



Clinical-chatbot AHP evaluation based on “quality in use” of ISO/IEC 25010

Vita Santa Barletta^a, Danilo Caivano^a, Lucio Colizzi^{a,*}, Giovanni Dimauro^a, Mario Piattini^b

^a University of Bari Aldo Moro, Computer Science Department, Via Edoardo Orabona, 4, 70125 Bari, Italy

^b University of Castilla-La Mancha, Alarcos Research Group, Information Systems and Technologies Institute, Ciudad Real, Spain

ARTICLE INFO

Keywords:

Medical-chatbot quality
Clinical pathway
AHP
ISO/IEC 25010

ABSTRACT

Background: Conversational agents are currently a valid alternative to humans in first-level interviews with users who need information, even in-depth, about services or products. In application domains such as health care, this technology can become pervasive only if the perceived “quality in use” is appropriate. How to measure chatbot quality is an open question. The international standard ISO/IEC 25010 proposes a set of characteristics (effectiveness, efficiency, satisfaction, freedom from risk, and context coverage) to be considered when the “quality in use” of a software system has to be measured.

Basic procedure: This study proposes a clinical chatbot comparison method based on quality. The proposed approach is based on Analytic Hierarchy Process methodology (AHP).

Findings: Our contribution is twofold. First, we propose a set of measures for each characteristic of ISO/IEC 25010 according to three classes of functionality: providing information, providing prescriptions and process management. Moreover a quantitative method is proposed for making homogeneous the pairwise weights when the AHP is used for the “quality-in-use” comparison. As a case study, a comparison of two versions of a chatbot was performed.

Conclusions: The results show that the proposed approach provides an effective reference base for performing quality comparisons of medical chatbots compliant with the ISO/IEC 25010 standard.

1. Introduction

Chatbots are special programs that interact with users by simulating a human conversation. Development platforms dedicated to chatbots are becoming increasingly established [1]. In the common sense, this software is based on artificial intelligence algorithms. Many solutions are based on the implementation of decision trees and rule-based conversation [2] or other simple mechanisms aiming to understand context. While chatbots have enormous success in areas such as product and service sales, marketing, entertainment and public administration [3], they are still not widespread in the field clinical domain [4]. The effectiveness of chatbots is certainly indisputable when it is necessary to bring users closer to information about a product or service. As far as the medical sector is concerned, situations tend to become more complicated because of the critical nature of the element on which it is necessary to make in-depth assessments: the responsibility for health and the risks that follow. As underlined in [5], chatbots in the health care domain can play an important role in optimizing resources only if

their quality is demonstrated and measured.

In the clinical domain, there are several chatbot-based experiments, but few applications have actually entered people’s lives [6]. Personalization of services, case management, user-centric dialogue, active and tireless guidance 24 h a day, real-time answers without having to wait in line, and immediate help without moving from home are all situations that are typical in the clinical domain, and they are examples of where chatbots could improve life only if they provide certified quality services. Surely, the aspect related to health safety and responsibility are the main reasons chatbots could be helpful. To lay the foundations for the diffusion of clinical chatbots, it is necessary to control all aspects related to health safety and user responsibility such as internal quality, external quality, and in-use quality [7]. This objective can only be achieved by applying quality assessment methods that helps bridge the gap between software metrics and software product quality factors [8]. In [9], it emerged that, from the quality-assessment point of view, it is often more effective to specialize a chatbot (by developing several of them), instead of having only one that discusses the whole domain.

* Corresponding author.

E-mail address: lucio.colizzi@uniba.it (L. Colizzi).

<https://doi.org/10.1016/j.ijmedinf.2022.104951>

Received 8 September 2022; Received in revised form 23 November 2022; Accepted 1 December 2022

Available online 13 December 2022

1386-5056/© 2022 Elsevier B.V. All rights reserved.

Some papers partially addressed the quality of the chatbot by measuring the quality of the dialogue [10] [11] [12]. In [13], a set of chatbots operating in the business environment were analysed according to 10 characteristics: visual look of the chatbot, form of implementation on the website, speech synthesis, basic or specialized knowledge base, presentation of knowledge and additional functionalities, conversational abilities, language skills and context sensitiveness, personality traits, personalization options, emergency responses in unexpected situations and possibility of rating chatbot. Therefore, it is required to define the design goals for the product (the first aspects of Quality by Design) and consequently the target product quality profile [14].

The ISO/IEC 25000 to ISO/IEC 25099 series of International Standards is entitled Systems and software engineering – Systems and software Quality Requirements and Evaluation, hence the acronym: 'SQuaRE'. SQuaRE has simplified the analysis from its predecessor, ISO 9126. The ISO/IEC 25010 standard [15] defines two models for quality measurement: "quality in use" and "product quality". The former is "the degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, freedom from risk and satisfaction in specific contexts of use" and includes characteristics that relate to the outcome of interaction when a product is used in a particular context of use, applicable to the complete human-computer system, including both computer and software in use. The latter aims to evaluate the factual quality characteristics of the software/system product." Moreover, since the "quality in use" is also closely related to the quality of the information that is conveyed to the user, it is important to consider some characteristics defined in the ISO/IEC 25012 standard [16] and the related measures reported in ISO/IEC 25024 [17].

ISO/IEC 25010: 2011 provides the leading models for assessing software product [18]. In [19] ISO/IEC 25010: 2011 is used to obtain the elements that will be tested by the method of Analytical Hierarchy Process (AHP). The research is using only 3 out of 6 characteristics contained in the ISO/IEC 25010: 2011 which is being transformed into a distributed questionnaire for 10 respondents of IT magister lecturers. Authors of [20], by means of the ISO/IEC 25010 quality model, lists a set of requirements of gamified Blood Donation (BD) Apps. A checklist was established to analyse the influence of the identified requirements on 30 software product quality characteristics. Some quality characteristics were more impacted by BD apps requirements than others, namely, functional suitability, operability, reliability, performance efficiency and security. The authors of [21] try to identify which are the quality measures defined in ISO/IEC 25010 that are currently being used in the software industry. For this purpose, a literature review has been carried out from 27 articles identifying 269 quality measures in total. Finally, a set of quality characteristics have been defined: Functional Suitability, Performance Efficiency, Compatibility, Usability, Reliability, Security, Maintainability, Portability, Safety, Evolvability, and Usable Security. In [22] quality testing on the website bios portal using the ISO 25010: 2011 method on email has been carried out. Also in this study, the test is done by calculating the weight calculation for the six parameters using the AHP method. ISO 25010: 2011 has also been the reference model for quality assessment of online gaming software [23]. The ISO 25010: 2011 standard has been considered as a reference model for quality measurement of other types of software such as information systems or the more complex Enterprise Resource Planning [24][25].

In this paper, we propose a method for evaluating the "quality in use" of clinical chatbots according to ISO/IEC 25010 standard. The proposed method is based on the analytic hierarchy process (AHP) [26].

The Analytic Hierarchy Process (AHP) is a multi-criteria decision analysis methodology which allows the best alternative to be selected from a discrete set of alternatives [27][28][29].

The method is based on the values and judgements, both quantitative and qualitative, of individuals and groups, determined according to a multi-level hierarchical structure in order to obtain priorities.

