations [6]. Regarding patient health, a symptom checker can help patients to perform self-care, avoiding unnecessary medical attention [8](e.g. visits can be managed by consulting with a pharmacist) [8], or it can help them to access and process health information rather than search Google, thus avoiding the problems involved in consulting raw information with different quality and technical levels [17][18]. Regarding health organizations, symptom checkers relieve the pressure of unnecessary visits by guiding patients to the appropriate health circuit. For example, in 2011, NHSDirect avoided 1.5 million unnecessary surgery appointments and 0.7 million emergency calls [15,19]. Although more evaluations are needed, recent studies have indicated that investments in web-based symptom checkers already have good outcomes for emergency cases but need improvement in non-emergency and

self-care cases [6,20]. This is interesting, since the investment needed to develop them is moderate compared to other health interventions. For example, Elliot et al. reported that the accuracy of web-based symptom checkers and telephone triage lines are comparable [21].

However, when direct human support is not provided by these systems, the appropriate communication of health information by the user is paramount, so the system provides appropriate guidance. This involves a challenge in the design of inquiry methods and user interfaces for symptom checkers since health information usually contains clinical terms, quantitative measures and time patterns [22] that users need to understand to provide accurate communication about their health conditions. In fact, little is known about how patients understand health information [1] or how patients perceive their conditions in contrast to how health professionals characterize and see them [3]. Therefore, assumptions about general user interface design cannot be readily applied and metrics for symptom recording Graphical User Interfaces (GUI) still need to be established. This makes the design and evaluation of each symptom checker´s user interface a unique process. That evaluation needs to effectively assess how successful the system is in communicating the clinical concepts that patients must understand to accurately communicate their health information. In fact, there may be many differences among users and many may have problems interpreting their health information considering that only 30% to 60% of citizens are health literate [23]. How successful that communication is will be the main factor influencing how accurate the system is in providing advice to the patient. Otherwise, even with advanced recommendation algorithms, if poor quality information is provided, the system will end up in a "garbage in, garbage out" situation. In such cases, a consumer CDSS may mislead the user rather than provide support for health related decision-making, driving some of them to increase unnecessary GP visits, or worse, advise others to perform self-care when they may be suffering a life-threatening condition. Therefore, besides measuring design usability flaws, techniques to evaluate Human Computer Interaction (HCI) between users and CDSSs are needed to determine if a cognitive gap exists between the clinical concepts that the GUI exposes and the user´s interpretation of the information requested. Only when that gap is minimized will it be effective and safe to deliver a symptom checker.

2. Background

2.1. Context: The symptom checker *Erdusyk*

Nowadays, most symptom checkers are in their first generation, meaning that they use an algorithm to diagnose or perform triage, but they still do not use information from external services (such as epidemiological ones) to improve their accuracy [6]. In North Norway, the symptom checker Erdusyk (in English, Are You Ill?) has been running since 2012 [24]. Erdusyk has evolved from this first generation of symptom checkers by introducing algorithms that leverage data provided by the patient (symptoms, demographics, etc.) and data from the incidence of gastrointestinal and respiratory infectious diseases datasets extracted from regional laboratory information systems [25]. By combining both, the system provides users with a list of the probabilities of the diseases that may be affecting them. This way they can access quality information to decide whether it is appropriate to perform self-care or that they need to visit their GP.

Recently, Norway has promoted a national initiative to evaluate openEHR and SNOMED-CT to enable the interoperability of clinical data across electronic health records [26,27,27–29]. As a consequence, the next version of Erdusyk should use Clinical Information Models (CIMs) to structure the information recorded by the patient [30] defined as openEHR archetypes. In addition, the system uses SNOMED-CT as clinical terminology [31].

To adapt Erdusyk to the new national scene and develop it into a second-generation symptom checker that can represent information using archetypes, we have accomplished several tasks. First, we have redefined its architecture to deal with archetypes [32]; second, we have used the national knowledge management center to drive the definition of archetypes for the new Virtual Medical Record (VMR) [31]; and third, we have developed data integration strategies to enable the secondary use of data from the laboratory information system in its inference engine [33]. The study was performed when the combination of different system components was being performed; therefore the interaction with the user had to be evaluated (user-task-system evaluation) [34]. According to the classification proposed by Yen and Bakken, this situates Erdusyk in Stage 3 of the development cycle, where aspects such as perception, acceptance, accuracy, and learnability must be evaluated in a laboratory setting [34]. This evaluation is of paramount importance since it will detect if there are significant usability barriers that will prevent users from using Erdusyk appropriately to record their symptoms. Specifically, this will determine the number of features from archetypes that the user is able to submit and will therefore determine which features from archetypes can be used by the symptom checker´s new algorithm. Figure 1 illustrates the archetype and medical ontology containing the medical concepts that are requested by Erdusyk´s user interface, and, on the right side, the cloud representing the cognitive process that users go through in order to understand those medical concepts.

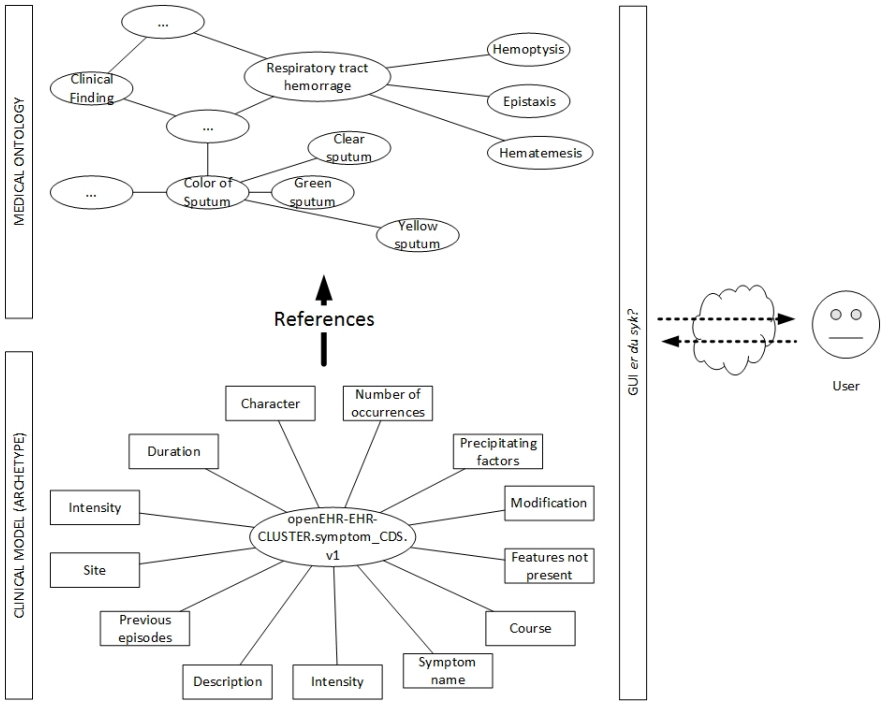

Figure 1. Schema of the medical ontology and clinical model that the user needs to populate to feed the CDSS.

2.2. Usability testing of CDSSs

Usability testing encompasses the evaluation of several dimensions that determine how well a software system can be understood, learned, and used and be attractive to the user [35]. The study of the cognitive process the user goes through when performing a task with the system is covered by the dimension that evaluates how well the system is understood. In symptom checkers, this concerns the identification and understanding of HCI barriers during the symptom recording process. Many techniques, including those performed by both experts and end-users, are available for usability testing in healthcare. Techniques such as cognitive task analysis, heuristic evaluation, and cognitive walkthrough involve testing with expert evaluators that examine the system while it performs some tasks to unveil usability problems [36]. Other methods involve end-users to test the system and perform objective and subjective measurements while they are using the system [34]. Examples of objective measurements can be eye-tracking or the time required to finish a task; examples of subjective measures can be interviews about the system or questionnaires that evaluate different parts of the system. Currently, standards such as ISO9241 cover usability and ergonomic aspects.

In the field of CDSS usability testing, mixed techniques have been proposed that combine two or more different types of techniques to improve the accuracy of tests and avoid bias. For example, Boland et al. [37] proposed a complete testing methodology with two main phases. The first

phase performed a cognitive walkthrough that compared the tested system with a previously selected reference system, and the second phase applied the think-aloud procedure and usability evaluation questionnaires [37]. Van Engen-Verheul also proposed a mixed method that 1) applies the think-aloud procedure to measure usability problems, and 2) analyzes interviews to measure deviations from the system´s predefined data entry. Li et al. proposed a method that combines the think-aloud procedure with near live scenarios to test a CDSS for primary care [38]. Davis and Jiang proposed a mixed analysis of a CDSS for people with diabetes by combining objective measurements (e.g. number of errors and completion time) with subjective measurements from usability questionnaires[39]. Lai et al. combined a heuristics evaluation with the think-aloud procedure to test a patient-oriented CDSS to prevent depression in chronically ill patients [40]. Although many of these techniques have been successfully used to test CDSSs oriented to clinical users or even chronic patients, they are not optimized to detect HCI barriers present in symptom checkers´ interfaces. There are two main factors that make testing of symptom checkers different from CDSSs oriented to clinical users: 1) expert methods only are not applicable provided that the end-users understanding of the system´s interface needs to be carefully assessed to avoid negative outcomes, and 2) symptom checker GUIs are very large and contain many different execution paths. This makes the cost of testing in controlled environments with end-users very high.

In the case of symptoms checkers, if an evaluation is performed only by an expert, it may bias the usability problems that are related to the understanding of clinical terms, temporality measures, and so on. Therefore, in this situation, testing with real users is necessary in order to understand the cognitive process that users go through when using a new system to record health data. A widely accepted technique that is based on the study of cognitive science is the think-aloud procedure [41,42]. Although some of the studies that have tested CDSSs used the think-aloud procedure as the gold standard for usability testing [43], this was done with a small sample of end-users that already had experience in the business process that the system covered. Even in the case of patient-oriented systems, their users are limited to a subset of one particular chronic disease and therefore they have a priori knowledge of the parameters that they need to submit to monitor their condition [39,40]. However, symptom checkers have a much more heterogeneous group of users accessing the system. Some of those users may have higher health literacy and experience in recording online information and some may have very little or no experience. This diversity of users, added to the size of a symptom checker´s interfaces, would require a large sample. However, the cost of testing with end-users in controlled environments may disallow the use of the think-aloud procedure for large samples. This is especially relevant in the case of Erdusyk, where an archetype-driven GUI contains a

high number of symptoms that have many details and paths that users may follow when they record symptoms. The archetype contains 14 elements per symptom. The respiratory disease module alone contains 9 symptoms. This involves 126 different sections, each of with have several subsections that can be covered by following several execution paths with different combinations.

To cope with this situation, we present a two-phase method for testing symptom checker interfaces with a high number of variables and execution paths. The first phase is oriented to detect which parts of the system present problems. This phase is performed through a freely accessible online system on the Internet where anyone can record a set of symptoms by answering a Technology Acceptance Model (TAM)-based questionnaire [44,45] to provide an evaluation of it. This allows for testing the system with a large sample of end-users without making them move to a controlled usability laboratory. The result of this phase is a set of areas where there are significant contributions to the users´ technology acceptance. In the second phase, that knowledge can be used to optimize the think-aloud procedure [42] since only a fraction of end-users are needed to cover the evaluation of significant areas. This strategy aims to keep the cost of the procedure at reasonable levels by restricting the use of the think-aloud procedure to key areas while keeping the study robust with a large sample size, providing an appropriate coverage of the interface.

