Journal of Medical Systems (2020) 44: 130
https://doi.org/10.1007/s10916-020-01583-w

MOBILE & WIRELESS HEALTH

Usability of an Intelligent Virtual Assistant for Promoting Behavior Change and Self-Care in Older People with Type 2 Diabetes

João Balsa[1] · Isa Félix[2] · Ana Paula Cláudio[3] · Maria Beatriz Carmo[3] · Isabel Costa e Silva[2] · Ana Guerreiro[4] · Maria Guedes[4] · Adriana Henriques[2,6] · Mara Pereira Guerreiro[2,5]

Received: 15 December 2019 / Accepted: 7 May 2020 / Published online: 13 June 2020

Abstract

In the context of the *VASelfCare* project, we developed an application prototype of an intelligent anthropomorphic virtual assistant. Designed as a relational agent, the virtual assistant has the role of supporting older people with Type 2 Diabetes Mellitus (T2D) in medication adherence and lifestyle changes. Our paper has two goals: describing the essentials of this prototype, and reporting on usability evaluation. We describe the general architecture of the prototype, including the graphical component, and focus on its main feature: the incorporation, in the way the dialogue flows, of Behavior Change Techniques, identified through a theoretical framework, the *Behaviour Change Wheel*. Usability was experimentally evaluated in field tests in a purposive sample of 20 participants (11 older adults with T2D and 9 experts). The Portuguese version of the System Usability Scale was employed, supplemented with qualitative data from open questions, diaries, digital notes and telephone follow-ups. The aggregated mean SUS score was 73,75 (SD 13,31), which corresponds to a borderline rating of excellent. Textual data were content analyzed and will be prioritized to further improve usability.

Keywords Intelligent virtual assistants · Usability · Relational agents · Behavior change · *mHealth* applications · Diabetes

Introduction

Worldwide T2D affects approximately 425 million adults (20-79 years) [24]. For example, in the United States around one-quarter of people over the age of 65 years have diabetes and one-half of older adults have pre-diabetes [16]. The global prevalence of T2D continues to increase steadily as more people live longer [49]. Difficulties in adhering to diabetes management are associated with lack of glycemic control [22], which, in turn, leads to a greater risk of heart attack and stroke, sight loss, foot and leg amputation, and renal failure [2, 3, 40]. Diabetes complications can be fatal and are associated with considerable direct and indirect costs for health systems [24].

Self-care entails tasks related, but not limited, to self-prevention, self-diagnosis, self-medication and self-management of chronic conditions. As articulated by other authors, based on the seminal work of Corbin and Strauss [30], self-management of chronic conditions encompasses three distinct sets of tasks: medical or behavioral management, role management and emotional management. One of the key dimensions of our work is supporting the medical management of older people with T2D; another key dimension is computer science (artificial intelligence and computer graphics) — how to develop a virtual assistant, as an intelligent relational agent to serve the aforementioned purpose. These dimensions are central

This article belongs to the Topical Collection: *Artificial Intelligence in Medicine (AIM)*
Guest Editor: Filipe Portela

✉ João Balsa
jbalsa@ciencias.ulisboa.pt

1 Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal

2 Unidade de Investigação e Desenvolvimento em Enfermagem (ui&de), Escola Superior de Enfermagem de Lisboa, Lisbon, Portugal

3 Biosystems & Integrative Sciences Institute (BioISI), Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal

4 Escola Superior de Enfermagem de Lisboa, Lisbon, Portugal

5 Centro de Investigação Interdisciplinar Egas Moniz, Instituto Universitário Egas Moniz, Monte de Caparica, Almada, Portugal

6 Faculdade de Medicina da Universidade de Lisboa, Instituto de Saúde Ambiental (ISAMB), Lisbon, Portugal

to the *VASelfCare* project, from which the application we describe emerged.

A recent systematic review, based on randomized controlled trials of eight self-management education programs for T2D, found these are likely to be cost-effective in the long-term [28]. Similarly, a systematic review of cost-effectiveness of self-management in chronic conditions, conducted as part of the European Union funded project PRO-STEP [42], identified diabetes as the top condition in which self-management was 'value added'. Albeit the scarcity of data, in particular when considering digital technology interventions, available evidence corroborates the notion that self-management of T2D is associated with health gains that can be cost-effective. In particular, PRO-STEP has called for support to innovation and digital technology as self-care facilitators [42].

Digital technology, and notably mobile health interventions, have demonstrated a positive effect on glycemic control in T2D patients [18, 23]. For example, the systematic review by Cui et al. [18] estimated an overall effect of mobile applications in HbA1c (*Glycated hemoglobin*), shown as mean difference (MD), of -0.40% (-4.37 mmol/mol, with a 95% confidence interval -0.69 to -0.11% [-7.54 to -1.20 mmol/mol]; p = 0.007). This effect can be particularly important if achieved by patients on their own, without the additional cost of involving healthcare professionals.

Another relevant issue is the personalization of Diabetes therapy, that can occur at different levels. Donsa et al. [20] identify a set of self-management scenarios and identify a number of challenges that are presented when trying to implement adequate technological solutions, noting the importance of mobile approaches in lifestyle promotion.

One issue that impacts on the use of digital technology is usability. The International Organization for Standardization (ISO), as cited by Maramba [31], defined usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Additionally, the ISO 9241 standard, as cited by Moumane [39], defines effectiveness, efficiency and satisfaction.

As noted by [50], *mHealth* applications are often used by people with little experience of technology, making usability a key factor in its adoption. When targeting older people, specific challenges arise, as identified by [48], where a set of aging barriers are analyzed in relation to the usability of *mHealth* applications. The development of the *VASelfCare* prototype was informed by the usability principles systematized in [4], such as the simple recognition of click-sensitive areas or the use of a *talkback* feature for visually-impaired people.

In [5], we presented the *VASelfCare* mobile application prototype, highlighting aspects related to artificial intelligence. Our prototype resorts to an anthropomorphic virtual assistant that supports medication adherence and lifestyle change (healthy diet and physical activity). Besides revisiting the description of our prototype, in this paper we extend the previous one by presenting and discussing results of the usability evaluation.