As shown in the Fig. 1 AHP is structured in a series of steps, by

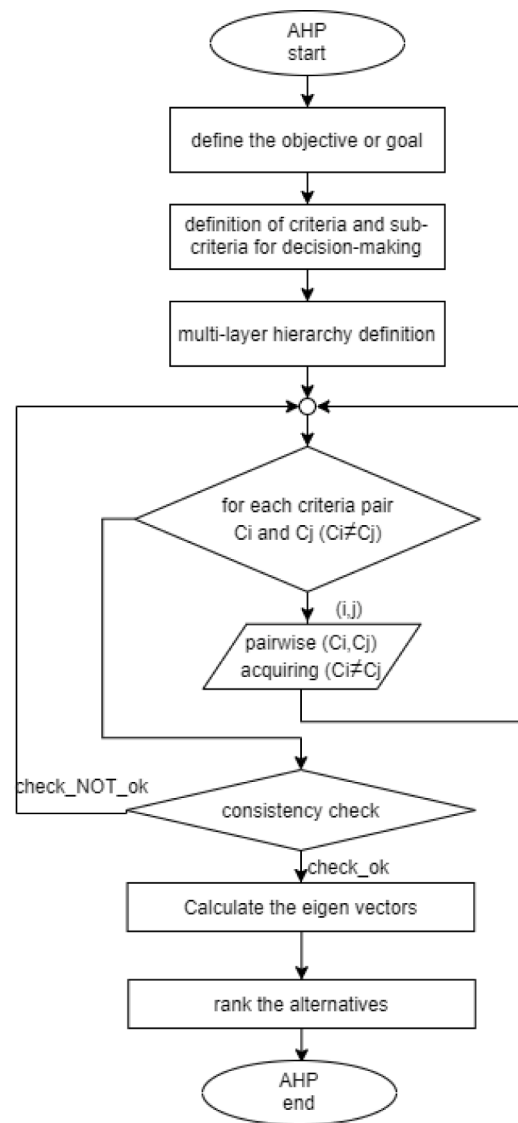


Fig. 1. analytical hierarchy process (AHP) flowchart diagram.

applying which the initial, usually complex and unstructured problem can be decomposed into a hierarchy that is easier to understand and evaluate.

The hierarchical structure is a linear structure formed as follows:

1. At the root is the final objective (goal) that the decision maker intends to achieve;
2. In the intermediate levels are the criteria and possible sub-criteria for decision-making.
3. In the leaves, are placed the alternatives.

The judgements are based on subjective interpretations, often expressed in verbal language and transformed into numbers by means of Saaty's ratio scale, which transforms the judgements into absolute scores between 1 and 9, where 1 represents the equality of the two criteria and the value 9 the extreme importance of one criterion over the other. The final judgement is calculated as a weighted average of the judgements of all decision-makers.

The numbers assigned by the decision maker in the comparisons are organised in a positive, reciprocal, square matrix on the main diagonal, called the pairwise comparisons matrix. All decision elements in a hierarchical level are compared to each other in pairs, by means of a preference ratio, to obtain local priorities. Then, applying the principle

of hierarchical composition, the priorities of the alternatives, called global priorities, are calculated.

AHP is just one of the so-called multi-criteria decision making (MCDM) methods. Other approaches are based on mathematical programming such as best worst method (BWM) or technique for order preference by similarity to ideal solution (TOPSIS) [30][31][32]. AHP is widely used for software product selection as the criteria definition phase is quite agile compared to other domains [33].

The novelty of our contribution is as follows: 1) a set of measures is proposed for each ISO/IEC 25010 characteristic and 2) a quantitative method is proposed for making homogeneous the pairwise weights when the AHP is used for the "quality-in-use" comparison.

The paper is organized as follows. Section 2 reports the state of the art of assessing clinical chatbot quality. In Section 3, three clinical chatbot dimensions for quality-in-use assessment are described. In Section 4, we propose a novel approach for the application of the ISO/IEC 25010 standard in the quality assessment of clinical chatbots. In Section 5, we report an example where we compare the "quality in use" between two versions of a clinical chatbot. Section 6 reports conclusions and outlines some future research work.

2. Clinical chatbots class of functionality and their quality assessment

In this section we will look at the most frequent classes of functionality that engage clinical chatbots and the latest approaches to assess their quality.

2.1. Clinical chatbot classes of functionality

Analysing the functionalities made available by the most famous chatbots operating in the clinical domain, the main field of application is diagnosis [34]. The chatbots are programmed to interview the user in a comprehensible language, providing insights to better understand their questions. The algorithmic approach of these chatbots is typically based on decision-trees, while the more advanced (few) use artificial intelligence algorithms to refine the interview strategy.

Another application for clinical chatbots is scheduling reported in [35]. In this case, the bot becomes an intelligent diary ready to remind the patient to take medicines, go to a scheduled medical examination or head out to a laboratory to perform one or more clinical examination prescribed by the physician.

Other interesting uses for chatbots (but here we are still in the world of basic research) are their application as facilitators within the so-called "integrated care pathways". A clinical pathway is a method for the patient-care management of a well-defined group of patients during a well-defined period [36]. The aim of a clinical pathway is to improve the quality of care, reduce risks and increase both patient satisfaction and efficiency in resource usage [37]. From this definition emerges the concept of a process that must be appropriately designed to describe a clinical pathway. To interpret the expected flows of activities prescribed by a clinical process, the chatbots have to deal with the absence of standardization, both in the description of each phase and the level of language used.

2.2. Clinical chatbot quality assessment

In [38], a hospitality index is defined for a set of specific quality attribute (modifiability, privacy and security, interoperability, reliability) as an indicator of how effective the platform is in achieving that attribute. In [39], the quality assessment of chatbots is addressed through the integration of AHP and quality function deployment (QFD) methods. In [40], the AHP method is proposed for quality assessment to compare either different versions of the same chatbot or the "as-is" version and others under development. This research work is general with respect to the way in which pairwise weights are established. In

[41], the naturalness evaluation of a chatbot system has been made by comparing human-to-human dialogues with human-to-machine dialogues. An ISO 9241-based questionnaire was used by the authors in [42] [43] to evaluate the usability of the eMMA chatbot. The evaluation criterion is qualitative, and the characteristics are all considered as a whole. In [44], the authors proposed a method for quality measurement with the aim of testing a framework of deviations from the correct text (divergents) to verify the correctness of the chatbot reaction. User satisfaction is a key feature of the quality evaluated in [45]. Through a comparison test, it is shown that quality increases if the chatbot integrates external knowledge compared to being closed. Additionally, in this case, the metrics are qualitative and evaluated by users through a posttest questionnaire. A recent study [46] reviewed the technical metrics used for the evaluation of chatbots applied in the health domain and revealed a "lack of standardization and paucity of objective measures". However, the authors underline that the quality assessment must be based not only overall but also on different perspectives. In fact, the metrics of the work included in the review were classified according to four different areas: global metrics regarding chatbots as a whole, metrics related to response generation, metrics related to response understanding and metrics related to aesthetics. Additionally, in [47], solutions based on conversational agents have been studied along three different dimensions: diseases, skills and technological enablers. The aim was to assess how much and how chatbots induce a change in patient behaviour. Beyond the interesting results on the three perspectives, the authors highlighted a future growing trend in the use of chatbots in health care by users of all ages. This happens both because of the ageing population and because the channels of access to these technologies are mobile channels of natural use by young people (*consumability*). It is crucial, therefore, to invest in methods and techniques that aim to measure and certify the quality of these technologies. The usability test proposed by [48] addresses ten topics. Each topic covers a specific class of functionalities (i.e., start anamnesis, change data, check protocol) or an interaction modality (i.e., say goodbye, feeling good dialogue, explanation modality). In [49] a chatbot solution based on predefined answer sets is proposed. The "quality in use" was also measured in this case through a 7-point Likert scale questionnaire. Based on the questionnaire feedback, different types of statistical methods were used for the quality assessment. Some contributions have helped to understand what to measure when assessing dialogue quality. In [50] a study was conducted on the linguistic accuracy of chatbots when interacting with English as a Second Language (ESL) students. The analysis of the responses provided by 5 chatbots focused on two evaluation perspectives: grammatical accuracy and meaning accuracy.