3. Materials and Methods

3.1. Principal Component Analysis

In usability studies, it is common to have a large number of independent (explanatory) variables corresponding to the different sections of the GUI, the characteristics of the users, and so on. In addition, it is common to have a response variable (e.g., usefulness perception, ease of use, efficiency, etc.) observed indirectly through questionnaires or scales composed of several items [36]. Therefore, the response is a latent variable observed through correlated variables (e.g., items in a questionnaire). These situations complicate data analysis. In scenarios where there are many correlated variables, “classical” statistical analysis (e.g., ANOVA) cannot be reliably applied since multicollinearity (inter-correlated variables) may lead to imprecise estimation of the effects of variables in the response variable and unstable estimation of the model´s parameters. In these cases, multivariate statistics can be beneficial since they provide optimal methods for visualizing the latent variable, dealing with multicollinearity and studying

the effect of the users´ different observations over the variance of the response variables. Among the different multivariate techniques available, Principal Components Analysis (PCA) deals with a large number of correlated variables that refer to the same underlying observed phenomena (e.g., usability)[46]. PCA is a dimensionality reduction technique that allows the representation of data described by a large number of variables, possibly correlated, as projections into a reduced set of linearly independent vectors known as principal components (PCs). Intuitively, PCA allows the separation of the true structure of data from random variation, concentrating the data structure in a few PCs [47]. When PCA is performed, a PC for each of the variables is estimated. The first PC is the one that represents the direction with the largest variance of data (see PC1 in Figure 2). The second PC corresponds to the second largest direction of variance, which is orthogonal to the first one (see PC2 in Figure 2)[48]. The following PCs correspond to the following directions of variance that are orthogonal to the previous ones [48]. Since PCs are orthogonal, there is no correlation among them. The proper selection of a minimal set of PCs allows for representing the observations in a reduced dimensional space, thus facilitating the visualization and analysis of complex multidimensional datasets. Figure 2 shows a minimal example where observations described by three variables are projected into a dimensional space defined by the two PCs that better summarize the direction of variance of the observations in the original 3D space.

To reduce the dimensionality, PCA only retains those PCs that explain the largest proportion of the total variance. A common method for selecting which PCs to retain is the Elbow method [47]. The Elbow method plots the eigenvalues (which correspond to the proportion of variance) vs. each PC and establishes that the PCs to retain are those prior to the change in the slope. Figure 3 shows another example where only the first two PCs from a total of five PCs are retained using the Elbow method.

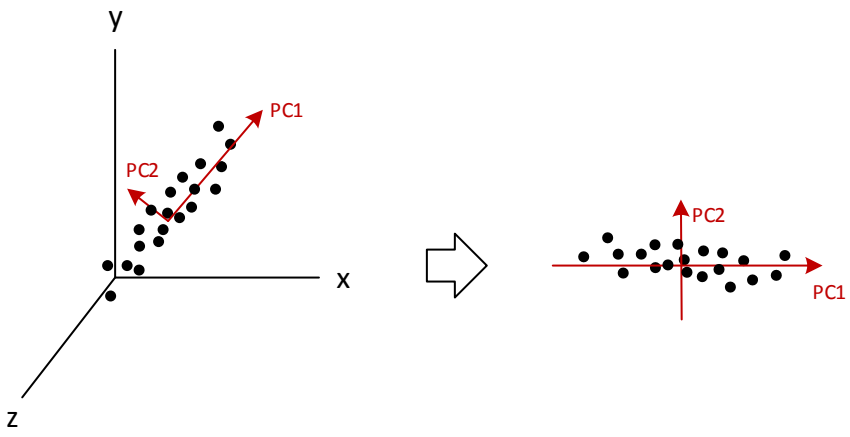

Figure 2. Dimensionality reduction with PCA.

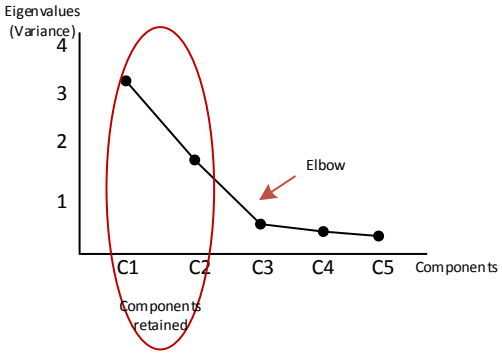

Figure 3. Selection of components with the "Elbow" method.

Beyond data visualization, PCA opens the door to applying statistical analysis, such as multiple regression or ANOVA [47]. PCA can be used as a previous step for regression in order to summarize correlated variables into a few orthogonal PCs. Since PCs are linearly independent, then regression can be applied to explain or predict the variation of observations across these PCs from a set of explanatory variables (e.g. users´ characteristics, GUI section etc.).

3.2. Usability methods

This study makes use of two well-established usability evaluation techniques:

TAM: TAM is a theoretical model that was developed to measure perceived usability [44]. TAM has two main blocks, which are related to usefulness perception and ease of use. TAM has been extensively used in many sectors to measure technology acceptance. Over the years, several extensions have been developed to include new factors that complement the measurement of technology acceptance [49,50]. Nowadays, TAM has been extensively used in many sectors, including Healthcare [45].

Think-Aloud: Think-aloud is a procedure that originated in cognitive psychology and was adapted to provide usability researchers with insights into the participant´s mental process when using a system [41]. When compared to expert-based examination, the think-aloud procedure allows the detection of more severe and recurring problems than expert-based methods [36]. In addition, it allows researchers to understand the reason for the problem directly from the user´s perspective. In this study, we used concurrent think-aloud because it is preferred version of the think-aloud procedure for diagnosing usability problems [36]. The think-aloud procedure can be complemented with retrospective interviews where the issues raised during the session are analyzed with the user. Interviewing the participants after the think-aloud procedure sessions provides a mean for deeper engagement with them compared to regular user observation. This

mechanism of revisiting problematic or noteworthy events allows both participants and researchers to examine and validate their interpretation and evaluation of the process collaboratively. The main drawbacks of the think-aloud procedure are its high cost and that it only reveals usability problems that intersect with the users.

3.3. Methodology

3.3.1. Overview

A two-phase methodology was designed to detect and understand the causes of HCI barriers in the Erdusyk interface. Phase I is concerned with detecting which sections have a significant positive or negative contribution to technology acceptance. Phase II concentrates on the execution of the think-aloud procedure in those significant areas to understand why their contribution is significant. Figure 4 shows the stages of the methodology.

Phase I (detection in Figure 4) aimed to deal with the large number of sections and possible execution paths and possible combinations of the GUI components, and to identify sections with significant contributions. For this, we designed a reduced cost study that was performed online with a large sample size (aiming for n=100) to guarantee the appropriate coverage of the interface. In Phase I, users went through the application freely, recording some symptoms of their choice and completing a TAM-based questionnaire at the end. Provided that the number of responses to the questionnaire measuring technology acceptance was not only one, PCA was used to reduce the seven response variables to only two PCs that summarized technology acceptance (TAM_PC) and familiarity of vocabulary (VOC_PC) respectively. These two variables were regressed with the variables that represented the symptoms, demographics, and other data provided by the users. This regression identified which of those variables were leading to significant negative or positive contributions to TAM_PC or VOC_PC.

Phase II (analysis in Figure 4) aimed to analyze the causes of those significant contributions with a more in-depth study of a smaller sample. A think-aloud procedure was executed that gave participants a set of vignettes that focused on the areas with significant contributions to TAM_PC or VOC_PC. The result allowed us to understand why the variables detected in Phase I had a negative or positive contribution to technology acceptance in order to establish directions of work to solve HCI barriers.

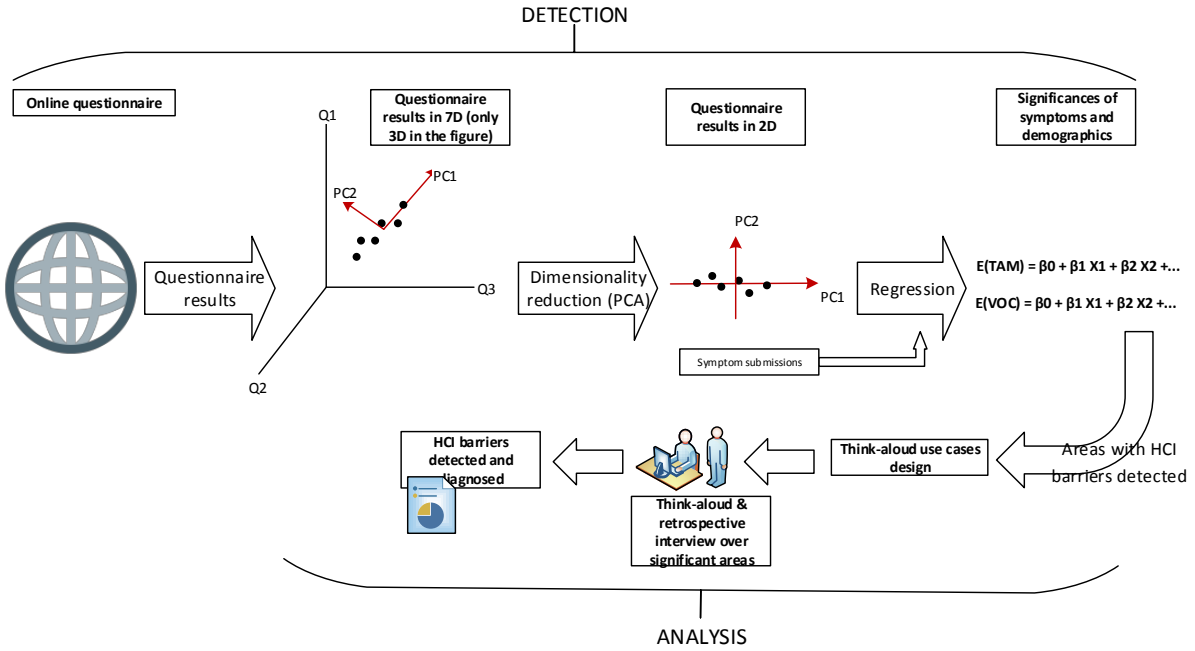

Figure 4. Methodology workflow.

3.3.2. Phase I: Problem detection

The problem detection phase consisted of a study performed with a large sample of citizens who tested the system and answered a subset of six questions adapted from TAM [45], plus one additional question that referred to the familiarity of the vocabulary used in the system. Participants were recruited through Facebook ads and the university website during April and May 2015. Table 1 shows the distribution of users by gender and age group.

In this phase, the participants carried out the study on the Internet through their own computers without direct contact with the research team. Provided that we were at Stage 3 of development [34], we aimed to explore the symptom recording cognitive process rather than the tool usefulness as a whole. Therefore, we selected a reduced subset of six questions adapted mainly from TAM's ease of use set. In addition to the TAM–based questions, a question that aimed to detect problems in the communication of clinical terms was added. The questions are displayed in Figure 5. To answer the questions, the users selected a value in a continuous 10-point Likert scale inspired by the procedure of Tedesco and Tullis [51] (with 0=Totally Disagree to 10 = Totally Agree).

The application asked the participants to record a set of symptoms among wheezing, shortness of breath, fever, weight loss, chest pain, headache, cough, and feeling generally unwell. They were instructed to go over the website workflow until the system informed them that their symptoms had been recorded. Once the users had recorded their set of symptoms, the evaluation questionnaire was displayed (i.e., the questionnaire was completed once during each session). The users could choose to record their real symptoms (if they were ill or had recently

been ill) or record a set of symptoms of their choice according to their previous experiences while being ill.

Table 1. Gender and age groups of participants.

Age group	Female	Male	Total
19-29	2	1	3
30-49	24	3	27
50-64	14	6	20
65+	2	2	4

Figure 5. Evaluation of the TAM-based questionnaire.

Originally, we aimed for total number of samples of 100. However, after cleaning and removing duplicates, a total of 53 subjects had completed the symptom recording process and submitted the usability evaluation questionnaire. Duplicates were detected based on IP addresses. All symptoms were checked to ensure that they had been covered by reviewing the data recorded in each section. The users´ responses are provided as additional material.

In reviewing gathered data, we saw that it was formed by variables of different natures; that is, qualitative vs. quantitative. Table 2 contains the independent variables considered in the study.

Four of them relate to demographic data (gender, age, chronic diseases, and ill), whereas the other nine relate to the symptoms that could be recorded.

Table 2.Independent variables.