The remainder of our paper is organized as follows. In "Related Work", we analyze the most relevant related work. In "Intelligent Relational Interaction", we present the core of our contribution, describing the way the interaction was designed. In "The *VASelfCare* application", we give an overall perspective of the application prototype. In "Usability Evaluation", we present the results regarding usability evaluation. Finally, in "Conclusions", we present some conclusions.

Related Work

Behavior change theory and behavior change techniques in the self-management T2D digital interventions

In their systematic review of computer-based T2D self-management interventions for adults, Pal et al. [41] pointed out that, in the interventions analyzed, the theoretical basis and the behavior change techniques (BCTs), which are the active ingredients of interventions, were not always described. This impairs the generation of evidence on the most effective BCTs to achieve the desired clinical outcomes and curtails replication.

Overall, there is a paucity of evidence on the effects of the active ingredients of multi-component interventions and their impact on changes in HbA1c, in particular concerning *mHealth* solutions using virtual assistants.

We address these issues by carefully designing an evidence-based and theory-driven intervention. The *Behaviour Change Wheel* (BCW) framework [34] offers a structured method of intervention development. The BCW was developed from 19 frameworks of behavior change identified in a systematic literature review. At the BCW core is a model of human behavior — the COM-B — which posits that human behavior (B) results from the interaction between physical and psychological capabilities (C), opportunities provided by the physical and social environment (O) and reflective and automatic motivation (M) [34]. In addition to a holistic scope, a key BCW advantage is the structured link of behavior determinants with relevant intervention functions and BCTs.

The BCW organizes content and components of behavioral interventions into nine *intervention functions* (IFs): restrictions, environmental restructuring, modeling, enablement, training, coercion, incentivization, persuasion and

education. IFs represent how an intervention might change the target behavior. To translate the general IFs into specific replicable techniques, Michie and colleagues [35] recommend the *Behaviour Change Techniques Taxonomy* (BCTTv1).

The BCTTv1 has been validated and is used to design and retrospectively evaluate and aggregate effect sizes of health interventions [36]. Developing an understanding of the theoretical basis of effective interventions can inform future development.

A scoping review [25], on the contents of the eHealth interventions targeting persons with poorly controlled T2D, identified 31 BCTs most frequently used based on the BCTTv1, such as *"instruction on how to perform a behaviour"*, *"adding objects to the environment"*, *"self-monitoring on outcomes of behaviour"*, *"social support (practical)"*, *"feedback on outcomes of behaviour"* and *"prompts/cues"*. These findings are corroborated by previous literature [41, 47]. Pal et al. [41] demonstrated that the most common BCTs with significant impact on HbA1c in computer-based interventions were: *"prompt self-monitoring of behavioural outcomes"* and *"provide feedback on performance"*.

The importance of *mHealth apps* (mobile Health applications, typically for iOS or Android devices), has been accompanying the development of mobile technologies. Bhuyan et al. [7] concluded 60% adults who were *mHealth* apps users considered them useful in achieving health behavior goals. In another study, Morrissey et al. [38] surveyed 166 medication adherence apps to ascertain whether they incorporated BCTs. The authors concluded that, from the 93 possible techniques [35], only a total of 12 were found in the evaluated apps. This result clearly shows that more work is needed in incorporating evidence on BCTs in available applications.

Digital interventions to support self-management of T2D: the case of virtual assistants

One form of virtual assistants are relational agents. These consist of virtual humans designed to build long-term socio-emotional relationships with users. Bickmore et al. (for instance [8–10]) have been the most active group, researching interventions in several areas. Notably, we could not find any published research from this group addressing adults with type 2 diabetes.

Others have developed relational agents as virtual coaches for T2D patients [1, 37]. However, these works lack data on usability or the effect on endpoints of interest.

The rationale for researching and developing an intelligent relational agent in our project was two-fold: firstly, the use of these agents has shown effectiveness and acceptability for older people, including those with limited health literacy [8, 10], and secondly, they may enhance engagement over time.

More recently, a European project [6, 19], researched a set of virtual multi-domain conversational non-anthropomorphic coaches, with the goal of improving life quality for people with T2D. Although this project has similarities with our work, its research is ongoing and evaluation data is currently scarce.

The virtual assistant dialogue manager

Two main types of approaches can be employed to manage the virtual assistant (VA) dialogues: rule-based and machine learning (data-based) approaches. Rule-based approaches to dialogue management have been explored, namely in the context of spoken dialogue systems. As mentioned by [15], examples of handcrafted rule-based systems to manage the *action-selection* process of a dialogue, that is, to determine *what to say next*, are reported by [11, 46] and [29]; the latter also using probabilities. As described in "Intelligent Relational Interaction", our approach differs from these by incorporating an additional (more abstract, also rule-based) level to the dialogue management.

As already explained, the VA we designed is not simply an *assistant*, in the sense that it is expected to answer the user's questions or give support to tasks execution. It is intelligent, as it has a goal (to promote beneficial behavior change in the patient) and proactively pursues that goal. As we opted to follow the relational agents approach (as detailed in "Intervention Design"), and wanted to incorporate specific BCT's, where appropriate, the structure of a daily interaction could not be completely free; instead, it had to obey to a set of well-defined guidelines, represented as rules. This kind of VA behavior could not have been obtained using machine learning approaches, since there is no data (actual dialogues) annotated with the features we need in this context (information regarding BCT application). There is published work regarding the production of annotated dialogues [14, 33], but none incorporating information regarding the use of BCTs. Our work in this first prototype expanded the frontiers of completely hardwired dialogue systems, by adding extra levels of dialogue management, that allow for a more modular and flexible definition of the daily interactions.

Intelligent Relational Interaction

We now describe our approach. We will first explain how the intervention was designed, and then how it was modeled using a rule-based approach.

Intervention Design

The intervention has two distinct phases: 1) *evaluation* and 2) *follow-up*. The main purpose of the evaluation phase is to collect data on the user's characteristics pertaining to the three components: *diet*, *physical activity* and *medication*, for future tailoring of the intervention. In the subsequent *follow-up* phase, the main goal is to promote the desired behavior or to maintain it. In this paper we focus on this *follow-up* phase, since it is the most interesting from an artificial intelligence perspective.