It is worth pointing out that a chatbot can be seen as an intelligent/adaptive interactive system. AHP is widely used also for the evaluation of solutions within the scope of this technology family [51]. From what we have amply reported above, it emerges that the need to measure the quality of chatbot technologies is an open question. The approaches are typically based on empirical experimentation. Without denying the effectiveness of tailor-made methodologies, it is important to invest increasingly in the direction of the standardization of the quality assessment process. This can be done by incorporating elements acquired from international standards that define guidelines (characteristics) on which domain-specific measures can be defined. The purpose of our work is to contribute to this direction. In particular, for clinical chatbots, we define as measures some features whose presence definitely improves the "quality in use". The measures refer to the characteristics of ISO/IEC 25010. With respect to these measures, we propose a quantitative method for making homogeneous pairwise weights in the application of AHP to determine the "quality in use". This implies that the calculation method is general with respect to the definition of specific measures that are defined according to the specific quality goal to be assessed.

3. The three clinical chatbot quality dimensions

For the purpose of this article, a set of clinical chatbots was studied to extract features to which the ISO/IEC 25010 quality model could be applied. Appendix B reports the entire set of analysed chatbots. Analyzing the functionalities of chatbots in Appendix B by means of UML (Unified Modelling Language) Reverse Engineering methodology applied to the chatbot use-cases [52], it emerged that the most recurrent classes of functionalities relate to user interactions are the following:

- **providing information.** This is perhaps the most widespread interaction in medical chatbots. The user wants to deepen knowledge on a topic, ask for information on a health issue, ask for an opinion, etc. The chatbot can reply with predetermined answers (possibly) enriched with semantic annotations, conveying certified information. A chatbot might also be able to improve its answers over time. Additionally, the activity of reminding the patient of events (taking medicine, doing a clinical analysis, meeting the doctor, etc.) can be seen as a form of information provision.
- **providing prescriptions.** As reported in Appendix B, many medical chatbots try to acquire a description of symptoms to guide diagnosis. Some chatbots are even able to provide recommendations on therapy or medical or specialist examinations. In this dimension, health safety must be guaranteed. This is achieved either through coded paths or through mediation by medical and clinicians.
- **process management.** It is a specific way of interacting with the patient with the objective of obtaining context information to understand the state of progress of a clinical process or a flow of activities. The typical situation is that of clinical pathways, clinical algorithms, guidelines expressed in the form of processes, etc. In these cases, the chatbot asks questions that have the purpose of:
 - understanding if a certain task has been performed,
 - reminding the next task to be performed,
 - intercepting if the patient has deviated from the standard clinical path,
 - collecting useful knowledge with the aim of foreseeing which are the next tasks in the standard process.

We will refer to these three classes of functionality as the clinical chatbot dimensions. Moreover, the intersection of these three classes of functionalities is not empty, which means that the chatbot can provide information during the recommendation of a therapy or integrate the therapy in a clinical process. Similarly, an integrated care pathway typically contains continuous feedback to the patient in terms of information and prescriptions. The question is: with respect to these classes of functionalities (chatbot dimensions), what is the software quality perceived by the user? To give an answer, we propose to cross-reference each chatbot dimension with the characteristics provided by the ISO/IEC 25010 standard to assess "quality in use". For each dimension-characteristic pair we define a set of quality measures. The method for quantifying these quality measures is described in Section 4.

4. Evaluation of "quality in use" for clinical chatbots

The proposed method is based on the procedure defined by [40] where a structured goal oriented approach (Analytic Hierarchy Process - AHP) is used for navigating complex decision-making processes that involve both qualitative and quantitative considerations. With respect to this work our contribution is twofold: tailoring the method to the quality characteristics defined by ISO/IEC 25010 and defining a quantitative method for calculating AHP pairwise. The steps in summary are:

1. creating a hierarchy of quality attributes;
2. selecting appropriate measures to represent each attribute;
3. constructing pairwise comparisons between the quality attributes;

4. creating comparison matrices and compute the first principal eigenvector of each one to assess relative and global priority;
5. combining the priorities and compute inconsistency factors to determine which product option best satisfies the hierarchy of quality attributes.

We will use the structure of the quality model defined in ISO/IEC 25010 for the "quality in use" evaluation of clinical chatbots. It is important to emphasize that the proposed method aims to compare the "quality in use" between two different chatbots or two versions of the same chatbot. To achieve this objective, we will focus on the characteristics and sub-characteristics identified by the above standard. The proposed method consists of defining quality measures for each characteristic indicated by ISO 25010 and for each dimension outlined in Section 3 for clinical chatbots: Providing information, Providing prescriptions, and Process management. The method consists of three phases:

Phase 1: Cross-reference each ISO/IEC 25010 characteristic with the three clinical chatbot dimensions identified (providing information, providing prescriptions, process management). Given a specific quality characteristic and a chatbot dimension, we define a set of measures (Table 1) on the basis of three distinct sources:

1. interviews with medical stakeholders and users;
2. measures proposed in other published contributions [50,13,40];
3. measures derived from other ISO standards and applicable to this context for data quality [16,17].

The ISO/IEC 25022 standard [53] already defines the measures for each of the characteristics of ISO/IEC 25010. Unfortunately, the generic nature of these measures does not allow the implementation of a quality comparison tailored to a specific application domain. This issue becomes even stricter when the quality evaluation is carried out along several perspectives. However, we can say that in our proposal, some measures can be merged to make them fall within some set of "quality in use" measures defined in ISO/IEC 25022. For example, in the "Efficiency" characteristics, the measures "real-time information" and "web service information instead of physical logistics" directly influence the measure "time efficiency". An example of matching between the measures of ISO/IEC 25022 and the measures proposed in Table 1 has been provided in Appendix C. Moreover, some characteristics of ISO/IEC 25012 [16] (and consequently, the relative measures in ISO/IEC 25024 [17]) have been used as measures transforming them into functionality, whose presence or absence represents lower or higher "quality in use".

Phase 2: Differences between the compared chatbots are weighed. For this purpose, a value is assigned to each measure reported in Table 1. The value $\{0, 1\}$ expresses a binary measure where true and false means, respectively, existence or nonexistence (a characteristic as whole, a behaviour, a certain functionality, etc.), while a discrete range measure such as $\{1, 2, 3, \dots, n\}$ expresses an ordinal categorical measure [scarce, insufficient, sufficient, discrete, good]. We refer to the degree of discretization (n) as score granularity, which depends on the measure and on the type of judgement it is necessary to express.

Phase 3: Determine the AHP Saaty score for each measure. For this purpose, we define in Table 2 a rule associating to each pair of compared chatbots a pairwise integer value belonging to the range $1 \dots 9$.

To address this scoring issue, and according to the procedure defined in [40], we exploited the AHP method. In our domain, however, we have two different perspectives of application: the former is the importance of the quality-in-use model characteristics and subcharacteristics of ISO/IEC 25010, and the latter aims to weigh the importance of the three chatbot dimensions. In this step, the AHP method is used on the data obtained in the previous step. The construction of the decision tree will depend on the objective to be achieved in the quality evaluation.

Considering two clinical chatbots, calculating the nine-level pairwise weight defined by our method requires considering the type of measure.

Table 1
Clinical chatbot "quality in use" proposed measures.