Variable	Type	Possible values
Ill	Qualitative	1/0
Gender	Qualitative	0 = male, 1 = female
Age	Quantitative ordinal (age ranges)	18-29 years ->1 30-49 years ->2 50-64 years->3 65+ years -> 4
Chronic disease	Qualitative	1/0 (presence or not of chronic diseases)
One additional variable per symptom (wheezing, cough, productive cough, shortness of breath, headache, chest pain, fever, weight loss, generally unwell)	Qualitative	1/0 (depending weather the user recorded that symptom or not)

The seven questions of the questionnaire shown in Figure 3 led to seven quantitative variables (Qi) representing the answer to each question and ranging from 1 (totally disagree) to 10 (totally agree). The 53 subjects, who provided seven answers each, led to a total of 371 answers for all questions. Among them, four missing values were present in the questionnaire responses dataset. We considered that dropping all the information from those subjects (six answers remaining from each) would lead to more information loss than imputating them. Therefore, we imputated the four missing values as the average of all the values provided for that question. Data from one subject was excluded for being considered as an outlier.

The questionnaire data were analyzed to identify factors influencing the results by types of symptoms registered, previous disease (diabetes, COPD, asthma, cardiovascular, or other), age range, and gender. To unveil the usability issues of the system, the data registered by the users (the independent variables corresponding to symptoms, demographics, etc.) needed to be related to the answers that they provided to the questionnaire (the dependent variables that identify the responses to the questionnaire). In this way, it was possible to determine how each independent variable influenced the questionnaire responses (positively or negatively).

TAM questions represent a way of measuring a variable that cannot be directly observed: the acceptance of the technology by the users. This involves a problem of multicollinearity among all the dependent variables since, in essence, they are measuring the same thing; presenting a challenge in dealing with the high dimensionality present (14 independent variables and 7 dependent variables). To deal with that situation, we proceeded to determine which independent variables influenced the users´ technology acceptance in two steps:

a) First, we applied PCA to reduce the dimensionality of the response variable (i.e., 7 Qi variables) to two uncorrelated PCs. As explained in the Results section, the first PC was associated with the variables derived from TAM (TAM_PC); the second component was associated with the familiarity of the vocabulary (VOC_PC).

b) Second, the scores derived from the PCA were used to estimate two regression models with the objective of quantifying the effects of the independent variables (X_i) over the mean values of TAM_PC and VOC_PC.

The statistical software packages used for the Phase II analysis were Stata 14 and R.

3.3.3. Phase II: Problem analysis

As a result of the problem detection phase we determined several areas that needed further investigation to analyze why they generated a negative or positive contribution to the PCs. Therefore, the think-aloud procedure was executed to provide insights into the cognitive process of users when they register their symptoms. Phase II uses the outcome of Phase I to constrain the areas of the GUI that must be tested to diagnose the causes of their significant contribution, thus minimizing the number of users needed for the think-aloud procedure. The execution of the think-aloud procedure relied on a set of vignettes that were designed from clinical resources and medical literature used to train clinicians [6]. Additionally, the vignettes were validated by a GP (JCA). The set of vignettes contained general symptoms of respiratory diseases, focusing on those symptoms that had been detected to have a significant contribution to TAM_PC or VOC_PC.

The think-aloud procedure was performed with 15 individuals between April and July 2016. The users were recruited via mailing lists and advertisements on the university website. This sample was independent from Phase I´s sample. Participants were native Norwegian speakers, were attached to the Norwegian healthcare system, and did not have an educational or professional background related to healthcare; most had a high educational profile, used a

computer on a daily basis, and did not show signs of cognitive impairment. No formal questionnaire was used to detect cognitive impairment; rather, we used the training stage during the think-aloud procedure for that. Of the 15 participants, 2 belonged to the [18-29] age group; 7 belonged to [30-49]; 5 belonged to [50-64]; and 1 belonged to the [65+] age group. Regarding their educational profiles, 3 had completed secondary education; 1 had a bachelor´s degree; 3 had master´s degrees and 3 had PhDs. Regarding gender, 5 were male and 10 were female. After the test, the participants were awarded with a lottery ticket. The data privacy delegate of the University Hospital of North Norway approved the study. The think-aloud procedure started by introducing the participants to the system´s objective; second, they continued training on an external website until they performed the think-aloud procedure properly; third, the session with Erdusyk was performed; and finally, a retrospective interview was conducted to analyze the user´s problems noted by the two interviewers during the procedure. The sessions were videotaped and the screen was recorded with ActivePresenter®. The data was transcribed verbatim and analyzed qualitatively by two independent reviewers (LMR, EB) in NVivo 11 following the Framework method [52,53]. The average weighted interrater agreement calculated using Cohen's Kappa was 0.82, almost perfect agreement [54]. The steps followed during the qualitative analysis were:

1) Familiarization: The two reviewers went through the interview materials independently, reading the notes taken during the interview, listening to the recordings and/or watching the videotaped sessions. The familiarization was performed freely and each reviewer wrote his or her own impressions separately and chose which material to review (audio, video and/or interview notes) without a defined guide. The verbatim transcripts were not used in the familiarization stage.

2) Open inductive coding of interviews: The two reviewers went through five interviews independently, coding them for HCI barriers and problems using NVivo 11. No predefined code list was used. The only agreement made before coding was that the reviewers would only code problems caused by the system and not problems caused by the user´s lack of attention to the assigned vignette. This provided the initial code sets used in the following stage to develop the framework index.

3) Development and application of the analytical framework: The stages of development and application of the analytical framework overlapped in the execution of the Framework method. The framework index was developed by iterating over the codes and notes taken in the coding stage until the reviewers agreed on a set of common codes and inductively identified hierarchical categories of the usability problems. The reviewers parallelized the coding and the

transcript tasks as much as possible by coding transcripts as they were provided by the external transcription service. For every three transcripts analyzed, the reviewers met and crosschecked their results. When both reviewers agreed that an update of the framework index was necessary, they updated the categories and codes. When a modification of the index was performed, all the coded interviews needed to be updated. The most common sources of disagreement between the reviewers included differences in how specific the code to identify an issue should be, differences in the interpretation of the codes in the framework index, and determining when it was necessary to add a new code. Disagreements were discussed until a consensus on how to proceed was reached by the two reviewers. Iterative modifications of the framework index resulted in the task continuing until the very end of the qualitative analysis. The use of the qualitative data analysis software was crucial for keeping track of the changes performed and re-coding when the framework index was updated.

4) Charting data into the framework matrix: Once all the transcripts were coded with the final version of the index, the framework matrix (containing users as rows and codes of the index as columns) was generated using the qualitative data analysis software. The framework matrix contained the issues that each user faced in verbatim text, retaining references to the original transcripts.

5) Interpretation of data: The results contained in the framework matrix were analyzed by the two reviewers to summarize the different issues and classify them as shown in Table 6, Table 7, and Table 8 **Error! Reference source not found.**in APPENDIX I. This summary was used to interpret and abstract the results further with partial support from other members of the research team, going back to the original text when necessary, and to write the conclusions reported in the Results section of the paper.

4. Results

4.1. Phase I: Problem detection

4.1.1. Dimensionality reduction of response variables with PCA

PCA was performed in order to reduce the seven response variables (Qi) to two PCs.

After scaling the Qi values, PCA was performed, which generated 7 PCs as a result. At this point, the minimum set of PCs that better represented the total variance needed to be selected. The scree plot of the proportion of variance explained by each PC is shown in Figure 6. The "elbow" is found between the second and third component.

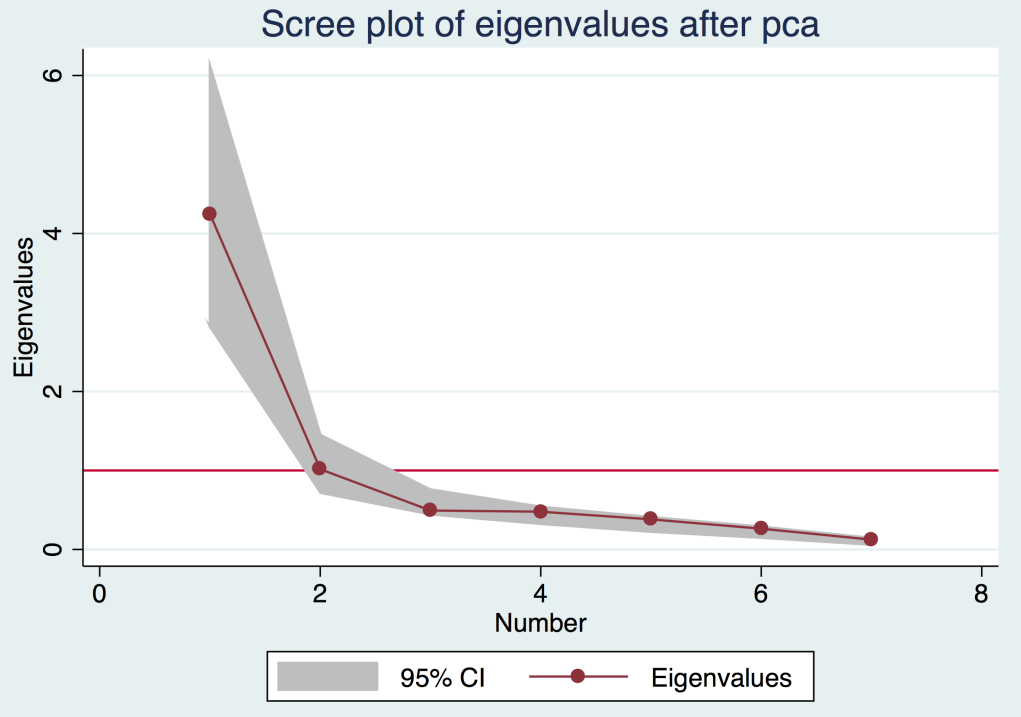

Figure 6. Scree plot of the variances represented by each PC.

The first component clearly represents a big fraction of the variance, whereas the second component lies on the borderline. Both were retained as together they explain nearly 75% of variance and, as discussed in the next section, it makes sense to retain PC2 as well according to the data domain. This way we have retained the two components that have an eigenvalue higher than one.

The two PCs (PC1 and PC2) selected are the dimensions that best represent the variation of the response data when the results of the set of Qis (the answers to the questions) are projected onto them. Based on this, a biplot can be built to observe how the subjects and their Qi values lay on this new two-dimensional space.

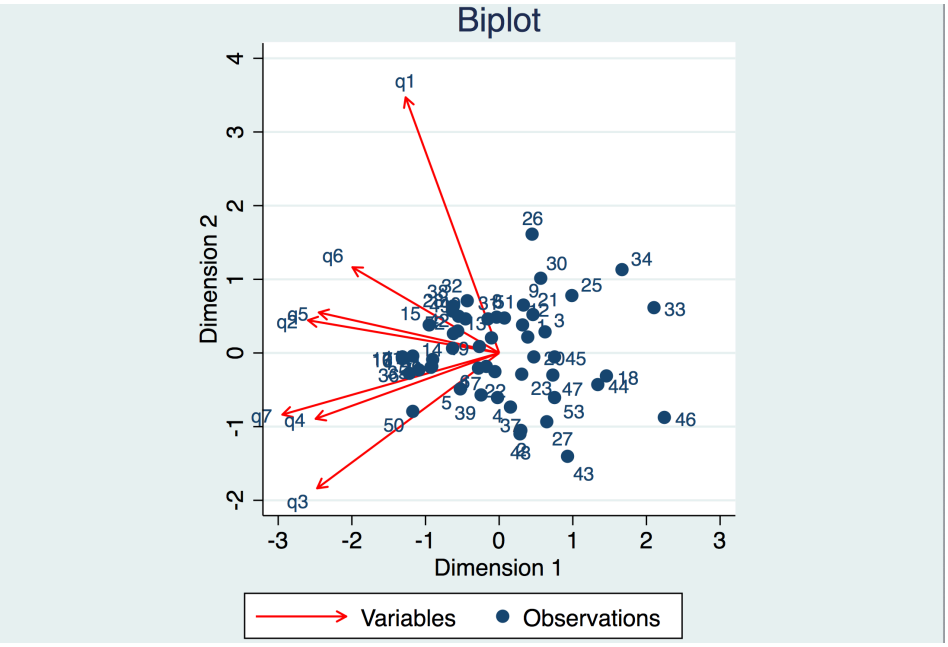

Figure 7. Biplot of Q_i variables projected onto the selected PCs.