Each daily interaction with the virtual assistant is structured according to the literature on relational agents [8] that recommends the dialogue should follow general sequential steps. These steps are: 1) *opening*; 2) *social talk*; 3) *review tasks*; 4) *assess*; 5) *counselling*; 6) *assign tasks*; 7) *pre-closing*; and 8) *closing*. A detailed description of each of these steps can be found in [13]. The design of the software prototype intervention and the dialogue creation was guided by the BCW [34]. This approach enhanced the development process based on an evidence-based selection of the intervention components (BCTs), ensuring that the intervention targets the underlying determinants of behavior.

Suitable BCTs were incorporated in different steps of the interaction. The selected BCTs are exclusive of a particular dialogue step in all intervention components. In our view, the operationalization in the first two and of the last two dialogue steps does not represent a behavior change technique, according to the BCT taxonomy v1 [35]. Specific BCTs are operationalized according to tailoring principles, including constructs such as lack of knowledge, or by a previously defined behavior target. Figure 1 represents the BCT's distribution according to the standard dialogue steps in the follow-up phase. In this figure, for each step of the dialogue, the specific BCT used is identified by its name and taxonomy code. Depending on the achievement of some behavior goal, the *Counseling* step may include

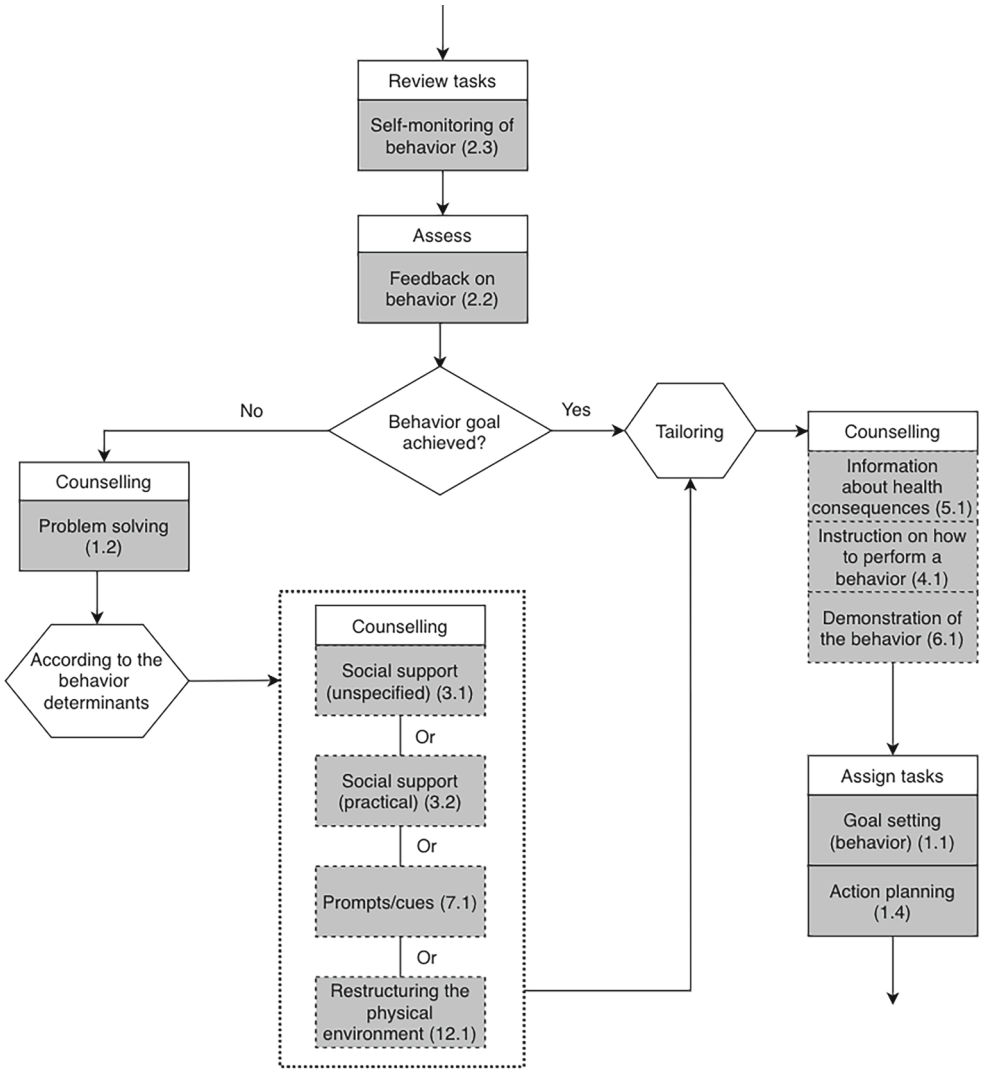

Fig. 1 BCTs distribution according to standard dialogue steps in an interaction. Dashed lines indicate BCTs that are operationalized depending on the context (tailoring or behavior targeting). Numbers represent the BCT code according to BCTTv1

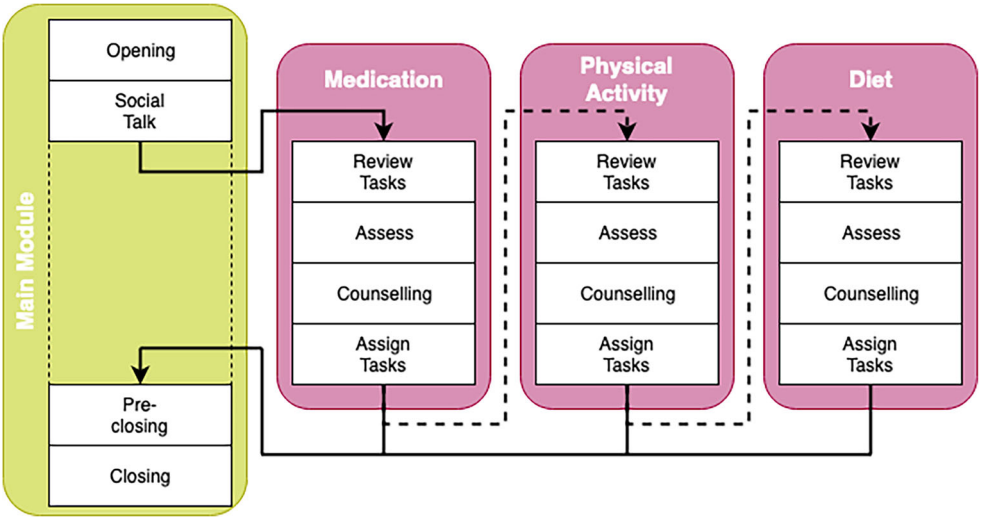

Fig. 2 Module interaction in the rule-based system

the application of some additional BCT, according to the identified behavior determinants.