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	PROVIDING INFORMATION	PROVIDING PRESCRIPTIONS	PROCESS MANAGEMENT
Effectiveness		<i>accuracy and completeness with which users achieve specified goals</i>	<ul style="list-style-type: none"> • text only • semantic annotation • figure & video • accurate speech synthesis • meets neurodiverse needs 	<ul style="list-style-type: none"> • provide prescription • provide suggestion • formal sending (pdf, email, legalmail) 	<ul style="list-style-type: none"> - indirect process information grasping for better answers and process management
Efficiency		<i>resources expended in relation to the accuracy and completeness with which users achieve goals</i>	<ul style="list-style-type: none"> • real time information • low cost/free information predilection • web service information instead of physical logistic 	<ul style="list-style-type: none"> - product/service suggestion 	<ul style="list-style-type: none"> • low finalized interaction for information grasping (indirect knowledge building) • patient/medics interaction • patient/PA interaction • tasks alignment • times alignment • costs alignment
Satisfaction	Usefulness	<i>degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the resultsof use and the consequences of use</i>	<ul style="list-style-type: none"> • accuracy related to lineguide • accuracy related to territory • completeness • consistency respect the EBM • personalized information 	<ul style="list-style-type: none"> • concreteness and practicability • lineguide resondance • personalized prescriptions/ suggestions • supplied in time 	
	Trust	<i>degree to which a user or other stakeholder has confidence that a product or system will behave as intended</i>	<ul style="list-style-type: none"> • certified by third medical-parties (credibility) • mediated by doctors • linked to the sources • personalized information • supported by feedback from others • psychological support • gracefgul degradation • effective function allocation • -gramatical fit • meaning fit • visual look 	<ul style="list-style-type: none"> • certified by doctors • linked to scientific, lineguide, EBM sources 	<ul style="list-style-type: none"> • real pathway state corispondance • completeness in the examination of the patient's datalog • predict in advance the next tasks to be performed • connect all the stakeholder in the clinical pathway • performing tasks privileging solutions, open, low cost, public heath based, obtaining the same autcome • effective function allocation
	Pleasure	degree to which a user obtains pleasure from fulfilling their personal needs			<ul style="list-style-type: none"> - effective function allocation
	Comfort	<i>degree to which the user is satisfied with physical comfort</i>	<ul style="list-style-type: none"> • multichanneling • human like interaction • linguistic accuracy of output • multimedia interaction • on demand and real time information retraveing 	<ul style="list-style-type: none"> - direct virtual interaction with clinical stakeholder 	<ul style="list-style-type: none"> - use of IoT for heath parameter measuring
Freedom from Risk	Economic Risk Mitigation	degree to which a product or system mitigates the potential risk to financial status, efficient operation, commercial property, reputation or other resources in the intended contexts of use	<ul style="list-style-type: none"> - information accompanied by economic and financial rights 	<ul style="list-style-type: none"> • consider whether an insurance policy has been taken out • providing price benchmark 	<ul style="list-style-type: none"> - case management
	Health and Safety Risk Mitigation	<i>degree to which a product or system mitigates the potential risk to people in the intended contexts of use</i>	<ul style="list-style-type: none"> • robustness to manipulation • certified information • care giver involvement • medics involvement • provide mechanisms to avoid interaction during travel 	<ul style="list-style-type: none"> • lineguide compliant • EBM compliant • validated by medics • care giver involvment 	<ul style="list-style-type: none"> • Avoid tasks/token error • protecd and respect privacy • care giver involvment • Avoid inappropriate utterrances and be able to perform damage control • off line personal health datalog access • ranking process state grasping • history of execution tracking • off road detection • patient/medics interaction

(continued on next page)

Table 1 (continued)

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	PROVIDING INFORMATION	PROVIDING PRESCRIPTIONS	PROCESS MANAGEMENT
	Environmental Risk Mitigation	<i>degree to which a product or system mitigates the potential risk to property or the environment in the intended contexts of use</i>	- provide mechanisms to avoid interaction during travel	- provide mechanisms to avoid interaction during travel	- provide mechanisms to avoid interaction during travel
Context Coverage	Context Completeness	<i>degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in all the specified contexts of use</i>	<ul style="list-style-type: none"> • providing information • linking information to other similar user feedback • providing mechanism to ranking information depending user objectives 	<ul style="list-style-type: none"> • providing prescriptions or recommendations or suggestions • provide mechanisms for formulating different hypotheses (ex. diagnosis) on which to give prescriptions or recommendations 	- providing integrated clinical pathway support
	Flexibility	<i>degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in contexts beyond those initially specified in the requirements</i>	<ul style="list-style-type: none"> • patient centered language • medical stakeholder language • care giver involvement 	- robustness to anespected input	<ul style="list-style-type: none"> • robustness to unclearness and enoughtness infomration in the datalog patient • robustness to anespected input

Table 2

Pairwise calculated for each pair Score granularity-dissimilarity.

		Score granularity (n)							
		2	3	4	5	6	7	8	9
d*(n-1)	0	1	1	1	1	1	1	1	1
	1	9	5	5	3	3	3	3	3
	2		9	7	5	5	5	5	3
	3			9	7	7	5	5	5
	4				9	9	7	7	5
	5					9	9	7	7
	6						9	9	7
	7							9	9
	8								9
	9								

If the measure is binary, it means that the pairwise is 1 if there are no changes between two compared chatbots. Otherwise, the pairwise is 9 if the behaviour expressed by the measure is present in one chatbot but not in the second. The sign (i.e., -9, +9) denotes which of the two compared chatbots has the behaviour.

In the case of a discrete range measure, it is proposed that the pairwise calculation is carried out as follows. Let M be an integer discrete measure [1, ..., n], where n is the score granularity. Let R1 and R2 represent the rank positions of the measure values associated with the first and second chatbots, respectively. The dissimilarity d between two chatbots is defined as $d = |R2 - R1| / (n - 1)$ [54]. For example, if the measure is [scarce, insufficient, sufficient, discrete, good] and the first and second chatbots have been evaluated as insufficient and discrete, respectively, then R1 and R2 are equal to 1 and 4, respectively, with dissimilarity $d = 3/4$. If the dissimilarity $d = 0$, this means that the two chatbots have the same rating on measure M. This case corresponds to a value of 1 on the AHP scale from 1 to 9. If $d = 1$, then the dissimilarity is maximum, and the corresponding AHP pairwise is 9.

Consider the two points on the Cartesian axis $P = (d, \text{pairwise})$: $P1 = (0,1)$; $P2 = (n-1,9)$. We assume that any $P = (d, \text{pairwise})$ can be modelled as a point of the linear segment $P1-P2$. Then, we have that the linear relationship between d and pairwise is the following:

$$\text{pairwise} = 8 * d + 1 \tag{1}$$

Since Formula (1) represents a continuous linear function, we have to select the AHP score value that minimizes the absolute error. For this reason, given a dissimilarity value d and a score granularity n, the corresponding pairwise value is rounded to the nearest Saaty score value. It

is worth noting that Formula (1) still holds when only odd values of the Saaty Score are considered (i.e., 1,3,5,7,9).

Furthermore, we observe that Formula (1) is also valid for a binary measure, i.e., when d is equal to 0 or 1 and $n = 2$. Therefore, since $2 \leq n \leq 9$ the pairwise values are reported in Table 22.

The proposed method is general and includes the possibility of using a perception-based approach in pairwise definition. The generalization lies precisely in the fact that the granularity n of a measure is not bound to a predefined value but can vary according to the particular measure defined for a particular characteristic.

In the case of a binary measure expressing the existence or absence of a feature or a chatbot behaviour, it is necessary to think carefully about the use of value 9. This value could be reduced or increased, according to the functionality importance, by accepting the consequences, as stated in [55].

Furthermore, it is important to note that if a measure has a granularity n expressed through a scale of ratios, Formula (1) determines the pairwise values simply through scale homogenization by linear stretch, which is a conventional method by which numerical response options are stretched to a common range (in our case, the 1 to 9 AHP levels). In the case of verbal response options, as described in the above example, homogenization is based on the rank number, regardless of the semantics of the wording used to label the options. This method may introduce errors if the responses to the specific measurement are not of single-peaked symmetric distribution. This limit can be circumvented by applying alternative scaling methods such as scale homogenization by semantic judgement of response options or scale homogenization using a reference distribution [56].

5. Example of chatbot "quality in use" comparison

Assistente Sanitario ("Health care Assistant") [36] is an experimental chatbot developed by our research team with the main task of managing clinical documentation in a patient-centric way. Originally, the chatbot represented a sort of documental suitcase that the patient could invoke, when necessary, without carrying all clinical documentation in paper mode. A further function of the chatbot was to provide clinical information of a generic nature by semantically annotating it on Wikipedia using an online service [57].

Our V. 1 version of the chatbot specialized in Providing information (Fig. 2 (a)), while in version V. 2, we plan to extend its functionality to implement Process management (Fig. 2 (b)), while nothing will change in the dimension Providing prescriptions.