The biplot in Figure 7 displays the subjects´ responses projected on the selected PCs and the directions of variation of each response variable. The numbers are the subjects´ IDs and the red vectors are the gradients that show the direction where the value of the Qi response variable grows faster. The similarity in the direction of the response variable vectors provides an idea of the correlation among variables. As depicted in the biplot, variable Q1 is less correlated to the other variables that have strong positive correlations. In looking at the domain of the study, it is observed that Q1 corresponds to the additional question introduced in the usability questionnaire related to the understanding of the vocabulary; whereas the other variables correspond to the questions adapted directly form TAM. In terms of the correlation of the TAM questions, it is possible to see how the TAM questions effectively allow indirect observation of the underlying technology acceptance factor.

When the directions of the response variables were checked, it was clear that PC1 seems to provide a general measure of the variation of all the TAM responses (Q2 to Q7) in one single dimension, while PC2 seems to summarize the variation of the responses to Q1. To confirm this, the correlation coefficients (r) of each PC with the Qi variables were checked. PC1 has a high correlation coefficient with Q2-Q7 (r between 0.73 for Q6, lowest; and 0.91 for Q7, highest). Q1 correlation is relatively low with PC1 (r = 0.45) and more correlated with PC2 (r = 0.81). For clarity, as previously done in the Methods section, PC1 and PC2 will be identified as TAM_PC and VOC_PC.

By keeping these two components, most of the effects of the information provided for symptoms and demographics over the questionnaire outcomes can be observed in two independent response variables: (a) TAM_PC, representing a summary of all TAM-related

questions (Q2-Q7); and (b) VOC_PC, representing the vocabulary question (Q1). At this point, we have the values (scores) that represent the projection of each subject on TAM_PC and VOC_PC. Since they are orthogonal, we can build two different regression models to study the effects of the input data on the values of TAM_PC and VOC_PC.

4.1.2. Analysis of the relationship between independent variables and TAM_PC/VOC_PC

In the previous section, the seven correlated response variables were reduced to the two independent PCs: TAM_PC and VOC_PC. In this section, the influence of the independent variables on the expectancy of the two PCs will be studied. To do so, it is possible to proceed by estimating two different regression models:

(a) One model to study the effect of the independent variables (Xi) on the expectancy of TAM_PC; and

(b) A second model to study the independent variables (Xi) on VOC_PC.

Stepwise regression was used to estimate both models, because it could deal with the high number of independent variables. The significance threshold for p-values was set to 0.05[1].

Study of the effect of Xi on the expectancy of TAM_PC

By applying stepwise regression, it is possible to estimate a model with the response variables explaining most of the total variance, as shown in Table 3.

Table 3. Regression model for TAM_PC response.

Variable	Coefficient	P-value	Confidence Interval		Model R^2
FEVER	1.946262	0.083	-0.265891	4.15841	0.2413 (model p-value=0.0034)
COUGH	1.161513	0.051	-0.004844	2.32787	
WHEEZING	-3.385491	0.015	-6.085208	-0.68577	
CONSTANT	-0.2892261	0.355	-0.911587	0.333134	

[1] This threshold was selected only because it is the default value in most studies. But in exploratory studies, especially in models involving psychological or sociological indirectly measured variables, relaxing it to 0.1 may be adequate. In fact, in our case, it would make all variables significant.

The coefficients of the model can be interpreted as follows:

$\beta_0 = -0.2892 \rightarrow constant\ of\ the\ model$
$\beta_1 = WHEEZING = -3.385$
$\rightarrow\ when\ WHEEZING\ is\ reported\ by\ the\ user, there\ is\ an\ average\ decrement\ \ in\ the\ TAM\_PC\ response\ of\ 3.385\ units$
$\beta_2 = COUGH = 1.16$
$\rightarrow when\ COUGH\ is\ reported\ by\ the\ user, there\ is\ an\ average\ increment\ \ in\ the\ TAM\_PC\ response\ of\ 1.16\ units$
$\beta_3 = FEVER = 1.946$
$\rightarrow\ when\ FEVER\ is\ reported\ by\ the\ user, there\ is\ an\ average\ increment\ in\ the\ expectancy\ of\ TAM\_PC\ of\ 1.1946\ units$

Table 3 shows a significant p-value ($0.0034<<0.05$) for the model in explaining TAM using the independent variables considered. In terms of R^2, the model is able to explain around a 24.1% of the variance in the response. The coefficient of WHEEZING is clearly significant (p-value=0.015) and negative, indicating a tendency of the users reporting it toward evaluating TAM more negatively. The coefficients of COUGH and FEVER are almost significant (p-values=0.051 and 0.083, respectively) with both being positive. This seems to indicate a tendency of the users that reported those symptoms toward evaluating TAM more positively. Therefore, they were considered for further investigation with the think-aloud procedure to confirm or dismiss their influence.

Study of the effects on the expectancy of VOC_PC

To study the effects of the independent variables on VOC_PC, stepwise regression was used to estimate the model in Table 4.

Table 4. Regression model for VOC_PC response.

Variable	Coefficient	P-value	Confidence Interval		Model R^2
ILL_PERSON_DATA	-0.5713112	0.048	-1.138047	-0.00457	0.2171 (model p-value=0.007)
WHEEZING	-1.15775	0.088	-2.494873	0.1793729	
AGE	0.4004904	0.028	0.045788	0.7551927	
CONSTANT	-0.7845108	0.099	-1.720681	0.151659	

The coefficients of the model can be interpreted as follows:

$\beta_0 = -0.784 \rightarrow model\ constant$
$\beta_1 = -0.571$
$\rightarrow the\ fact\ of\ having\ an\ ill\ subject\ recording\ real\ data\ produces\ an\ average\ decrement\ of\ VOC\_PC\ of\ 0.571\ units$

$\beta_2 = 0.4$
$\rightarrow$ *each increment in the range of age produces an average increment of VOC_PC of* 0.4 *units*

$\beta_3 = -1.157$
$\rightarrow$ *the fact of reporting WHEEZING produces an average decrement of VOC_PC of* 1.157 *units*

Table 4 shows a significant p-value ($0.007 << 0.05$) for the model in explaining VOC_PC using the independent variables considered. In terms of R^2, the model is able to explain 21.7% of the variance in the response. Reporting wheezing contributed to a worse outcome in the evaluation of the understanding of the vocabulary (Q1). Also, the user being ill at the moment of using the application led to worse outcomes in the evaluation of the understanding of the vocabulary (Q1). The coefficients of ILL_PERSON and AGE were clearly significant (p-values 0.048 and 0.028, respectively). WHEEZING was almost significant (p-value = 0.088) and it was investigated further in Phase II to clarify its significance.

4.2. Phase II: Problem analysis

Taking the variables from Phase I that produced significant contributions to TAM_PC and VOC_PC expectancy into account, we designed a set of vignettes that contained realistic cases where those variables would be present. Then we assigned them to a set of users and performed the think-aloud procedure to identify the causes of those contributions. Figure 8 shows the analytical framework that resulted from iterating over the codes and notes taken in the coding stage, and later proceeding inductively to classify them as hierarchical categories. Once it was stable, the framework was used to code all the interview transcripts of the think-aloud procedure. The parentheses of each node contain two numbers: the first number corresponds to the number of different users that mentioned each code; the second number corresponds to the total number of times that the code was mentioned, irrespective of the user. Three main axes are present encompassing HCI observations, namely, design issues, interpretation issues, and general user opinions. Table 6, Table 7, and Table 8 in APPENDIX I present the subcategories of each of the axes and a summary of the problems related to them found with the think-aloud procedure.

Table 5 contains the variables detected in Phase I with significant contributions (see the Contribution to PC column) to TAM_PC or VOC_PC mapped to the causes found during Phase II. The Code column contains the code from the Framework index. The Reason column contains an explanation of the cause for the significant contribution.

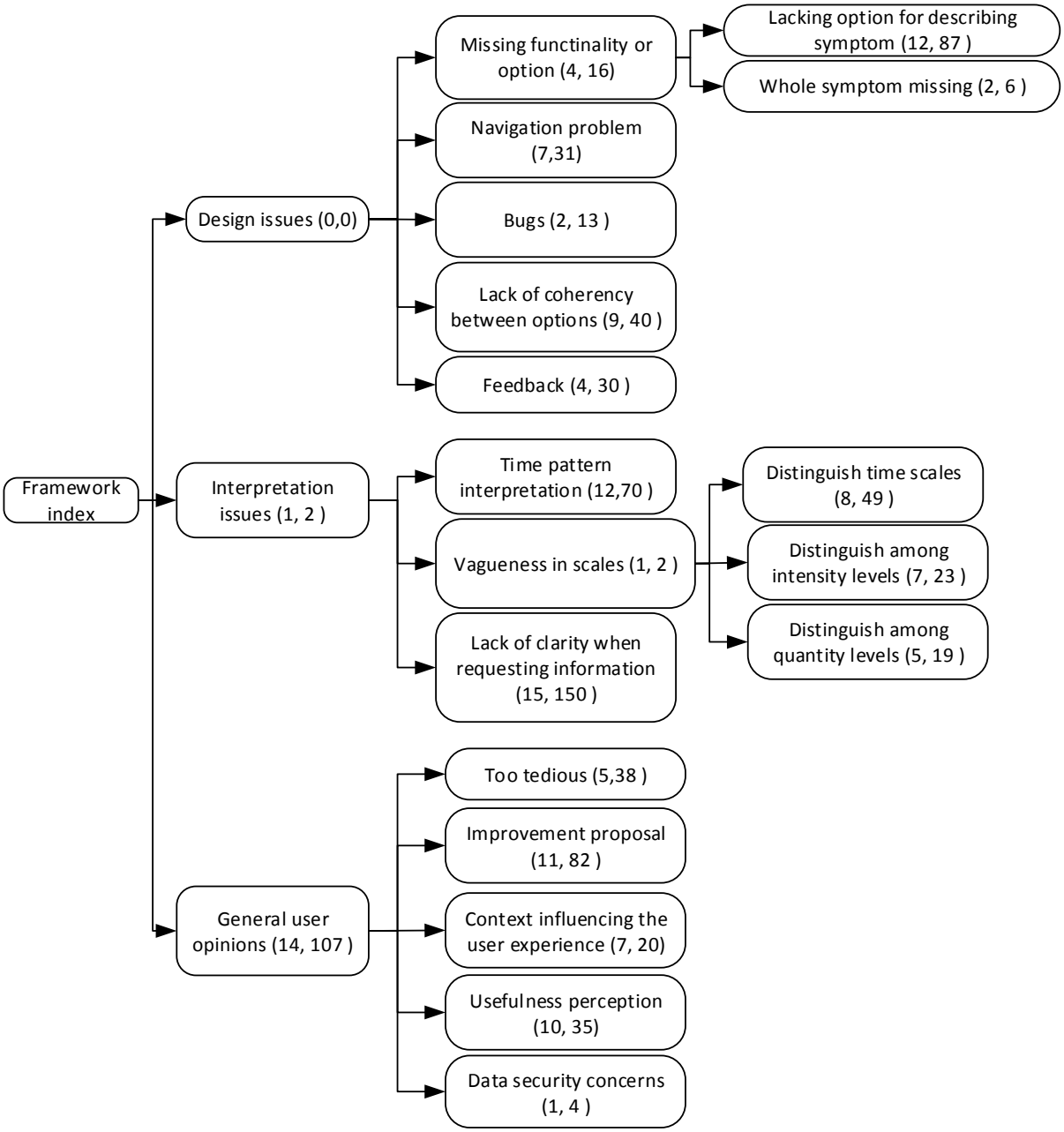

Figure 8. Framework index.

Table 5. Significant areas for technology acceptance (found in Phase I) mapped to their causes (found in Phase II).