Rule-based component

In order to incorporate the desired BCTs in the dialogue flow, the interaction is controlled by a rule engine.

This rule-based component corresponds to the definition of a set of *if–then* rules (rules of the form "if *some conditions hold* then *execute some action*"), where the conditions may include contextual information regarding the interaction (e.g., user characteristics, such as age, the date when the interaction takes place, or the answer to some question) and the action can represent the subsequent locutionary act performed by the virtual assistant or the update of some condition (that will allow the execution of other rules). This approach allows greater flexibility in the dialogue definition and in the generation of diversified interactions.

Rules define the flow of the dialogue in two ways. At a lower level, they have the role of representing a handcrafted portion of the dialogue that might, for instance, materialize the local application of a specific BCT. At a higher level, they are responsible for managing the overall interaction flow, for instance, the realization that the dialogue should go to next step, or that a different component should be addressed.

The way this works is illustrated in Fig. 2. This figure represents the flow of the dialog for one day of interaction. There is an organization in modules. The main module includes the four steps — *Opening*, *Social Talk*, *Pre-Closing* and *Closing* — that are always executed; the first two at the beginning and the other two at the end. Then, the modules that incorporate some BCT are started according to the defined protocol (first, *medication*, then *physical activity*, and finally *diet*). The duration of the evaluation phase varies between one and three days, depending on the component. For each component, the follow-up phase depends on patient adherence to the target behavior and can be delivered ranging from eight days to several weeks. In Fig. 2, the dotted lines represent the possibility, inferred or not by the rule engine, of including more than one component in a single interaction. Multi-behavior interactions do not go through all the dialogue steps for each component on a daily basis.

The *VASelfCare* application

The central element of the *VASelfCare* application is an anthropomorphic female character (called Vitória) which plays the role of an intelligent virtual assistant. In Fig. 3 we show a screenshot of the application (taken from the evaluation phase, and with the text translated to English).

Vitória is capable of speaking (articulates speech while the corresponding subtitles appear on the screen) and expresses emotions through facial and body animations. The overall development of the application interface was guided by usability principles for older people with T2D [4]. For example, redundancy of both audio and written information may help reduce any communication shortcomings, such as lower eyesight accuracy and hearing deficits. The virtual assistant chosen for this first prototype is a female 3D model, obtained from Daz3.[1]

Application Architecture

The core of *VASelfCare* technological solution is implemented in *Unity*.[2] Although best known as a game engine, it suits adequately the requirements of our application, with

[1] https://www.daz3d.com

[2] https://unity.com/

Fig. 3 Screenshot of the *VASelfCare* application

the possibility of defining multiple views, each with their own purpose, as well as the transitions between them.

The architecture, firstly introduced in [13], comprises 3 main components: the *Core*, the *Dialogue Creator* and the *Speech Generator* (Fig. 4). The *Core* controls the interface and the flow of the execution. The *Dialogue Creator* is the component used to define the speech of the VA and the choices presented to the user while the interaction with the user is taking place. Finally, the *Speech Generator* creates the audio and viseme files to support the VA's articulated speech.

The *User Interface* module provides several views for the patient: (i) to log-in; (ii) to interact with the virtual assistant; (iii) to register personal data, such as daily number of steps walked, blood glucose levels or weight; (iv) to access personal information, such as, prescribed anti-diabetic medication or charts with the registered data over time; (v) to view the assigned self-care plans; and (vi) to access advice information about the diet, physical activity and medication.

The *Application Controller* manages the flow of the execution being responsible for the logical sequence of the application and for communicating with the other components of the *VASelfCare* Core. The *Dialogue Engine* corresponds to what was described in the previous section.

The *Dialogue Creator* component corresponds to the definition of the handcrafted portion of the dialogues (the lower level mentioned earlier).

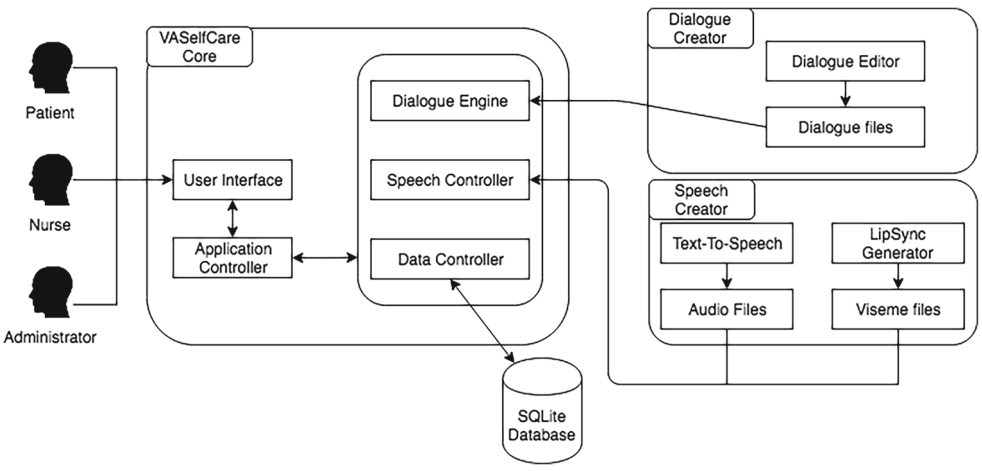

Fig. 4 *VASelfCare* application architecture

Table 1 Data of a fictitious patient for usability testing

−−	Diogo Santos, male, 72 years old
−−	Last HbA1c record: 7.5%
−−	Medication: metformin 1 g (1 tablet after breakfast and dinner) + acarbose 50 mg (1 tablet before lunch and dinner)
−−	Height: 1.76 m
−−	Weight: 97 kg
−−	Waist circumference: 73 cm
−−	Hip circumference: 99 cm
−−	Performs self-monitoring of blood glucose

To convert the written dialogues into audio files a Text-To-Speech software is used, *Speech2Go*.[3] The speech rate has been slowed down, taking in consideration the target population. The *LipSync Generator*, developed in the context of a previous project [17] is used to convert text to viseme files. Audio and viseme files are read simultaneously to provide Vitória's articulated speech.