The 'quality in use' assessment has been conducted by means of AHP

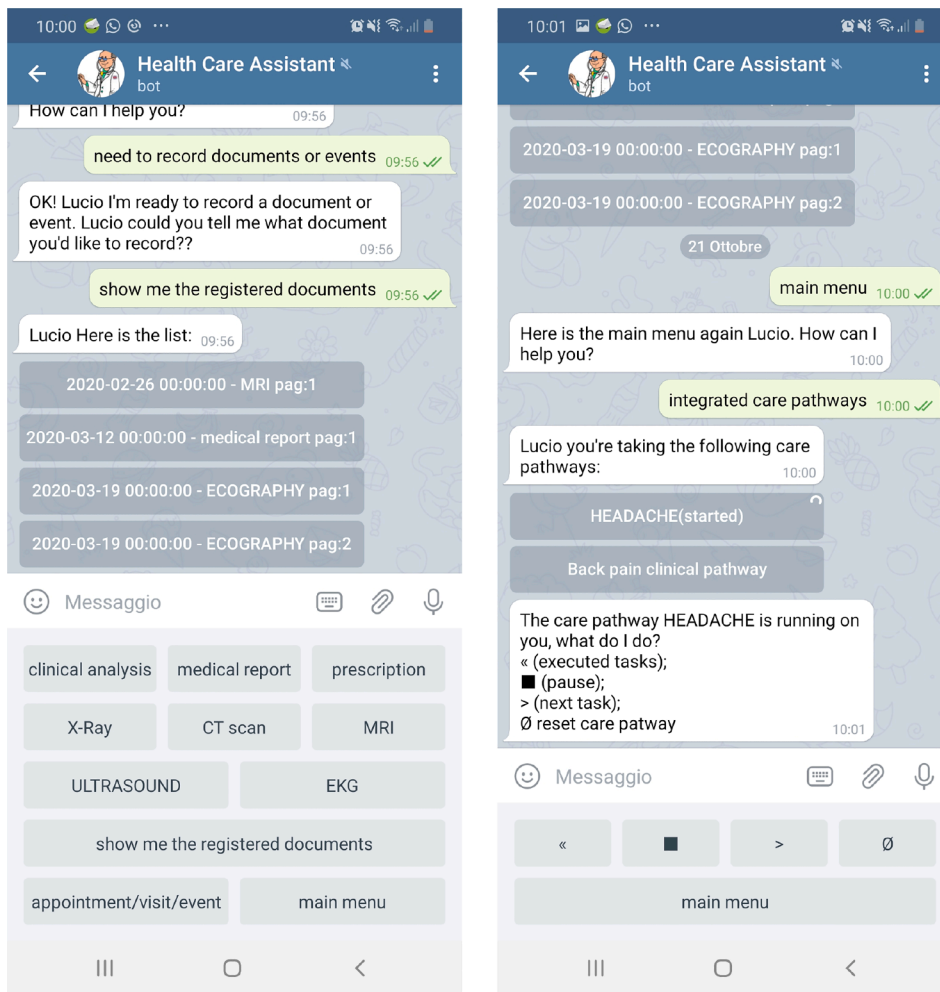


Fig. 2. Health Care Assistant GUI. V. 1 (a); V. 2 (b).

Table 3
"effectiveness" values for all measures proposed.

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	Providing information (measures)	Assistente Sanitario V. 1		Assistente Sanitario V. 2	
			Score granularity (n)	Q_Score	pairwise	Q_Score
Effectiveness	accuracy and completeness with which users achieve specified goals	text only	2	1	1	1
		semantic annotation	2	1	1	1
		figure & video	2	1	1	1
		accurate speech synthesis	5	1	1	1
		meets neurodiverse needs	2	0	1	0
ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	providing prescription (Measures)	Assistente Sanitario V. 1		Assistente Sanitario V. 2	
			Score	Q_Score	pairwise	Q_Score
		Effectiveness	granularity (n)			
Effectiveness	accuracy and completeness with which users achieve specified goals	provide prescriptions	2	0	1	0
		provide suggestions	2	1	1	1
		portal sending (pdf, email, legalmail)	2	0	1	0
ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	process management (measures)	Assistente Sanitario V. 1		Assistente Sanitario V. 2	
			Score	Q_Score	pairwise	Q_Score
Effectiveness	accuracy and completeness with which users achieve specified goals	indirect process information grasping for better answers and process management	2	0	9	1

where the pairwise value was calculated using the method introduced in Section 4. All values have been assigned by a group of 15 patients who used the chatbot for their migraine illness on the basis of a questionnaire. In Table 3, we report the measure values only for the "effectiveness" characteristic together with the pairwise values calculated according to formula 1 (see Appendix A for all remaining measures and characteristics).

If the Score granularity is 2, then the measure is binary. Otherwise, it is an ordinal categoric, and the value it takes (Q_score) belongs to the range 1 to Score granularity. Therefore, when the value of Score granularity is 2 and the Q_Score value is 1, the chatbot exhibits the behaviour expressed by the measure (e.g., a specific function has been implemented); otherwise, it is 0. On the other hand, if the measure is of the ordinal categorical type, the Q_Score represents its specific rank position where $1 \leq Q_Score \leq Score\ granularity$. The column pairwise is calculated according to Formula (1), and it is important to emphasize that in the Process management section, there has been a major upgrade of the chatbot from V. 1 to V. 2.

The previously calculated data can now be used for quality evaluation between the two versions of the same chatbot. In V. 1 the chatbot was exclusively developed to provide information, while in the future, V. 2 will also be enabled to manage clinical processes.



For calculation purposes, the free Superdecision software [58] implementing the AHP method was not used. The hierarchical model designed for the considered case is shown in Fig. 3. The model contemplates not only the characteristics of ISO 25010 but also the three medical chatbot dimensions: *providing information*, *providing prescriptions*, and *process management*.

As it was easy to predict, the version V. 2 version of the Chatbot Health care Assistant has a measured "quality in use" definitely higher than version V. 1. Indeed, the result of the comparison is reported in Table 4.

6. Discussion

The result derived from the proposed method seem to be reasonable, but the following question must be asked: does this result reflect reality? If we consider the chatbot in its entirety and measure it against the total characteristics of the ISO/IEC 25010, it certainly does. Let us see what happens if we compare the two chatbots exclusively on the *providing information* dimension. To do this, we removed the other two dimensions

Table 4 Superdecision output. Alternative rankings for Providing Information, Providing prescriptions and Process management.

Graphic	Alternatives	Total	Normal	Ideal	Ranking
	Assistente Sanitario V. 2	0.5265	0.8320	1.0000	1
	Assistente Sanitario V. 1 0.1063	0.1680	0.2019	2	

from the hierarchical model and reran the calculation. The result is shown in Table 5.



As we can see, in this case, Chatbot V. 2 has a worse "quality in use" than V. 1. This is explained by the following chain of events. Enabling a chatbot to manage integrated clinical processes implies the continuous involvement of medical stakeholders to ensure medical safety.

In fact, the *process management* dimension has implied a remarkable improvement for subcharacteristic health and safety risk mitigation. However, adding *process management* functionality in chatbot version V. 2 also introduces penalties on *providing information* in real time, that is, a lower value for "on demand and real time information retrieving" or "real time information". This is not a technological limit but an application domain constraint, where some human validation steps are mandatory in the integrated care pathway.

These results imply that a global improvement of "quality in use" does not necessarily mean an improvement in each single dimension.

Although the proposed method has been tested by comparing two versions of the same chatbot, it can be generalized to n chatbots in a quite natural way. In particular, for each alternative, it is necessary to acquire pairwise values with respect to its parent level in the AHP hierarchy. For example in Fig. 3 it has to be acquired the pairwise value with respect to the measures defined for each of the following quality dimensions: *providing information*, *providing prescriptions* and *process*

Table 5 Superdecision output. Alternative Rankings for Providing Information.