Principal Component (PC)	Variable	Contribution to PC	Code	Reason
TAM_PC (summary of the responses to questions from TAM)	WHEEZING	NEGATIVE	*Time pattern interpretation *Too tedious	*Bad localization of the archetype for this symptom
	FEVER	POSITIVE	(Positive aspect; no code related)	*Good localization
	COUGH	POSITIVE	(Positive aspect; no code related)	*Good localization
VOC_PC (summary of the responses to the understanding of the vocabulary)	ILL	NEGATIVE	*Context influencing user experience *Too tedious	*Could not be determined. Users have diverse opinions on how being ill would influence the use of the system.
	WHEEZING	NEGATIVE	*Lack of clarity requesting information *Difference among intensity levels (trivial, mild, moderate) *Difference in time scale (suddenly, rapid, gradually) *Distinguish among intensity levels *Bugs	*Intensity levels and scales from SNOMED-CT cannot be properly interpreted without examples *Error in the headings of some sections
	SPUTUM (detected during the think-aloud procedure)	NEGATIVE	*Lack of clarity when requesting information *Differences among quantity levels *Missing functionality or option (cannot specify color properly) *Improvement proposal	*Term *sputum* not understood *Intermediate scales of colors cannot be specified *Examples are needed for specifying quantities, color, and so on.

	AGE	POSITIVE	*Context influencing user experience *Too tedious	*Young users had more attention to detail and they pointed out sources of confusion in the interface

4.2.1. Diagnosis of negative contributors to the PCs

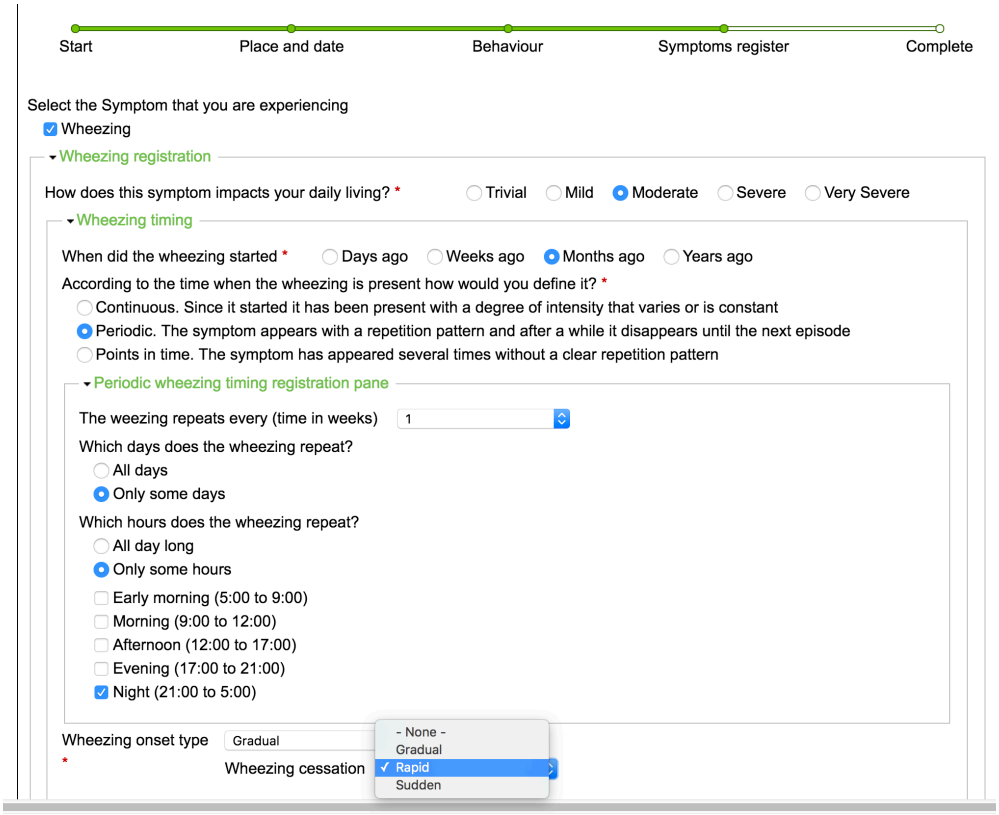

Figure 9. English version of the wheezing symptom screen.

As shown in Table 5, the think-aloud procedure revealed the causes for the variables contributing negatively to TAM_PC. For the variable representing WHEEZING (depicted in Figure 9), the interviews revealed that the negative contribution to VOC_PC was caused by bad

localization of the symptom archetype and an error in the names of the sections that created confusion when recording the symptom information. Regarding localization, the archetype for symptom is a generic maximum dataset for all symptoms and needs to be constrained to deal with only the entities that are relevant for each context (symptom) [55]. We had kept a minimum level of localization that led to attributes such as the character of onset/cessation (e.g., gradual, sudden, rapid etc.) to appear for wheezing when they were not relevant for that symptom. Besides, users complained that those scales were not natural to them. Again, this was a problem of bad localization from SNOMED-CT terms to express velocities in symptom onset/cessation and a lack of an appropriate explanation with examples as discussed in the following sections. Wheezing is perceived as a continuous symptom where the character of the onset/cessation is irrelevant. One user said:

"Some choices caused some trouble for me to understand: the distinction between 'rapid' and 'sudden',... it may be very hard to distinguish between those two parameters."

It was not possible to identify the causes for the negative contribution of the variable ILL to VOC_PC. While some users (e.g., user 7) considered that being ill would make them less tolerant of providing detailed information; others (e.g., users 13 and 11) considered that if they were ill, they would be willing to devote more time and effort to provide all the detailed information requested.

4.2.2. Diagnosis of positive contributors to the PCs

In addition to the negative contributions to TAM_PC and VOC_PC, the think–aloud procedure allowed the identification of the causes of positive contributors (see Table 5). The positive contribution of the variable AGE to VOC_PC indicated that the older the users were, the more positively they evaluated the question represented by VOC_PC. The interviews revealed that the cause for this positive contribution of AGE was that six out of the seven navigation problems detected had been pointed out by users younger than 50. Also, the code being too tedious was more frequent in the interviews of users younger than 50, since it was pointed out by three out of five. Additionally, the density of the code per interview was also higher in users younger than 50 (six out of eight). The cause for this difference in the problems detected depending on age was revealed during the think-aloud procedure. On the one hand, it was found that young users showed more attention to detail and devoted more time trying to understand the complex navigation across subsections; on the other hand, older users tended to navigate in a more superficial way through those parts that had complex navigation or had a high level of detail

(e.g., time pattern subsections). This caused young users to be more aware of the existing problems while the older users made assumptions about correct navigational behavior without analyzing the section in depth.

The positive contribution of the variable COUGH was explained during the interviews as a consequence of the good localization of the symptom archetype. In the case of the symptom cough, all attributes of the model were relevant to it. For example, the periodicity fit very well since many types of cough are present with some pattern (e.g., early morning cough). Additionally, its onset/cessation may differ depending on the condition causing it. Despite the fact that COUGH had a positive contribution, the interviews performed after the think-aloud procedure detected problems in a subsection of COUGH that was displayed when the productive cough option was marked. The subsection was intended to record the characteristics of sputum, but had not been detected in Phase I as a problem. Problems related to sputum were linked to the specification of its color and quantity. Users needed examples to quantify volume and more flexibility to decide about color. Additionally, the term used in the Norwegian language was considered too medical. One user said about sputum volume:

"…you are suppose to describe how much is moderate, how much is normal; I don't know, many people have different notions about quantity."

Regarding the positive contribution of FEVER, the interviews performed after the think-aloud procedure revealed that it was better localized than other symptoms, including characteristics like the body location of the measurement. This made the symptom features easier to interpret for users. Despite this better localization, users identified the cessation and onset character as irrelevant attributes for fever.

4.2.3. Additional issues unveiled by the think-aloud procedure

In addition to allowing us to understand why some sections and symptoms had negative or positive contributions to TAM_PC and VOC_PC, the think-aloud procedure helped us to detect and understand many other usability issues. Below, are some other usability issues that the think-aloud procedure helped to identify.

Navigation axe:

Several issues related to problems with the navigation were detected. The most relevant was a bug in the system that deleted the information that had already been completed and drove the

user back to the start screen when a specific combination of options was pressed. Second, users pointed to the need for providing better feedback and guidance across all sections so they knew which section they were completing at all times and were aware when they had finished one section and began another one. For example, one user mentioned:

"Yeah you need a little guidance, I think...well you will find out as you are doing it, but its so easy to lose those (the symptoms) on the top especially if you start with cough."

In addition, the amount of detail and number of sections made the users lose their sense of where they were at each point. The users appreciated the navigation bar, but also commented on the need for additional feedback informing them about how much information they needed to provide to finish each section and to better differentiate each of the sections they were going through.

Users like reassurance that they have finished completing a section with an explicit indicator. In addition, for inner subsections, they proposed using different headers and text sizes to better identify the nesting structure of the subsections. The need for better guidance was also identified with problems about understanding when a feature refers to each symptom episode or to the whole history of the symptom. More guidance was also needed when requesting complex information, such as time patterns. This was linked to the positive contribution of AGE since older adults may need better guidance to record details appropriately. Users pointed out that it would be appropriate to start filling out information according to the symptom that is most concerning and continuing in a decreasing order of importance.

Finally, although mandatory fields were indicated by a red star, some users wanted to be informed about this more explicitly to avoid having to go back to search for them when the error was displayed.

<u>Lacking options or functionalities:</u>

Regarding missing options and functionalities, the think-aloud procedure made us reconsider adding the section for reporting the precipitating factor of a symptom, which had not been implemented. Users considered this to be paramount since it would allow them to link causes or factors that worsen or improve a symptom. The users complained that they could not express such factors.

When asked about missing signs linked to respiratory symptoms, the users mentioned that they would like to mention joint pain that is often present in flu episodes.

Regarding the complexity of the system, the users encouraged us to reduce the level of detail when possible since they needed more flexibility in providing some details that were difficult to

remember. For example, some users pointed out that having so many subsections for recording each symptom would make them feel anxious if they felt ill. User 7 said:

"Too many choices, too many questions, I would say. I think in a realistic situation I would be a little impatient with all these issues and options, so I had enough… perhaps I might only choose something to make it go faster."

The problem of sections that were too detailed was also noticed in the sections about behavioral information, such as tobacco consumption, where users preferred to be able to include cigars and casual smoking behaviors in general rather than using accurate pack-year measures as in the clinical domain.

Users pointed out that when leaving a section unanswered they thought they were doing something wrong. An example of this is the section for recording chronic diseases, where no field must be completed if no disease is present. The users preferred to have a default option to explicitly specify that the condition was not present rather than leaving the section unanswered. For example:

"Ok. There should be an alternative that you don´t have. No chronic diseases."

In addition, users pointed out that, in some situations, the user may be providing information on behalf of another person. Therefore, this should be considered as an option, and that accessibility should be taken into account.

<u>General user opinions:</u>

Users also provided valuable feedback about general topics during the think-aloud procedure. The most important general issue detected was the need for examples to differentiate among the levels of the scales. For example, users recommended that the scales taken from SNOMED-CT terms to specify the volume of sputum or intensity levels should be explained with examples. With regards to volume, the users wanted examples that specified quantity (e.g., "half a teaspoon").

With regards to onset/cessation, the users recommended that types should be described, and for intensity levels, which were selected from the SNOMED-CT sub-concepts of symptom severity (i.e., trivial, mild, moderate, etc.), they should be illustrated by examples of impact on daily living (e.g., "you are not able to go to work"). About intensities, one user said:

"Moderate… does that mean I can't go to work? Or does it mean I feel bad at work? Or does it mean that I stay in bed all day? … It is hard to know seriously and if you are in the system and wondering whether you should go to the hospital or not or whether you should go to the doctor or not…"

Another issue identified was the need to request information more clearly in sections by stating the sentence as a question rather than as a section title. For example "Do you have any chronic diseases?" was preferred to "Chronic diseases" as a section title. Also, some words were identified as too medical and unnecessary; users pointed out that they did not need the medical term as long as they had a good description that allowed them to understand the information requested. An example is headache types (e.g., cluster, tension, or migraine), where a good description was enough for users to understand and appropriately communicate the headache type without the need to know the medical term. Some users forgot some symptoms; one user proposed avoiding this by starting with the symptom he was the most worried about and continue recording symptoms in order of their importance. Finally, only one user mentioned that he or she would like to be told explicitly that the data would be appropriately handled.