Usability Evaluation

Participants and procedure

Usability tests were conducted on a purposive sample of end-users and experts. The former included 11 patients recruited in five primary care units of the Portuguese National Health Service. Inclusion criteria were having T2D, age equal or above 65 years, being able to speak and write in Portuguese, and being a frequent user of digital technology (e.g. possessing a smartphone or using social media or a computer). The sample of experts included nine academic nurses with expertise in community health and elderly care, recruited in the Lisbon School of Nursing.

Tablets deployed for the tests had the following characteristics: operating system Android 7.0, CPU - Snapdragon 435, RAM 3 GB, ROM 32 GB.

A test case (Table 1) was depicted in the prototype and presented to participants. Then, end-users were asked to undertake a trial period of 26 days at home, guided by a list of suggested tasks, which included talking to Vitória daily. Experts were asked to test the prototype independently over a shorter period (8 to 10 days), by means of multiple interactions with Vitória on the same day. In both cases the trial period covered the evaluation and follow-up phases for anti-diabetic agents, physical activity and healthy eating.

While end-users could provide answers in the intelligent dialogues as they wished, experts were requested to provide these answers in accordance to one of the profiles described in Table 2. The purpose of this approach was covering the full range of possibilities in the intelligent dialogues, yielding potentially more meaningful data.

Both end-users and experts were given the option of reporting problems and suggestions throughout the trial in a paper-based diary or a digital notepad, embedded in the prototype. A telephone number for support and weekly follow-up was also made available to end-users.

Data were collected at the end of the trial period through the Portuguese version of the System Usability Scale (SUS) [32]. Additionally, the end-users' questionnaire had three open questions (what they most and less enjoyed, suggestions) and two closed-questions (e.g. future training). Experts' questionnaire had the same three open questions plus a single closed-question on the dialogues' subtitles, similar to the end-users' questionnaire. Questionnaires were administered face-to-face by a member of the research team (end-users) or on-line (experts).

Numeric data were inputted into a database, individually double-checked for accuracy and subjected to descriptive and bivariate statistical analysis with the aid of SPSS (version 25). Textual data from the open questions, diaries, digital notes and follow-up calls were gathered in a single file and subjected to content analysis, by deriving main themes and categories under these themes. Phrases or text segments were then labelled under each category.

Results and Discussion

Table 3 summarizes the sociodemographic characteristics of the participants.

The aggregated mean SUS score was 73.75 (SD 13.31). There were no significant differences between mean end-users' and experts' scores (76.59, SD 12.26 versus and 70.2; SD 14.43; U = 37.0; $p = 0.34$).

Based on Brooke [12], the aggregated mean SUS score is boarder-line for *excellent* usability; the mean end-users' score corresponds to an *excellent* rating whilst the mean experts' score rates *good*. Another approach to interpret results is looking at percentiles; the aggregated mean SUS score of 73.75 corresponds to a percentile of 70%, meaning that the prototype has higher perceived usability than 70% of all systems products in the Sauro [43] database (as cited in [44]).

Table 4 presents the themes and categories that emerged from the analysis of the textual data, with examples. The numbers between brackets represent the count of textual data elements in that category. Categories not represented in the table include typos or other language mistakes (40). Other comments referred to the test conditions. For example, some participants felt that it would have been preferable to have their own medication in the application prototype. This may have influenced their subjective perception of usability.

[3] https://harposoftware.com

Table 2 User profiles for usability testing by experts

Profile	Adherence to medication	Adherence to physical activity	Adherence to diet
A	+	+	+
B	+	+	-
C	+	-	+
D	+	-	-
E	-	+	+
F	-	+	-
G	-	-	+
H	-	-	-

Another group of comments pertained to prototype features, such as Victória's "metallic voice", the pronunciation of some words, or the fact that it is not possible to repeat the interaction more than once every day.

The next immediate step is prioritizing usability problems uncovered during field testing to define the urgency of remedial actions. One approach to prioritization is using a four-point scale, from "1" most severe problems to "4", the least severe [26]. Another approach is resorting to severity categories; *high* (failure in task execution), *medium* (not so severe, task can be executed) and *low* (minor problems) [26].

One of the strong points of our work is the multi-method evaluation of usability, comprising a questionnaire (SUS) and qualitative data from different sources. Questionnaires, albeit the most common method for assessing usability, have been criticized for yielding only an overall measure without indicating the issues that need to be tackled [31]. Another strength is resorting to a validated version of the SUS, which is a frequently used questionnaire [31]. This approach enables comparison across usability tests.

Many usability tests are conducted in a laboratory setting, where participants perform tasks in a controlled environment [26, 39]. Whilst this enables control of the experiment and facilitates data collection, especially when ascertaining metrics of effectiveness or efficiency through observation, it is not representative of real-world use. Our choice of field testing appears to be endorsed by the extent of qualitative data obtained, associated with a more prolonged contact with the prototype.

In relation to the sample size, calculations tend to rely on aspects such as problem discovery and completion rate. The former was of relevance for our study. For example, if the goal is to find 85% of all problems (or, at least of the most obvious problems), in the assumption that there is a 0.3 probability that a user detects the occurrence of a usability problem, a sample of five users suffices [27]. A sample size calculator available at https://measuringu.com/problem_discovery/ may be of interest to those conducting usability studies. While a 5-users sample is a good baseline from a problem-discovery perspective, we pursued additional suggestions and insights, as evidence by the multimodal data collection instruments employed (open questions, diaries, digital notes and telephone follow-ups). Therefore, sample size estimation was based on previous experience and resources available. Published studies using SUS show large variations in sample size, from 2 to 373 users [31].