Graphic	Alternatives	Total	Normal	Ideal	Ranking
	Assistente Sanitario V. 2	0.1253	0.4720	0.8938	2
	Assistente Sanitario V. 1	0.1402	0.5280	1.0000	1

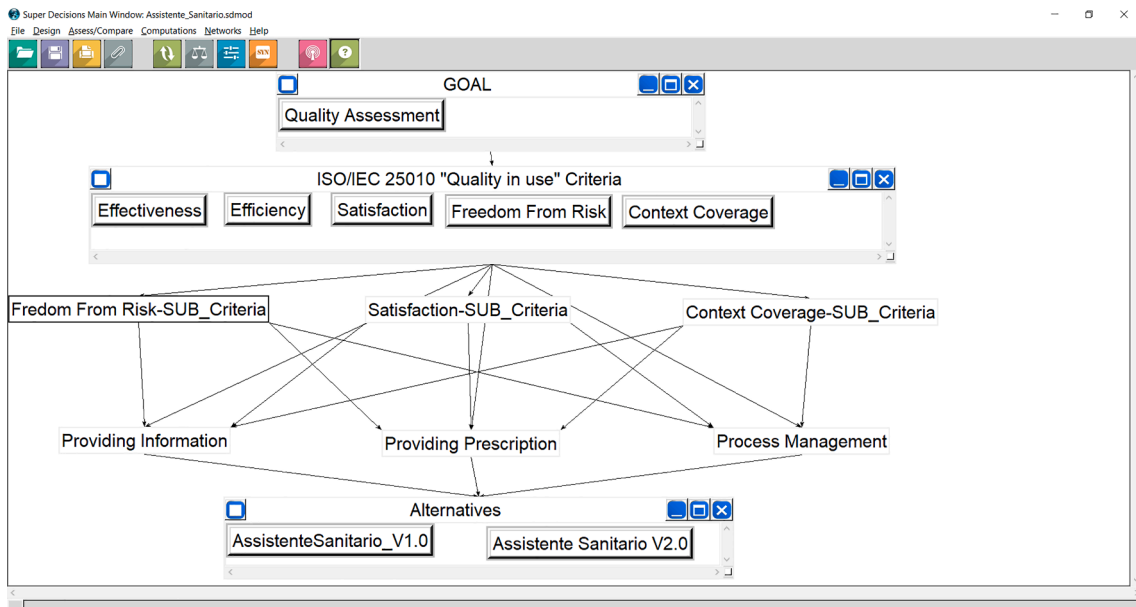


Fig. 3. Superdecision output. Alternative rankings for providing information, providing prescriptions and process management.

management.

It is important to point out that the proposed measures are developed according to the three dimensions that represent the main interactions between the user and conversational agent. The main implication of this organization of quality assessment lies in the potential for analysis that will be possible.

The proposed method certainly contributes to providing a reference base for performing a quality comparison of clinical chatbots compliant with the ISO/IEC 25010 standard. An evaluation fully compliant with the standard should also include the measures that in the proposed approach have been identified outside the ISO framework. The standardization of the set of proposed measures represents the first issue that should be addressed in future research work.

7. Conclusions and future research

The transition from the classic search for information on the web to a human-like interaction with a chatbot certainly introduces issues about process design and interaction quality. In this paper, we have highlighted the importance of assessing the quality of chatbots operating in the clinical domain.

Our contribution is twofold. First, we proposed a set of measures for each characteristic of ISO/IEC 25010 according to three classes of functionality: 1) providing information, 2) providing prescriptions and 3) process management. Moreover a quantitative method is proposed for making homogeneous the pairwise weights when the AHP is used for the "quality-in-use" comparison.

We tested the proposed approach on the comparison of two different implementations of a clinical chatbot over time. The results showed that improving the values of some measures in one dimension could lead to the deterioration of other quality-in-use in other dimensions. This tells us that a total quality evaluation or comparison cannot ignore the verification of quality for each single dimension.

In future research, the implications of having different weights of importance associated with characteristics of the ISO/IEC 25010 standard should also be analysed. The only characteristic weighted more than the others in terms of importance is "freedom from risk" since it is imposed by the clinical domain. Another area of research for future study is certainly the measurement of "product quality" (the other quality model of ISO/IEC 25010), wherein the proposed pairwise calculation method could give interesting results. Moreover, in Section 4, it has been pointed out that the method allows a comparison between chatbots working on the same classes of functionality. This represents a limitation of our research work.

Finally, a recent review [59] highlighted that there is a technological trend in the development of chatbots. Rule-based conversational agents (which interact through precise rules often encoded within databases) are giving way to development techniques enabled by artificial intelligence algorithms. It is interesting to note that in both cases, the stated assessment methodologies are based on three specific aspects: content

evaluation, user satisfaction, and functional aspects. This means that a quality assessment cannot disregard the intersection of these three specific areas. The method proposed in this research work is in line with this trend, as the measures defined cross characteristics encoded in ISO/IEC 25010 standard precisely for the evaluation of "quality in use" (also including user satisfaction). As we have seen, these measures are based on functional specifications, some of which are related to content evaluation. In the future, our work can be further developed to cluster these measures (possibly extending them) on both rule-based and AI chatbot technologies, applying the AHP method to these two classes based on how important the rule component is compared to the AI component. This extension could also pave the way for quality evaluation of hybrid chatbots that have both rule-based components (providing rigor to the dialogue) and AI components that contribute to the dynamic nature of the dialogue.

8. Summary Table

1. What is already known on the topic

- Chatbots are currently a valid alternative to humans in first-level interviews with users, but how to measure their quality is an open question.
- There is a set of characteristics for measuring quality in use of software systems proposed by the international standard ISO/IEC 25010.
- There is a multi-criteria decision analysis methodology, dubbed Analytic Hierarchy Process (AHP), which allows to select in a given discrete set of alternatives the best one.

2. What this study added to our knowledge

- Identification of a set of measures for each characteristic of ISO/IEC 25010 according to three classes of functionality: providing information, providing prescriptions and process management.
- Definition of a quantitative method for making homogeneous the pairwise weights when the AHP is used for the "quality-in-use" comparison.

Funding This work was funded by Italian Ministry of Education, University and Research (MIUR) through D.M. 1062/2021 - Programma Operativo (PON) Ricerca E Innovazione 2014-2020 – Azione IV.6 "Contratti di ricerca su tematiche Green" ed Azione IV.4 "Dottorati e Contratti di ricerca su tematiche dell'Innovazione" (CODICE CUP DM 25/06/2021 N.737 H95F21001470001 of University of Bari Aldo Moro).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. AHP pairwise calculated for all measures.

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	Providing INFORMATION (measures)	Measure type	Assistente Sanitario Version 1 Score granularity (n)	Assistente Sanitario Version 2 pairwise	Q_Score	Q_Score
Effectiveness		accuracy and completeness with which users achieve specified goals	text only	x	2	1	0	1
			semantic annotation	x	2	1	0	1
			figure & video	x	2	1	0	1

(continued on next page)

(continued)