4.2.4. Perceptions of users about the usefulness of the symptom checker

When asked generally about the system, the users perceived consumer CDSSs for symptom checking as a positive initiative (see **Error! Reference source not found.**). All users acknowledged the usefulness of symptom checkers to avoid unnecessary visits to the GP and to be more informed about health issues. Two users also mentioned the specific case of parents that need to access reliable health information to decide about whether their children need to visit a doctor or not. Several users mentioned the problem of checking raw information on the Internet since this often creates anxiety and unnecessary concern leading to stress and unnecessary GP visits [56]. For them, initiatives involving symptom checkers that drive the user to reliable health information may have a positive impact on their quality of life by making better use of health resources and avoiding the effort that making an appointment and getting to a GP office involves. For example:

"I think that is a very good idea. I think most people if they know that it exists they would use it, because they will trust that more than a random search on the Internet, because most people know that a lot of information on the Internet is a bit scary."

Finally, no differences were detected according to educational level or gender.

5. Discussion

5.1. Multivariate statistics in usability testing

As in psychology studies, usability data interpretation involves the study of latent variables (e.g., usefulness, ease of use, satisfaction, etc.) that are observed indirectly. This indirect measure is usually carried out with methods that contain many correlated items (e.g., questionnaires or interview sections). Moreover, this response variable depends on many dimensions determined by the characteristics of users (age, profession, computer literacy, etc.) and systems (design choices, layout, color coding, etc.). This makes usability studies multivariate in nature. Multivariate statistics represent a way for usability practitioners to observe how many variables involved in a study interact at a time. Some usability studies have employed the power of multivariate statistics to perform comparisons of different products or settings. Davis and Jiang made use of another multivariate statistical tool (MANOVA) to study the significance of differences among three diabetes websites whose usability was measured as correlated variables [40]. Similarly, Smith et al. used MANOVA to study the effect of an intervention with a website for helping patients that had suffered from stroke [57]. These studies used multivariate statistics to check for significant differences among groups. When it comes to PCA, Sauro and Kindlund proposed the use of PCA with the objective of unifying several usability metrics into a single PC [58]. However, an even more appealing use of these techniques is as a part of the usability methodology itself, thus providing insights into the true structure of the variance in the study data. In specific, dimensionality reduction techniques such as PCA or factor analysis help to observe the latent structure of data, separating it from the random variance introduced by the indirect observation. This was seen in the usability questionnaire of Phase I, where all TAM-related questions could be reduced to only two PCs that could be regressed to detect variables with a significant influence on technology acceptance. Understanding this variation allows for investigating all the relationships among the variables in a quantitative manner, thus maximizing the knowledge extracted from the study dataset.

A very related multivariate technique that could have been used as an alternative to PCA is factor analysis. However, factor analysis assumes prior knowledge of the latent variable [59]. PCA was preferred over factor analysis since we had adapted TAM questions and we had added a question that might not be aligned with the underlying model. Therefore, we preferred to use PCA, which reduces dimensionality by focusing on explaining the maximum variance possible without making assumptions about the underlying model.

Multivariate statistics allow for quantitatively analyzing the studies that support the conclusions of usability studies. Therefore, if they are appropriately combined with current usability methods, multivariate statistics can help to make optimal use of resources by concentrating testing efforts in those areas where there is statistical evidence of usability problems. Nevertheless, it is important to note that although multivariate statistics are a

powerful data visualization and analysis tool, they are not a "one size fits all" approach to analyzing data from usability studies. These kinds of statistics are complex and require trained professionals to interpret their results. Furthermore, they need a large amount of data to provide useful results. Therefore, they are appropriate for testing complex scenarios with heterogeneous users and complex GUIs such as those of symptom checkers, but they may not provide a significant benefit in environments such as CDSSs for professionals where users are experts in the business process and the size of the GUI is moderate.

5.2.Findings about users

In general terms, our method detected that the usefulness perception of Erdusyk was high with all users, acknowledging that it is a useful technology that can help to reduce unnecessary visits to the GP, avoid anxiety when searching the Internet with search engines such as Google, and facilitate communication with the GP by making the patient reflect on symptom details.

In most sections, users understood the information requested by the system correctly. Nevertheless, some important barriers were detected. The main barriers were issues related to the interpretation of symptoms´ time patterns (e.g., recurrence). Most users failed to record them properly and many were confused due to the level of detail required. Another common barrier found by most users was uncertainty about deciding what level to select in scales (e.g., intensity levels such as trivial, mild, moderate, etc.). This was caused by a lack of examples to be used as references to select one level or another. Examples are needed to allow users to understand complex concepts. However, other studies have pointed out that providing examples that are too broad about a symptom may influence the user leading to biased information [20]. Therefore, examples need to be concrete and linked to particular sections. The users also found that the GUI presented some unnecessary medical terms that could be explained with definitions rather than using the medical name.

Navigation was another area where improvement is needed. Users appreciated the ability to do all symptom recordings in one screen, but missed more guided navigation across symptoms and their sections. They pointed to the need for explicitly marking the navigation over symptom sections at all times to avoid confusion about what section belongs to each symptom. Another finding related to navigation was that the users preferred to start with the symptom they were most concerned about and continue recording symptoms in a decreasing order of importance. In relation to this, other studies have documented the preference to record one symptom at a

time [20]. Although navigation can be improved, previous experiences show that navigation problems may be minimized but not always be fully eliminated [43].

Regarding the way information was requested, the users preferred that information be requested as short questions rather than section headings. They also pointed out that although the amount of information requested was acceptable, they would have been much more comfortable if the level of detail in some sections was reduced.

In terms of problems related to the adoption of clinical models, the think-aloud procedure revealed that some symptoms (e.g. wheezing) needed a complete redesign to localize the archetype properly. Users pointed to the possibility of avoiding some sections and also pointed for the need of including other sections or symptoms. Another aspect of relevance was related to the participant's characteristics; the second regressed model showed that older users evaluated the interface more positively. The think-aloud procedure revealed that the cause for this was the attention to detail that younger users demonstrated, unveiling more problems in understanding how information should be recorded in some complex navigation areas, whereas older users tended to navigate in a more superficial way, making presumptions about the application behavior or directly ignoring sections that involved complex terminology or navigation.

Our method failed to diagnose the cause for the negative contribution of the variable representing ill users. The users provided contradictory reasons to explain it during the think-aloud phase. We believe that the use of vignettes could not simulate the real setting in this case and studies with real patients may be needed to unveil how an illness affects the user. As in Luger et al. [20], the participants pointed out that the vignettes were an added complication when recording symptoms since the users did not really have the symptoms. Related to this issue was the symptom wheezing. The study identified that some onset/cessation characteristics should not be linked to the symptom wheezing. In addition, the users pointed out to problems understanding its repetition pattern. In some diseases, wheezing has a periodic presentation; however, the users did not identify it in that way. Wheezing is usually associated with chronic diseases such as asthma or COPD and real users are needed to further understand how its time pattern can be communicated.

These findings have allowed for the determination of which areas may lead to a misuse of the system, which could result in inaccurate recommendations to users. For example, appropriately recording time patterns, intensity scales, and onset/cessation characters is needed in order to differentiate cough episodes that have viral infections as a cause (where self-care is appropriate) from possible underlying asthma that should be further investigated by a doctor

[60]. At the moment, most symptom checkers only measure the main features of symptoms, but the knowledge derived from the think-aloud procedure about which elements of archetypes can be appropriately communicated may be used to include more detailed information, improving triage algorithms for future symptom checkers. Additionally, the effectiveness of HCI interaction in symptom checkers is highly dependent on health literacy. Therefore, the development of systems like Erdusyk must be coordinated with interventions to improve citizens´ health literacy, which is relatively low right now [23]. Nevertheless, through our experience, we learned that users are aware of the misuse of health services and they are willing to be educated to make better use of them.

5.3. Comparison with other studies

Several studies have presented approaches to test CDSS usability [37–40,43]. Lately, some studies have proposed combinations of different techniques to increase the problem detection rate and the robustness of the testing process [37,40,43,61]. Although the combination of techniques for CDSS usability testing offers very robust usability evaluation frameworks, there are some issues that may limit their adequacy to evaluate symptom checkers. The think-aloud procedure has been applied in CDSSs testing with end-users, but in environments where users were experts in the workflow of the business process that the system implemented. In the case of symptom checkers, evaluating the system with end-users and the think-aloud procedure will always be adequate for detecting communication barriers and avoiding negative outcomes. However, symptom checkers provide services to users with different health and computer literacy levels [23]. As a consequence of that heterogeneity, a large sample size is necessary to test symptom checkers, thus boosting the cost of executing the think-aloud procedure. Our method attempts to deal with CDSS end-user testing challenges with a detection phase using a large sample and a diagnosis stage that restricts the use of the think-aloud procedure to areas where problems have been detected. Nevertheless, it does not intend to be an alternative to other methods, but rather a cost-effective user-based method to complement them. As a matter of fact, TAM is one method that focuses on perceived usability rather than actual usability [62]. This implies that users may perceive a component of the interface as easy to use, while it has a usability problem (i.e., false negative). In the case of Erdusyk, Phase I performed well in detecting the issues related to the axes *Interpretation issues* and *General user opinion* as shown in Table 5. However, false negatives from Phase I were detected in aspects related to the *Design issues* axe. Again, inspection-based methods are appropriate to detect this type of issue. Therefore, our method can be applied to gain knowledge about the user's cognitive process and

it can be complemented with expert-based methods to avoid false negatives related to design issues. This is aligned with standards such as ISO 9241 and ISO 16982 that recommend the combination of both types of methods when the system must guarantee high quality. For example, Boland et al.´s [37] and Lai et al.´s [40] methods may be combined with our method, using their first stage with a cognitive walkthrough and/or heuristics to ensure that the GUI is acceptable from the usability expert´s point of view, and later applying our method to detect whether significant HCI barriers are present for end-users. Similarly, the application of the second stage of Van Engen-Verheul [43] could be useful after applying our method to detect which issues caused higher deviations from the optimal execution path. As explained later, all these methodologies can be used effectively in the evaluation stage of user-centered design (UCD) developments working iteratively toward the final product [63].

Another type of related research includes those that have studied the process of self-diagnosis with online resources outside the usability arena. Understanding such a process is paramount to guarantee the positive impact of symptom checkers. Luger et al. investigated the cognitive process of online self–diagnosis in older-adults using the think-aloud procedure [20]. Several of the findings presented in our study are consistent with the results of Luger et al. First, they detected that younger users seemed to be more accurate in using online tools for self-diagnosis; this may be a consequence of the more attention to detail and thorough navigation of younger users detected in our study. Second, our findings regarding the difficulties of navigation and feedback are consistent with Luger's findings using WebMD and Google for self-diagnosis, where the users found problems with navigation, layout, and error feedback. Although most of our results are consistent with Luger et al.´s, an important difference was found. In our case, most of the participants were aware of the risks of finding low quality information and appreciated having a symptom checker that could filter information well to avoid the "cyberchondria" [64] often derived from free searches in Google.

5.4. Strengths and Limitations

The method presented was used during Stage 3 of the development cycle, where system-user-task evaluation is performed [34]. But we believe that it is robust enough to be adapted to later stages of development where the system is evaluated in a real setting [34]. In such a case, the first phase should be performed using the complete TAM questionnaire and real patients should

be included in the second stage. Testing with real settings may, for example, explain why ill patients evaluate the system more negatively.

An important choice in every usability study is the sample size. We considered our sample size to be insufficient for Phase I and sufficient for Phase II. In Phase I, we aimed for a sample size of $n = 100$, but after the five campaigns of Facebook Ads and posts on the university website, we had gathered $n = 53$. This led to weak models in Phase I. Although psychology models usually have low but significant R^2, as in our case, some variables in both models were not significant (at a 95% significance level). The main problem was related to the VOC_PC response model, which was demonstrated as brittle. When the subjects containing imputated values were dropped and the statistical analysis was repeated, the TAM_PC response model did not vary; however, the significance of the variable ILL_PERSON_DATA in VOC_PC model became not significant. This is a consequence of a small sample size in Phase I. The small sample effect is also seen in the significance of WHEEZING in the model to quantify the effect on VOC_PC. Its coefficient in absolute value (1.157) indicated that it had a large effect on the mean of VOC_PC; however, it was not significant as a consequence of the small number of subjects reporting it. To avoid these problems, the methodology would be improved if instead of setting a pre-defined number of ad campaigns and closing the recruitment after executing them, Phase I was performed iteratively until no improvement in the models appears. That iterative procedure can be applied directly in UCD development environments, as depicted in Figure 10. The figure shows how, after the product is redesigned based on the feedback gathered in the previous iteration, the usability method is executed leading to more robust statistical models (clearer significances) with fewer significant areas. The iteration cycle stops when the method converges into a product that meets all user requirements; that is, the method does not detect any significant HCI barrier. This strategy may enhance the quality of user-centered design practice by introducing an effective feedback loop involving user participation in the overall development process while it reduces the cost of the extensive usability test and the number of iterations to achieve a higher quality of usability. As depicted in the figure, the methodology presented here is complementary to expert-based evaluation methods.