Some authors [39] attempted to use a maximum variability sample in usability tests, considering characteristics such as experience with technology, gender, level and the nature of education and the occupation. Our patient sample was skewed towards expert technology users. While these users may potentially raise more issues and suggestions, future

Table 3 Sociodemographic features of participants

Feature		**End-users**	**Experts**
Gender	Female	27.3% (n=3)	88.9% (n=8)
	Male	72.2% (n=8)	11.1% (n=1)
Mean age (years)		70.91	54.33
		[max: 80, min: 67]	[max: 63, min: 40]
Education	college degree	63.6% (n=7)	-
	secondary education	36.4% (n=4)	
Mean experience in the use of technology usage (years)		6.09 [max: 20, min:0]	-
Mean professional experience (years)		-	31.78
			[max: 40,min: 19]

Table 4 Themes and categories that emerged from textual data analysis

Themes	Categories	Examples
Positive aspects (29)	Global opinion (26)	"Easy to use" [SCOND02]
	Language (3)	"Simple language" [SJUL02]
Aspects for further improvement (213)	Technology (20)	"Buttons for recording medication are too small"
		"the graph for number of daily steps is out of scale" [DAF02]
	New functionalities (53)	"Strategy to remind users to use the app" [SJUL01#1]
		"Development of recipes for people with diabetes" [DEL01]
	Content	Repetitions (16): "Daily questions about medication intake feel repetitive" [DAF01]
		Clarification (16): "Clarify what 'two doses of alcohol' means" [Profile E]
		Incongruences (10): "After recording the number of daily steps, the app asks: 'Are you sure that you intend to record the number of intake?' [Profile C]
		Amount of information (4): "4th day: interaction with too much information" [Profile E]
		Meals (7): "(. . .) it doesn't include soup. Most elders eat soup at meals" [Profile A]
		Medication (14): "Too much focus on medication" [Profile G]
		Physical activities (9): "Include 'tiredness' as a possible reason for not having met the daily steps goal" [Profile G]
		Others (21): "Reply options for the user always affirmative, without option for refusal" [Profile G]
	Training (3)	"There should exist a more extended session to demonstrate the use of the app" [SJUL02#V]

usability tests should target also novice technology users, which may bring different perspectives. The level of education of our sample is not representative of the Portuguese senior population either. The average schooling of older adults in Portugal is primary education [21]. Future research should also encompass less educated users, who may have a different perspective on usability.

Another area of future research is measuring efficacy and efficiency, which in addition to satisfaction are key attributes of usability. This could be done, for example, by examining task completion.

Reviews on the methods for usability testing call for the deployment of automation, which remains infrequent [31, 50]. For example, eye tracking, used to see the movement of users' eyes when performing tasks, has been explored in the usability evaluation of an online diabetes exercise system [45]. Another possibility is monitoring the application usage and task completion remotely. The feasibility of using these automated tests merits exploration in future research.

Conclusions

The Intelligent Virtual Assistant we developed is underpinned by a robust theoretical approach in behavior change, concurring to original features of the dialogue flow and the model supporting the interaction with users. The potential benefits of the human appearance of our virtual assistant, designed to communicate verbally in a helpful-cooperative style and express emotions according to the dialogue context, will be ascertained in future research.

The aggregated mean SUS score obtained, which is borderline between good and excellent, is encouraging.

Content analysis of textual data yielded insights on issues and opportunities for improvement, which will be prioritized and implemented to further strengthen usability.

Altogether, usability data will contribute to develop a new version of the application prototype, which will be subjected to evaluation in the next phase of the project: a non-randomized non-controlled feasibility trial.

Acknowledgments The authors are indebted to other *VASelfCare* team members for their contribution to the software development (http://vaselfcare.rd.ciencias.ulisboa.pt/); Pedro Alves' and Pedro Neves' contributions were instrumental for the work herein reported.

Funding This work was supported by FCT and Compete 2020 (grant number LISBOA-01-0145-FEDER-024250). It is also supported by UID/MULTI /04046/2019 Research Unit grant from FCT, Portugal (to BioISI).