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	Providing INFORMATION (measures)	Measure type	Assistente Sanitario Version 1 Score granularity (n)	Q_Score	Assistente Sanitario Version 2 pairwise	Q_Score
Efficiency		resources expended in relation to the accuracy and completeness with which users achieve goals	accurate speech synthesis		5	1	0	1
			meets neurodiverse needs	x	2	0	0	0
			real time information		5	5	0	3
			low cost/free information predilection web service information instead of physical logistic		3	1	0	1
Satisfaction	Usefulness	degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the results of use and the consequences of use	accuracy related to territory completeness consistency respecting the EBM		5	3	0	3
			personalized information	x	2	1	0	1
			certified by third-party medical (credibility)	x	2	0	0	0
			mediated by doctors linked to the sources	x	2	0	0	1
	Trust	degree to which a user or other stakeholder has confidence that a product or system will behave as intended	personalized information	x	2	1	0	1
			supported by feedback from others	x	2	0	0	0
			psychological support	x	2	0	0	0
	Pleasure	degree to which a user obtains pleasure from fulfilling their personal needs	graceful degradation	x	2	0	0	0
			effective function allocation		5	2	0	2
			grammatical fit		5	5	0	5
meaning fit				5	5	0	5	
Comfort	degree to which the user is satisfied with physical comfort	visual look		5	2	0	2	
		multichanneling	x	2	0	0	0	
		human like interaction		5	3	0	3	
		linguistic accuracy of output		5	4	0	4	
Freedom from Risk	Economic Risk Mitigation	degree to which a product or system mitigates the potential risk to financial status, efficient operation, commercial property, reputation or other resources in the intended contexts of use	multimedia interaction	x	2	0	0	0
			on demand and real time information retrieval		5	5	0	1
			information accompanied by economic and financial rights	x	2	0	0	0
			robustness to manipulation		5	5	0	5
	Health and Safety Risk Mitigation	degree to which a product or system mitigates the potential risk to people in the intended contexts of use	certified information		5	3	0	3
			care giver involvement	x	2	0	0	0
			medics involvement	x	2	0	0	0
			provide Safe Driving Mode	x	2	0	0	1
Environmental Risk Mitigation	degree to which a product or system mitigates the potential risk to property or the environment in the intended contexts of use	provide Safe Driving Mode	x	2	0	0	0	

(continued on next page)

(continued)

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	Providing INFORMATION (measures)	Measure type	Assistente Sanitario Version 1 Score granularity (n)	Q_Score	Assistente Sanitario Version 2 pairwise	Q_Score
Context Coverage	Context Completeness	degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in all the specified contexts of use	providing information	x	2	1	0	1
			linking information to other similar user feedback	x	2		0	
			providing mechanism to rank information depending on user objectives	x	2		0	
	Flexibility	degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in contexts beyond those initially specified in the requirements	patient centered language	x	2	1	0	1
			medical stakeholder language	x	2		0	
		care giver involvement	x	2		0		

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	PROVIDING PRESCRIPTION (Measures)	Measure Type	Assistente Sanitario Version 1 Score granularity (n)	Q_Score	Assistente Sanitario Version 2 pairwise	Q_Score
Effectiveness		accuracy and completeness with which users achieve specified goals	provide prescriptions	x	2	0	0	0
			provide suggestions	x	2	1	0	1
			formal sending (pdf, email, legalmail)	x	2	0	0	0
Efficiency		resources expended in relation to the accuracy and completeness with which users achieve goals	product or service suggestions	x	2	1	0	1
Satisfaction	Usefulness	degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the results of use and the consequences of use	concreteness and practicability		5	4	0	4
			lineguide response	x	2	1	0	1
			personalized prescriptions and suggestion	x	2	1	0	1
	Trust	degree to which a user or other stakeholder has confidence that a product or system will behave as intended	supplied in time	x	2	1	0	1
			certified by doctors	x	2	0	0	0
	Pleasure	degree to which a user obtains pleasure from fulfilling their personal needs	linked to scientific lineguide and EBM sources	x	2		0	
			take into account the user's inclinations or ethical choices	x	2	0	0	0
Freedom from Risk	Economic Risk Mitigation	degree to which the user is satisfied with physical comfort degree to which a product or system mitigates the potential risk to financial status, efficient operation, commercial property, reputation or other resources in the intended contexts of use	effective function allocation		5	2	0	2
			direct virtual interaction with clinical stakeholders	x	2	0	0	0
			consider whether an insurance policy has been taken out	x	2	0	0	0
			providing a price benchmark	x	2	0	0	0

(continued on next page)

(continued)

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	PROVIDING PRESCRIPTION (Measures)	Measure Type	Assistente Sanitario Version 1 Score granularity (n)	Q_Score	Assistente Sanitario Version 2 pairwise	Q_Score
	Health and Safety Risk Mitigation	degree to which a product or system mitigates the potential risk to people in the intended contexts of use	lineguide compliantness	x	2	1	0	1
			EBM compliantness	x	2	1	0	1
			validated by medics	x	2	1	0	1
			care giver involvement	x	2	0	0	0
	Environmental Risk Mitigation	degree to which a product or system mitigates the potential risk to property or the environment in the intended contexts of use	provide Safe Driving Mode	x	2	0	0	0
Context Coverage	Context Completeness	degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in all the specified contexts of use	providing prescriptions or recommendations or suggestions	x	2	1	0	1
			provide mechanisms for formulating different hypotheses (ex., diagnosis) on which to give prescriptions or recommendations	x	2		0	
	Flexibility	degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in contexts beyond those initially specified in the requirements	robustness to unexpected input		5	2	0	2

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	PROCESS MANAGEMENT (measures)	Measure Type	Assistente Sanitario Version 1 Score granularity (n)	Q_Score	Assistente Sanitario Version 2 pairwise	Q_Score
Effectiveness		accuracy and completeness with which users achieve specified goals	indirect process information grasping for better answers and process management	x	2	0	0	1
Efficiency		resources expended in relation to the accuracy and completeness with which users achieve goals	low finalized interaction for information grasping (indirect knowledge building)	x	2	1	0	1
			patient/medics interaction	x	2		0	1
Satisfaction	Usefulness	degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the results of use and the consequences of use	patient/PA interaction	x	2	1	0	3
			tasks alignment		5		0	
	time alignment			5	1	0	3	
	cost alignment			5	1	0	3	
	Trust	degree to which a user or other stakeholder has confidence that a product or system will behave as intended	real pathway state correspondance		5	1	0	3
	Pleasure	degree to which a user obtains pleasure from fulfilling their personal needs	completeness in the examination of the patient's data		5	1	0	3
			predict in advance the next tasks to be performed		5	1	0	4

(continued on next page)

(continued)

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	PROCESS MANAGEMENT (measures)	Measure Type	Assistente Sanitario Version 1	Assistente Sanitario Version 2				
Freedom from Risk	Comfort	degree to which the user is satisfied with physical comfort	connect all the stakeholders in the clinical pathway performing tasks		5	2	0	3		
			privileging solutions, open, low cost, public health based, obtaining the same outcome		5	3	0	3		
			effective function allocation		5	3	0	3		
			use of IoT for health parameter measuring	x	2		0			
			case management		5		0	3		
			Health and Safety Risk Mitigation	degree to which a product or system mitigates the potential risk to people in the intended contexts of use	avoid tasks/token error		5	1	0	4
					protect and respect privacy		5	5	0	5
					care giver involvement	x	2		0	1
					avoid inappropriate utterances and be able to perform damage control		5	4	0	4
					offline personal health datalog access	x	2		0	
ranking process state grasping		3			1	0	3			
history of execution tracking	x	2			0	0	0			
off-road detection		3			1	0	3			
patient/medics interaction	x	2			0	0	1			
provide Safe Driving Mode	x	2				0				
Context Coverage	Context Completeness	degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in all the specified contexts of use	providing integrated clinical pathway support	x	2		0	1		
			Flexibility	degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in contexts beyond those initially specified in the requirements	robustness to unclarity and enoughness information in the patient datalog		5	1	0	3
					robustness to unexpected input		5	2	0	5

Inconsistency	Effectiveness	Efficiency	Freedom Fr-	Satisfacti-
Context C-	↑ 5	↑ 5	↑ 9.0000	← 5
Effectiveness		← 5	↑ 9.0000	↑ 5
Efficiency			↑ 9.0000	← 1
Freedom Fr-				← 9

ISO/IEC 25010 "Quality in use" Criteria

Inconsistency	Pleasure ~	Trust ~	Usefulness~
Comfort ~	← 0	← 0	← 0
Pleasure ~		← 0	← 0
Trust ~			← 0

Satisfaction-SUB Criteria

Inconsistency	Environmen~	Health and~
Economic R~	← 7	↑ 9.0000C
Environmen~		↑ 9.0000C