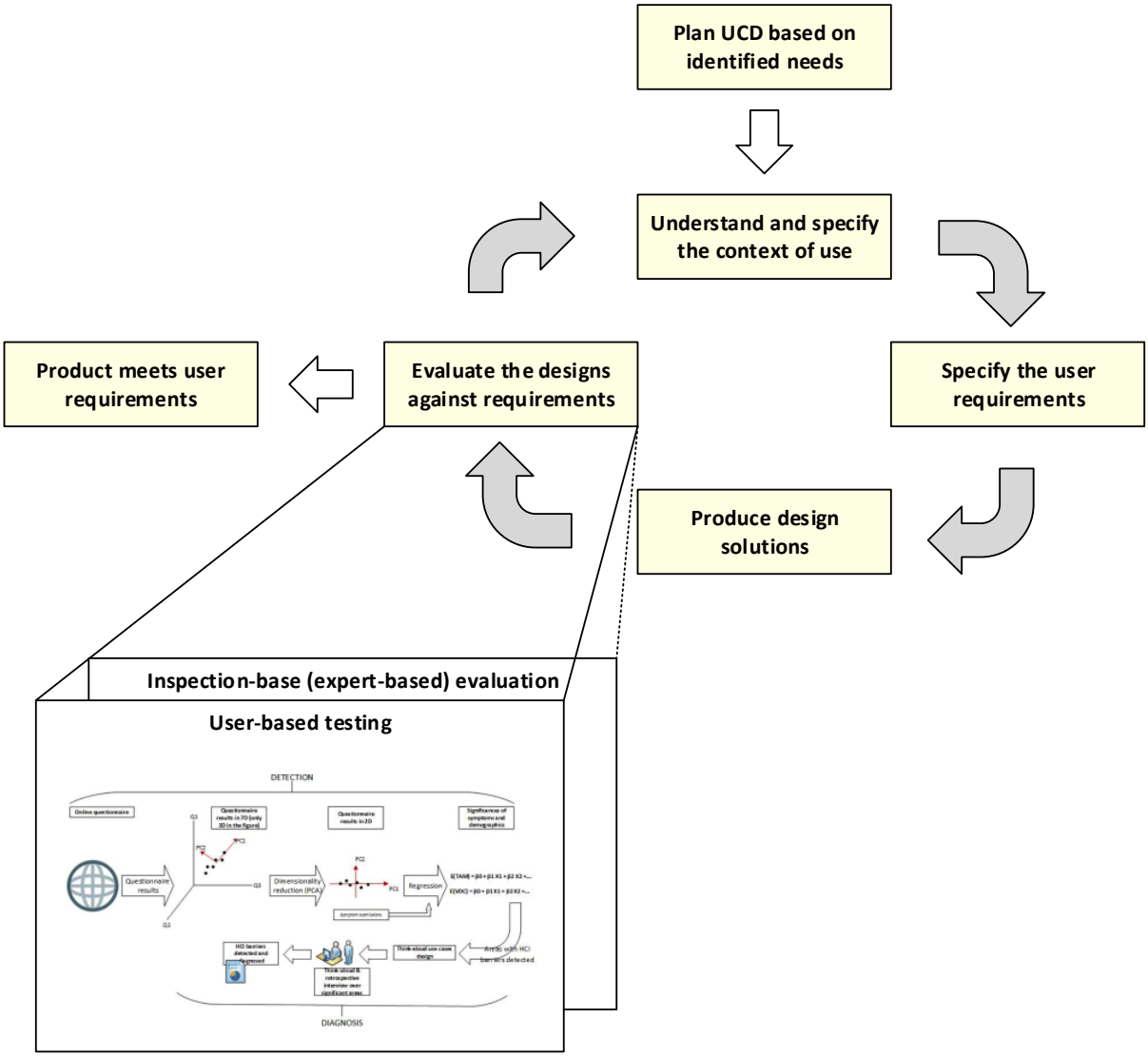

Figure 10. Inclusion of the methodology as a part of user-centered design developments.

In Phase I, we checked the logs and verified that all symptoms had been covered. As a consequence, we are confident that the areas explored were the ones with the highest concentration of HCI barriers. In some sections, such as that of the wheezing symptom, despite resulting in significance in the analysis, only two users had entered all the requested information. Nevertheless, the application of the think-aloud procedure as a gold standard served to verify that there were HCI barriers associated with that section. Regarding Phase II, our experience in performing the think-aloud procedure was that from user 6 onwards, the findings were redundant for most participants. Therefore, once the areas with the highest concentrations of problems have been detected, eight users may suffice to determine the causes using the think-aloud procedure, as recommended by the literature [36]. In Phase II, increasing the sample size would not have provided significant insights into variables that were not

explained, since real patients were needed for that. Real patients would be needed to understand why ill users evaluated the system more negatively, since the use of vignettes could not.

One limitation of this study is that most of the participants in the think-aloud procedure were recruited through the university website and most of them had higher education. Although this did not influence the methodology developed in this research, it may jeopardize the generalization of results for the evaluation of Erdusyk because most health consumers may have very different understandings of the terms and their meanings. Nevertheless, we did not find differences between users with different educational levels. Additionally, other studies suggest that the main users of online health information are in fact highly educated adults [65], which may indicate that we have covered a good proportion of Erdusyk target users.

6. Conclusion

The positive outcomes of symptom checkers depend on the seamless communication of medical concepts between the users and the system. The detection of HCI barriers in the user interfaces of symptom checkers is paramount to avoid providing misleading advice that may result in negative consequences for users. Testing with end-users is needed to assess how good a symptom checker is in communicating with users. However, the potential users of symptom checkers are very diverse and their interfaces typically contain a large number of different symptoms and possible execution paths. This may result in high cost when involving end-users to test the system. We have presented a method that aims to deal with those challenges in a cost-effective manner. The method allows, first, for detecting areas of the user interface that make significant contributions to technology acceptance; and, second, to analyze the causes of such contributions, limiting the use of the think-aloud procedure to significant sections of the user interface. Multivariate statistics allow for analyzing the results of remote tests performed by large samples of users, thus maximizing the coverage of the GUI. The results of the statistical analysis can be used during the second stage to determine which areas of the GUI the think-aloud procedure should be focused on to diagnose the causes of the problems found.

Acknowledgements

This work was supported by Helse Nord [grant HST1121-13], the Faculty of Health Sciences from UIT The Arctic University of Norway [researcher code 1108], and the Norwegian Research

Council [grant 248150/O70]. We thank Professor Emeritus Rafael Romero-Villafranca for reviewing the statistical analysis of this paper.

APPENDIX I

Table 6. Design issues axe.

Design issues	
Lack of option for describing symptom (OR) Whole symptoms missing	Section to mark "no chronic diseases"
	Symptoms such as shortness of breath should be linked to the precipitating factor
	Specify if you are recording on someone's behalf
	Users need more colors to specify the features of sputum. Cannot specify gray.
	In headaches, users advised to set only a description or, first the description, and later, the medical name in parenthesis
Navigation problem	Bug in tobacco reporting makes users jump back to the previous screen
	Help button not seen
	More marked navigation signaling when the symptom starts and ends is needed
	Allow to start by the symptom the user is most concerned
	Clarify the hierarchy of sections signaled with text size is important
Missing functionality or option	Not possible to specify cigars or snus consumption
	Not possible to specify casual smoking
	Not possible to specify that the user has traveled to more than one foreign county
	No option to say that no chronic disease is present
	Missing option to indicate the precipitating factor (e.g. physical activity)
	Option to indicate that information is provided on behalf of someone
	Free text field
	Allow to point out the position of the headache rather than describe the location
	Joints pain symptom missing
Feedback	Mark in red mandatory sections, searching them is difficult
	Provide feedback to mark the change among symptoms and sections
	Say in advance when new subsections are appearing
	Explain well the system in the introduction/welcome

Table 7. General user opinions axe.

General user opinions	
Usefulness perception	"Nothing really to improve, you can leave it as is, seems like a good pattern"
	Good to decide when going to the GP or not rather than search in Google
	Useful for people in general, but specially for parents with children to avoid unnecessary visits to the emergency room
	Useful in areas with poor health infrastructure
	It is considered good that the GP is on the loop and reads what the user reports
	This is better than looking in the internet freely because you get scared, more scientific
	"It's obviously a useful and welcome technology, and it makes the patient reflect on some details that the doctor would have to ask. Puts the patient in the right path"
	Helps to structure symptoms to express them better to the doctor
Too tedious	Too much text in the welcome page
	Many subsections that open gradually to fill in and too much information
	Cant remember so many tobacco details
	It can be difficult to remember all the details of the time patterns of the symptom
Improvement proposal	One page per symptom: delimit better the start and end of each symptom (e.g. "now we start reporting the symptom wheezing…") etc.
	Rephrase chronic disease as a question
	Users considered that medical terms are not fully necessary. A description could be enough
	Order information in drop downs alphabetically
	Add links to websites for those users that want further understanding of a symptoms or condition
	Need examples for intensity and quantity scales to know what each level represents (e.g. sputum volume)
	Only one place (fever) requires the use of the keyboard. Use a drop down so keyboard is not needed at all.
	Help button missing
	Allow to start with the symptom the user is most concern about and continue in order of importance
	Buttons continue & back on the sides of the website should be added
	For medical terms such as headache, the description should go first and the medical name of the disorder should go in parenthesis
	For list of options, a help button should be offered with additional information to understand the data requested
	An extra free text field would be nice to leave a

	note, but it would be useless for automatic systems
	Some users had to repeat the recording of demographics because of the bug in tobacco
	Mark better the hierarchy of sections with different text headers
	Add graphic metaphors for sections such as intensity
	Set units and rephrase when asking about how many weeks elapse until a symptom repeats
Data security concerns	Want to know where data will be stored in a organization of confidence such as the hospital
Context influencing the user experience	"I think in a realistic situation would be a little impatient of all these issues and options, so I had enough perhaps only chosen something to make it go faster."
	If you are color blind the system may have some barriers
	People who are ill would take more time to read details and try harder when blocked
	Gender of the user
	Mood of user
Others (General User opinions)	Some parts difficult to understand what info is requested
	Fairly o no big troubles, accessible and intuitive, user friendly, easy to understand
	Not too medical
	Quite comprehensive
	Easy to read
	A bit too much detail
	Complete
	Scale "generally unwell" considered weird
	Not too much room for improvement

Table 8.Interpretation issues axe.

Interpretation issues		
Lack of clarity when requesting information	Word sputum not understood	
	Reformulate chronic disease request as a question	
	Word trivial is not familiar language	
	Wheezing cough instead of wheezing confuses in wheezing subsections	
	Cluster headache, tension headache and migraine not familiar terms, just describe better (help button was not seen)	
	Some questions formulation not easily understood or are too long	
	English-like words "timing", rephrase in proper Norwegian language	
	"Behavior" is not a good name for the tobacco section	
Time pattern interpretation	Problems expressing that a symptom repeats every X weeks	
	Difference continuous and periodic	
	Unsure about the time pattern of the symptom wheezing	
Vagueness in scales	Difference among	Distinction between

	intensity levels	trivial, mild, moderate…
	Difference among quantity levels	Sputum moderate, copious …not possible to determine the difference among levels
	Difference among time scale	Difference among suddenly, rapid, gradually

References

[1] Col N, Correa-de-Araujo R. Chapter 27 - Consumers and Clinical Decision Support A2 - Greenes, Robert A. Clinical Decision Support (Second Edition), Oxford: Academic Press; 2014, p. 741–69.