Compliance with Ethical Standards

Conflict of interests All authors declare that they have no conflicts of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Besides, the study protocol, which encompasses iterative tests with older people with T2D and health professionals during the software development phase, was granted ethical approval from the Portuguese authorities[4] (6104/CES/2018 ARSLVT). This article does not contain any studies with animals performed by any of the authors.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. op den Akker, R., Klaassen, R., Lavrysen, T., Geleijnse, G., van Halteren, A., Schwietert, H., and Van der Hout, M., A Personal Context-Aware Multi-Device Coaching Service That Supports a Healthy Lifestyle. In: *Proceedings of the 25th BCS Conference on Human-Computer Interaction, BCS-HCI '11, pp 443-448, BCS Learning & Development Ltd., Swindon, GBR*, 2011.
2. American Diabetes Association: 12. older adults: Standards of medical care in diabetes—2019. Diabetes Care 42 (Supplement 1), S139–S147. https://doi.org/10.2337/dc19-S012, 2019. https://care.diabetesjournals.org/content/42/Supplement_1/S139.
3. American Diabetes Association: 5. lifestyle management: Standards of medical care in diabetes—2019. Diabetes Care 42 (Supplement 1), S46–S60. https://doi.org/10.2337/dc19-S005., 2019. https://care.diabetesjournals.org/content/42/Supplement_1/S46.
4. Arnhold, M., Quade, M., and Kirch, W., Mobile Applications for Diabetics : A Systematic Review and Expert-Based Usability Evaluation Considering the Special Requirements of Diabetes Patients Age 50 Years or Older Corresponding Author :. *Journal of Medical Internet Research* 16(4):1–18, 2014. https://doi.org/10.2196/jmir.2968.
5. Balsa, J., Neves, P., Félix, I. B., Guerreiro, M. P., Alves, P., Carmo, M. B., Marques, D., Dias, A., Henriques, A., and Cláudio, A. P., Intelligent Virtual Assistant for Promoting Behaviour Change in Older People with T2D. In: P.M. Oliveira, P. Novais, and L.P. Reis (Eds.) *Progress in Artificial Intelligence - 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3-6, 2019, Proceedings, Part I, Lecture Notes in Computer Science, vol. 11804, pp. 372–383. Springer*, 2019. https://doi.org/10.1007/978-3-030-30241-2_32.
6. Beinema, T., op den Akker, H., and Hermens, H., Creating an artificial coaching engine for multi-domain conversational coaches in eHealth applications. In: E. André, T.W. Bickmore, S. Vrochidis, and L. Wanner (Eds.) *Proceedings of the AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications co-located with the Federated AI Meeting (FAIM 2018), ICAHGCA@AAMAS 2018, Stockholm, Sweden, July 15, 2018, CEUR Workshop Proceedings, vol. 2338, pp. 35–39. CEUR-WS.org*, 2018. http://ceur-ws.org/Vol-2338/paper5.pdf.
7. Bhuyan, S. S., Lu, N., Chandak, A., Kim, H., Wyant, D., Bhatt, J., Kedia, S., and Chang, C. F., Use of Mobile Health Applications for Health-Seeking Behavior Among US Adults, Vol. 40. https://doi.org/10.1007/s10916-016-0492-7, 2016.
8. Bickmore, T. W., Caruso, L., and Clough-Gorr, K., Heeren, T.: 'It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers* 17(6):711–735, 2005. https://doi.org/10.1016/j.intcom.2005.09.002.
9. Bickmore, T. W., Puskar, K., Schlenk, E. A., Pfeifer, L. M., and Sereika, S. M., Maintaining reality: Relational agents for antipsychotic medication adherence. *Interacting with Computers* 22(4):276–288. https://doi.org/10.1016/j.intcom.2010.02.001, 2010.
10. Bickmore, T. W., Silliman, R. A., Nelson, K., Cheng, D. M., Winter, M., Henault, L., and Paasche-Orlow, M. K., A randomized controlled trial of an automated exercise coach for older adults. *Journal of the American Geriatrics Society* 61(10):1676–1683. https://doi.org/10.1111/jgs.12449, 2013.
11. Boye, J., Dialogue management for automatic troubleshooting and other problem-solving applications. In: *Proceedings of the 8th SIGDial workshop on discourse and dialogue, pp. 247–255.*, 2007. https://sigdial.org/files/workshops/workshop8/Proceedings/SIGdial45.pdf.
12. Brooke, J., SUS: A retrospective. *Journal of Usability Studies* 8(2):29–40, 2013.
13. Buinhas, S., Cláudio, A. P., Carmo, M. B., Balsa, J., Cavaco, A., Mendes, A., Félix, I., Pimenta, N., and Guerreiro, M. P., Virtual Assistant to Improve Self-care of Older People with Type 2 Diabetes: First Prototype. In: *Gerontechnology: First International Workshop, IWoG 2018, pp. 236–248. Springer, Cham.*, 2019. https://doi.org/10.1007/978-3-030-16028-9_21. http://link.springer.com/10.1007/978-3-030-16028-9_21.
14. Bunt, H., Petukhova, V., Malchanau, A., Fang, A., and Wijnhoven, K., The dialogbank: dialogues with interoperable annotations. *Language Resources and Evaluation* 53(2):213–249. https://doi.org/10.1007/s10579-018-9436-9, 2018.
15. Burgan, D., Dialogue Systems & Dialogue Management. Tech. rep., National Security & ISR Division Defence Science and Technology Group, 2017.
16. collab=Centers for Disease Control and Prevention, C.D.C.: National diabetes statistics report: estimates of diabetes and its

[4] *Administração Regional de Saúde de Lisboa e Vale do Tejo* (Regional Health Administration of Lisbon and Tagus Valley)

burden in the united states. Atlanta GA: centers for disease control and prevention; 2017, 2017.
17. Clàudio, A. P., Carmo, M. B., Pinto, V., Cavaco, A., and Guerreiro, M. P., Virtual humans for training and assessment of self-medication consultation skills in pharmacy students. In: *10Th international conference on computer science and education, pp. 175–180. ICCSE*, 2015.
18. Cui, M., Wu, X., Mao, J., Wang, X., and Nie, M., T2DM self-management via smartphone applications: a systematic review and meta-analysis. *PLoS ONE* 11(11):1–15, 2016.
19. Das, K., Beinema, T., Akker, H. O. D., and Hermens., H., Generation of multi-party dialogues among embodied conversational agents to promote active living and healthy diet for subjects suffering from type 2 diabetes. In: *Proceedings of the 5th International Conference on Information and Communication Technologies for Ageing Well and e-Health - Volume 1: ICT4AWE*, pp. 297-304, INSTICC, SciTePress, 2019. https://doi.org/10.5220/0007750602970304.
20. Donsa, K., Spat, S., Beck, P., Pieber, T. R., and Holzinger, A., towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges. In: *Smart health, pp. 237–260. Springer international publishing*, 2015. https://doi.org/10.1007/978-3-319-16226-3_10.
21. Fundação Francisco Manuel dos Santos, Pordata: base de dados Portugal contemporâneo. Accessed: 2020-03-10. https://www.pordata.pt/DB/Ambiente+de+Consulta/Tabela.
22. Garcia-Pèrez, L. E., Àlvarez, M., Dilla, T., Gil-Guillèn, V., and Orozco-Beltràn, D., Adherence to therapies in patients with type 2 diabetes. *Diabetes Therapy* 4(2):175–194, 2013. https://doi.org/10.1007/s13300-013-0034-y.
23. Hou, C., Carter, B., Hewitt, J., Francisa, T., and Mayor, S., Do Mobile Phone Applications Improve Glycemic Control (HbA1c) in the Self-management of Diabetes? A Systematic Review, Meta-analysis, and GRADE of 14 Randomized Trials. *Diabetes Care* 39(11):2089–2095, 2016.
24. International Diabetes Federation : IDF Diabetes Atlas, 8th edn. International Diabetes Federation, Brussels, Belgium, 2017.
25. Kebede, M. M., Liedtke, T. P., Möllers, T., and Pischke, C. R., Characterizing Active Ingredients of eHealth Interventions Targeting Perso ns With Poorly Controlled Type 2 Diabetes Mellitus Using the Behavior Change Techniques Taxonomy: Scoping Review, Vol. 19. https://www.jmir.org/2017/10/e348/, 2017.
26. Kekàlàinen, A., Kaikkonen, A., Kankainen, A., Cankar, M., and Kallio, T., Usability testing of mobile applications: a comparison between laboratory and field testing. *Journal of Usability Studies* 1(1):4–17, 2005.
27. Lewis, J. R., Sample sizes for usability tests. *interactions* 13(6):29. https://doi.org/10.1145/1167948.1167973, 2006.
28. Lian, J. X., McGhee, S. M., Chau, J., Wong, C. K. H., Lam, C. L. K., and Wong, W. C. W., Systematic review on the cost-effectiveness of self-management education programme for type 2 diabetes mellitus. *Diabetes Research and Clinical Practice* 127:21–34. https://doi.org/10.1016/j.diabres.2017.02.021, 2017.
29. Lison, P., A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech &, Language* 34(1):232–255, 2015.
30. Lorig, K. R., and Holman, H. R., Self-management education: History, definition, outcomes, and mechanisms. *Annals of Behavioral Medicine* 26(1):1–7, 2003. https://doi.org/10.1207/S15324796ABM2601_01.
31. Maramba, I., Chatterjee, A., and Newman, C., Methods of usability testing in the development of eHealth applications: A scoping review. *International Journal of Medical Informatics* 126:95–104. https://doi.org/10.1016/j.ijmedinf.2019.03.018, 2019. http://www.sciencedirect.com/science/article/pii/S1386505618313182.
32. Martins, A. I., Rosa, A. F., Queirós, A., Silva, A., and Rocha, N. P., European portuguese validation of the system usability scale (sus). *Procedia Computer Science* 67:293–300. https://doi.org/https://doi.org/10.1016/j.procs.2015.09.273, 2015. Proceedings of the 6th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion.
33. Merdivan, E., Singh, D., Hanke, S., Kropf, J., Holzinger, A., and Geist, M., Human annotated dialogues dataset for natural conversational agents. *Applied Sciences* 10(3):762, 2020. https://doi.org/10.3390/app10030762.
34. Michie, S., Atkins, L., and West, R., The behaviour change wheel : a guide to designing interventions Silverback Publishing, 2014.
35. Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M. P., Cane, J., and Wood, C. E., The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine* 46(1):81–95, 2013. https://doi.org/10.1007/s12160-013-9486-6.
36. Michie, S., Wood, C. E., Johnston, M., Abraham, C., Francis, J. J., and Hardeman, W., Behaviour change techniques: The development and evaluation of a taxonomic method for reporting and describing behaviour change interventions (a suite of five studies involving consensus methods, randomised controlled trials and analysis of qualitative da. *Health Technology Assessment* 19(99):1–187. https://doi.org/10.3310/hta19990, 2015.
37. Monkaresi, H., Calvo, R. A., Pardo, A., Chow, K., Mullan, B., Lam, M., M.Twigg, S., and Cook, D. I., Intelligent diabetes lifestyle coach. In: *Proceedings of the Fifth International Workshop on Smart Healthcare and Wellness Applications (SmartHealth'13), pp. 1–4*, 2013.
38. Morrissey, E. C., Corbett, T. K., Walsh, J. C., and Molloy, G. J., Behavior change techniques in apps for medication adherence: A content analysis. *American Journal of Preventive Medicine* 50(5):e143–e146. https://doi.org/10.1016/j.amepre.2015.09.034, 2016.
39. Moumane, K., Idri, A., and Abran, A., Usability evaluation of mobile applications using iso 9241 and iso 25062 standards. *SpringerPlus* 5(1):548, 2016. https://doi.org/10.1186/s40064-016-2171-z.
40. OECD/EU: Health at a Glance: Europe 2018: State of Health in the EU Cycle. OECD Publishing, Paris, https://doi.org/10.1787/fd41e65f-es, 2018. https://doi.org/10.1787/health_glance_eur-2018-en.
41. Pal, K., Eastwood, S. V., Michie, S., Farmer, A. J., Barnard, M. L., Peacock, R., Wood, B., Inniss, J. D., and Murray, E., Computer-based diabetes self-management interventions for adults with type 2 diabetes mellitus. Cochrane Database of Systematic Reviews 2013, Issue 3. Art. No.: CD008776. DOI: 10.1002/14651858.CD008776.pub2.
42. PRO-STEP Project Consortium: PRO-STEP project: Promoting Self-management for chronic diseases in Europe - Final Report, 2018.
43. Sauro, J., A practical guide to the system usability scale: background, benchmarks & best practices CreateSpace Independent Publishing Platform, 2011.
44. Sauro, J., and Lewis, J. R., Quantifying the user experience : practical statistics for user research Elsevier/Morgan Kaufmann, 2012.
45. Schaarup, C., Hartvigsen, G., Larsen, L. B., Tan, Z., Årsand, E., and Hejlesen, O. K., Assessing the potential use of eye-tracking triangulation for evaluating the usability of an online

diabetes exercise system. In: I.N. Sarkar, A. Georgiou, and P.M. de Azevedo Marques (Eds.) *MEDINFO 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics, São Paulo, Brazil, 19-23 August 2015, Studies in Health Technology and Informatics, vol. 216, pp. 84–88. IOS Press*, 2015. https://doi.org/10.3233/978-1-61499-564-7-84.

46. Smith, C., Crook, N., Dobnik, S., Charlton, D., Boye, J., Pulman, S., de la Camara, R. S., Turunen, M., Benyon, D., Bradley, J., Gambàck, B., Hansen, P., Mival, O., Webb, N., and Cavazza, M., Interaction strategies for an affective conversational agent. *Presence:, Teleoperators and Virtual Environments* 20(5):395–411, 2011.
47. Van Vugt, M., De Wit, M., Cleijne, W. H., and Snoek, F. J., Use of behavioral change techniques in web-based selfmanagement programs for type 2 diabetes patients: systematic review. *J Med Internet Res* 15(12):e279. https://doi.org/10.2196/jmir.2800, 2013.
48. Wildenbos, G., Peute, L., and Jaspers, M., Aging barriers influencing mobile health usability for older adults: A literature based framework (mold-us). *International Journal of Medical Informatics* 114:66–75. https://doi.org/10.1016/j.ijmedinf.2018.03.012, 2018. http://www.sciencedirect.com/science/article/pii/S1386505618302454.
49. World Health Organization: Diabetes, 2018. https://www.who.int/news-room/fact-sheets/detail/diabetes.
50. Zapata, B. C., Fernández-Alemán, J. L., Idri, A., and Toval, A., Empirical studies on usability of mhealth apps: A systematic literature review. *Journal of Medical Systems* 39(2):1. https://doi.org/10.1007/s10916-014-0182-2, 2015.