[2] Rigby M, Koch S, Keeling D, Hill P, Alonso A, Maeckelberghe E. Developing a New Understanding of Enabling Health and Wellbeing in Europe. European Science Foundation; 2013.

[3] Institute of Medicine (US) Roundtable on Evidence-Based Medicine. The Learning Healthcare System: Workshop Summary. Washington (DC): National Academies Press (US); 2007.

[4] Patient Protection and Affordable Care Act, HR 3590, 2010.

[5] Chewning B, Bylund CL, Shah B, Arora NK, Gueguen JA, Makoul G. Patient preferences for shared decisions: A systematic review. Patient Education and Counseling 2012;86:9–18. doi:10.1016/j.pec.2011.02.004.

[6] Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480.

[7] Jimison HB, Sher PP, Jimison JJB. Decision Support for Patients. In: FHIMSS ESBE FACMI, editor. Clinical Decision Support Systems, Springer New York; 2007, p. 249–61.

[8] NHS. 51 million unnecessary GP visits….NHS investigate why n.d. http://www.selfcareforum.org/wp-content/uploads/2011/07/chhosewellsummercampaignpressrelease.pdf.

[9] Hammond T, Clatworthy J, Horne R. Patients' use of GPs and community pharmacists in minor illness: a cross-sectional questionnaire-based study. Fam Pract 2004;21:146–9.

[10] Strategy Directorate, Department of Health. Support for Self Care in General Practice and Urgent Care Settings A baseline Study. Department of Health; 2006.

[11] AskMD - Get Answers - Manage Conditions and Symptoms. Sharecare n.d. https://www.sharecare.com/askmd/get-started (accessed August 29, 2016).

[12] Enriksen TS, Skrøvseth SO, Yitbarek Yigzaw K, Bellika JG. Er du Syk? n.d. www.erdusyk.no (accessed December 2, 2013).

[13] Symptom Checker - Drugs.com n.d. https://www.drugs.com/symptom-checker/ (accessed August 29, 2016).

[14] Symptom Checker - Mayo Clinic n.d. http://www.mayoclinic.org/symptom-checker/select-symptom/itt-20009075 (accessed March 2, 2015).

[15] National Health System. Symptom checkers - NHS Choices 2014. https://www.nhs.uk/symptomcheckers/pages/symptoms.aspx (accessed January 28, 2015).

[16] Symptom Checker from WebMD. Check Your Medical Symptoms. n.d. http://symptoms.webmd.com/#introView (accessed June 23, 2016).

[17] Traver M, Basagoiti I, Martínez-Millana A, Fernández-Llatas C, Traver V. Experiences of a General Practitioner in the daily practice about Digital Health Literacy. Proceedings of the 38th

Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Orlando (Florida): 2016.
[18] Silver MP. Patient perspectives on online health information and communication with doctors: a qualitative study of patients 50 years old and over. J Med Internet Res 2015;17:e19. doi:10.2196/jmir.3588.
[19] Fraser H, Wyatt J. Chapter 8 - International Dimensions of Clinical Decision Support. In: Greenes RA, editor. Clinical Decision Support (Second Edition), Oxford: Academic Press; 2014, p. 241–67.
[20] Luger TM, Houston TK, Suls J. Older Adult Experience of Online Diagnosis: Results From a Scenario-Based Think-Aloud Protocol. Journal of Medical Internet Research 2014;16:e16. doi:10.2196/jmir.2924.
[21] Elliot AJ, Kara EO, Loveridge P, Bawa Z, Morbey RA, Moth M, et al. Internet-based remote health self-checker symptom data as an adjuvant to a national syndromic surveillance system. Epidemiol Infect 2015;143:3416–22. doi:10.1017/S0950268815000503.
[22] Schwartz LM, Woloshin S, Welch HG. Can patients interpret health information? An assessment of the medical data interpretation test. Med Decis Making 2005;25:290–300. doi:10.1177/0272989X05276860.
[23] Quaglio G, Sørensen K, Rübig P, Bertinato L, Brand H, Karapiperis T, et al. Accelerating the health literacy agenda in Europe. Health Promot Int 2016:daw028. doi:10.1093/heapro/daw028.
[24] Bellika JG, Marco-Ruiz L, Wynn R. A Communicable Disease Query Engine. Studies in Health Technology and Informatics 2015:1012–1012. doi:10.3233/978-1-61499-512-8-1012.
[25] Bellika JG, Hasvold T, Hartvigsen G. Propagation of program control: A tool for distributed disease surveillance. International Journal of Medical Informatics 2007;76:313–29. doi:10.1016/j.ijmedinf.2006.02.007.
[26] Nasjonal IKT. Nasjonal IKT - Tiltak 15.5 Folkeregisteret i helsenettet n.d.
[27] Lærum H, Bakke SL, Pedersen R, Valand JT. An update on OpenEHR archetypes in Norway: Response to article Christensen B & Ellingsen G: “Evaluating Model-Driven Development for large-scale EHRs through the openEHR approach” IJMI May 2016, Volume 89, pages 43-54. Int J Med Inform 2016;93:1. doi:10.1016/j.ijmedinf.2016.05.002.
[28] Bakke SL. National governance of archetypes in Norway. Stud Health Technol Inform 2014;216:1091–1091.
[29] Christensen B, Ellingsen G. Evaluating Model-Driven Development for large-scale EHRs through the openEHR approach. Int J Med Inform 2016;89:43–54. doi:10.1016/j.ijmedinf.2016.02.004.
[30] Moreno-Conde A, Moner D, Cruz WD da, Santos MR, Maldonado JA, Robles M, et al. Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J Am Med Inform Assoc 2015;22:925–34. doi:10.1093/jamia/ocv008.
[31] Marco Ruiz L, Maldonado JA, Karlsen R, Bellika JG. Multidisciplinary Modelling of Symptoms and Signs with Archetypes and SNOMEDCT for Clinical Decision Support. Stud Health Technol Inform., Madrid: IOS press; 2015.
[32] Marco-Ruiz L, Maldonado JA, Traver V, Karlsen R, Bellika JG. Meta-architecture for the interoperability and knowledge management of archetype-based clinical decision support systems. 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2014, p. 517–21. doi:10.1109/BHI.2014.6864416.
[33] Marco-Ruiz L, Moner D, Maldonado JA, Kolstrup N, Bellika JG. Archetype-based data warehouse environment to enable the reuse of electronic health record data. International Journal of Medical Informatics 2015;84:702–14. doi:10.1016/j.ijmedinf.2015.05.016.
[34] Yen P-Y, Bakken S. Review of health information technology usability study methodologies. Journal of the American Medical Informatics Association 2012;19:413–22. doi:10.1136/amiajnl-2010-000020.

[35] ISO/IEC 9126-1:2001 - Software engineering -- Product quality -- Part 1: Quality model. ISO n.d. http://www.iso.org/iso/catalogue_detail.htm?csnumber=22749 (accessed September 13, 2016).
[36] Dumas J, Fox J. Usability Testing. Human–Computer Interaction Handbook, vol. 20126252, CRC Press; 2012, p. 1221–42.
[37] Boland MR, Rusanov A, So Y, Lopez-Jimenez C, Busacca L, Steinman RC, et al. From expert-derived user needs to user-perceived ease of use and usefulness: a two-phase mixed-methods evaluation framework. J Biomed Inform 2014;52:141–50. doi:10.1016/j.jbi.2013.12.004.
[38] Li AC, Kannry JL, Kushniruk A, Chrimes D, McGinn TG, Edonyabo D, et al. Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. International Journal of Medical Informatics 2012;81:761–72. doi:10.1016/j.ijmedinf.2012.02.009.
[39] Davis D, Jiang S. Usability testing of existing type 2 diabetes mellitus websites. International Journal of Medical Informatics 2016;92:62–72. doi:10.1016/j.ijmedinf.2016.04.012.
[40] Lai T-Y. Iterative refinement of a tailored system for self-care management of depressive symptoms in people living with HIV/AIDS through heuristic evaluation and end user testing. International Journal of Medical Informatics 2007;76:S317–24. doi:10.1016/j.ijmedinf.2007.05.007.
[41] Ericsson; KA. Protocol Analysis: Verbal Reports as Data. 2nd Revised edition edition. MIT Press; 1993.
[42] Lewis, C. H. Using the "Thinking Aloud" Method In Cognitive Interface Design (Technical report). IBM. RC-9265; 1982.
[43] van Engen-Verheul MM, Peute LWP, de Keizer NF, Peek N, Jaspers MWM. Optimizing the user interface of a data entry module for an electronic patient record for cardiac rehabilitation: A mixed method usability approach. International Journal of Medical Informatics 2016;87:15–26. doi:10.1016/j.ijmedinf.2015.12.007.
[44] Davis FD. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Q 1989;13:319–340. doi:10.2307/249008.
[45] Kim J, Park H-A. Development of a health information technology acceptance model using consumers' health behavior intention. J Med Internet Res 2012;14:e133. doi:10.2196/jmir.2143.
[46] Jackson JE. A User's Guide to Principal Components. Edición: New edition. Hoboken, N.J: Wiley-Interscience; 2003.
[47] Greenacre MJ, Primicerio R. Multivariate analysis of ecological data. First Edition. Bilbao: Fundación BBVA; 2013.
[48] Flach P. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. 1 edition. Cambridge ; New York: Cambridge University Press; 2012.
[49] Venkatesh V, Davis FD. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Manage Sci 2000;46:186–204. doi:10.1287/mnsc.46.2.186.11926.
[50] Venkatesh V, Bala H. Technology Acceptance Model 3 and a Research Agenda on Interventions. Decision Sciences 2008;39:273–315. doi:10.1111/j.1540-5915.2008.00192.x.
[51] Tedesco DP, Tullis TS. A Comparison of Methods for Eliciting Post-Task Subjective Ratings in Usability Testing ABSTRACT. Proceedings of the UPA 2006 Conference, 2006.
[52] Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Medical Research Methodology 2013;13:117. doi:10.1186/1471-2288-13-117.
[53] Ritchie J, Lewis J, Nicholls CM, Ormston R. Qualitative Research Practice: A Guide for Social Science Students and Researchers. Edición: 2. SAGE Publications Ltd; 2013.
[54] Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977;33:159–74. doi:10.2307/2529310.
[55] The Template Object Model. London, UK: The OpenEHR foundation; 2007.

[56] Medlock S, Eslami S, Askari M, Arts DL, Sent D, de Rooij SE, et al. Health Information–Seeking Behavior of Seniors Who Use the Internet: A Survey. J Med Internet Res 2015;17. doi:10.2196/jmir.3749.
[57] Smith GC, Egbert N, Dellman-Jenkins M, Nanna K, Palmieri PA. Reducing Depression in Stroke Survivors and their Informal Caregivers: A Randomized Clinical Trial of a Web-Based Intervention. Rehabil Psychol 2012;57:196–206. doi:10.1037/a0029587.
[58] Sauro J, Kindlund E. A method to standardize usability metrics into a single score, ACM Press; 2005, p. 401. doi:10.1145/1054972.1055028.
[59] Byrne BM. Factor analytic models: viewing the structure of an assessment instrument from three perspectives. J Pers Assess 2005;85:17–32. doi:10.1207/s15327752jpa8501_02.
[60] McGarvey LPA. Patterns of cough in the clinic. Pulmonary Pharmacology & Therapeutics 2011;24:300–3. doi:10.1016/j.pupt.2011.01.014.
[61] Lanzola G, Parimbelli E, Micieli G, Cavallini A, Quaglini S. Data Quality and Completeness in a Web Stroke Registry as the Basis for Data and Process Mining. Journal of Healthcare Engineering 2014;5:163–84. doi:10.1260/2040-2295.5.2.163.
[62] Grudin J. A moving target: The evolution of HCI. The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications, CRC Press - Taylor & Francis group; 2012, p. xxvii–lxi.
[63] International Organization for Standardization. ISO 9241-210:2010 Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems n.d.
[64] Starcevic V, Berle D. Cyberchondria: towards a better understanding of excessive health-related Internet use. Expert Rev Neurother 2013;13:205–13. doi:10.1586/ern.12.162.
[65] Fox S, Duggan M. Health Online. Pew Research Center's Internet & American Life Project. PewResearchCenter; 2013.