Freedom From Risk-SUB Criteria

Inconsistency	Flexibilit~
Context C~	← 3

Context Coverage-SUB Criteria

Appendix B. Analysed clinical chatbots

Chatbot name	Channel	Main functions	Reference
SafedrugBot	Telegram	helps doctor access right information about drug dosage and guide patience	[60]
Florence chatbot	Messenger Skype Kik	reminds patients to take pills tracks body weight tracks moods finds a doctor or pharmacy nearby provides information on any medical issue	[61]
Izzy	Messenger	helps women track their period provides information on users' sexual issues and menstrual health reminds them when to take birth control pills	[62]
Forksy	Messenger	assists in tracking calories promotes healthy eating habits food diary	[63]
Babylon Health	mobile App	remote consultation with health care professionals and doctors patient's medical history database symptom checker	[64]
Buoy Health	website	assist patients in diagnosing	[65] [66]
CancerChatbot	Messenger	offers detailed information on cancer and related topics	[67]
Sensely	mobile App	tracks health symptoms using both text and speech communication diagnosis formulation tries to understand the level of emergency	[68]
GYANT	Messenger Alexa	symptom checker	[69]
Woebot	mobile App	studies patient mood, personality and suggests remedies as a therapist for depression	[70] [71]
HealthTap	Messenger	physician-patients communication channel via bot make its vast repository of knowledge available to patients using the app	[72]
Your.Md	Messenger Slack KIK Telegram	symptom checker	[73]
Ada Health	mobile App Alexa	symptom checker	[74]

(continued on next page)

(continued)

Chatbot name	Channel	Main functions	Reference
Symtomate	mobile App	symptom checker	[75]
Bots4Health	mobile App	sexual and reproductive health chat about a wide range of health issues	[76]
Assistente Sanitario	Telegram	Clinical data repository Provides semantically annotated information	[36]

Appendix C. ISO/IEC 25022 and Table 1 measures example matching

	ISO/IEC 25022 measures	clinical chatbot quality measure combination
Effectiveness	Tasks completed	Formal sending; Indirect process information grasping for better answers and process management
	Objectives achieved	Provide prescriptions; Provide suggestion; Formal sending
	Errors in a task	indirect process information grasping for better answers and process management (on task errors)
	Tasks with errors	Indirect process information grasping for better answers and process management (on task errors)
	Task error intensity	Indirect process information grasping for better answers and process management (on task errors intensity)
Efficiency	Task time	Real time information; Web service information instead of physical logistic
	Time efficiency	Real time information, Web service information instead of physical logistics, Task alignment
	Cost-effectiveness	Patient/medics interaction; Patient/PA interaction
	Productive time ratio	Real time information; Low finalized interaction for information grasping (indirect knowledge building)
	Unnecessary actions	Low finalized interaction for information grasping (indirect knowledge building); Low cost/free information predilection
Satisfaction	Fatigue	Web service information instead of physical logistic
	Overall satisfaction	Completeness; Personalized information; Visual look; Gramatical fit; Meaning fit
	Satisfaction with features	Effective function allocation
	Discretionary usage	Connect all the stakeholders in the clinical pathway; Personalized information
	Feature utilisation	Effective function allocation; Use of IoT for health parameter measuring; multichanneling
	Proportion of users complaining	Linguistic accuracy of output; Gramatical fit; Meaning fit
	Proportion of user complaints about a particular feature	Effective function allocation; Linguistic accuracy of output; Gramatical fit; Meaning fit
Economic risk	User trust	Certified by third-parties medical (credibility); Certified by doctors
	User pleasure	Graceful degradation; Visual look; Effective function allocation; Predict in advance the next tasks to be performed
	Physical comfort	Human like interaction; Multimedia interaction; Direct virtual interaction with clinical stakeholder
	Return on investment (ROI)	Providing price benchmark; Case management
	Time to achieve return on investment	Information accompanied by economic and financial rights; Providing price benchmark
	Business performance	Providing price benchmark; Avoid tasks/token error; Performing tasks privileging solutions, open, low cost, public health based, obtaining the same outcome
	Benefits of IT Investment	Providing price benchmark; Connect all the stakeholders in the clinical pathway; Multichanneling; Direct virtual interaction with clinical stakeholder
Health and safety risk	Service to customers	Case management; Care giver involvement; medics involvement
	Website visitors converted to customers	Case management; Care giver involvement; Medics involvement; certified information
	Revenue from each customer	Providing price benchmark; Case management; Care giver involvement; Medics involvement; Information accompanied by economic and financial rights
	Errors with economic consequences	Avoid tasks/token error; Avoid inappropriate utterances and be able to perform damage control; Medics involvement; Robustness to manipulation
	User health reporting frequency	Use of IoT for health parameter measuring; history of execution tracking
	User health and safety impact	Avoid tasks/token error; Avoid inappropriate utterances and be able to perform damage control; Medics involvement; Robustness to manipulation; Care giver involvement; Provide Safe Driving Mode
	Safety of people affected by use of the system	Validated by medics; Avoid inappropriate utterances and be able to perform damage control; Care giver involvement; Medics involvement
Environmental risk Context completeness	Environmental impact	Provide Safe Driving Mode
	Context completeness	Providing mechanism to rank information depending user objectives; Provide mechanisms for formulating different hypotheses (ex. diagnosis) on which to give prescriptions or recommendations
Flexibility	Flexible context of use	Robustness to anespected input; Patient centered language
	Product flexibility	Robustness to anespected input
	Proficiency independence	Providing information, providing prescriptions or recommendations or suggestions; Providing integrated clinical pathway support

References

[1] S. Pérez-Soler, S. Juárez-Puerta, E. Guerra, J. de Lara, Choosing a chatbot development tool, *IEEE Softw.* 38 (4) (2021) 94–103.

[2] J. Pereira, Ó. Díaz, Chatbot dimensions that matter: Lessons from the trenches, in: *International Conference on Web Engineering*, Springer, 2018, pp. 129–135.

[3] T. Makasi, A. Nili, K.C. Desouza, M. Tate, A typology of chatbots in public service delivery, *IEEE Softw.* 39 (3) (2021) 58–66.

[4] L. Müller, J. Mattke, C. Maier, T. Weitzel, Conversational agents in healthcare: using qca to explain patients' resistance to chatbots for medication, *Int. Workshop Chatbot Res. Des.*, Springer (2019) 3–18.

[5] J.-E. Bibault, B. Chaix, P. Nectoux, A. Pienkowski, A. Guillemasé, B. Brouard, Healthcare ex machina: Are conversational agents ready for prime time in oncology? *Clin. Transl. Radiat. oncology* 16 (2019) 55–59.

[6] N. Bhirud, S. Tataale, S. Randive, S. Nahar, A literature review on chatbots in healthcare domain, *International journal of scientific & technology research* 8 (7) (2019) 225–231.

- [7] I.O. for Standardization/International Electrotechnical Commission, et al., Iso/iec 25010 - systems and software engineering — systems and software quality requirements and evaluation (square) — systems and software engineering, ISO/IEC.
- [8] M. Yan, X. Xia, X. Zhang, L. Xu, D. Yang, S. Li, Software quality assessment model: A systematic mapping study, *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, 2018, pp. 1–18.
- [9] A.P.S. Alves, D.O.G. de Alencar, A.M. Gonalo Filho, S.C. Paiva, D.B.F. Carvalho, Development and evaluation of a chatbot for the regional museum of sao joao del-rei. 2018 XLIV Latin American Computer Conference (CLEI), IEEE, 2018, pp. 388–397.
- [10] S.K. Yuwono, B. Wu, L.F. D’Haro, Automated scoring of chatbot responses in conversational dialogue, in: 9th International Workshop on Spoken Dialogue System Technology, Springer, 2019, pp. 357–369.
- [11] M. Shmueli-Scheuer, T. Sandbank, D. Konopnicki, O.P. Nakash, Exploring the universe of egregious conversations in chatbots, in: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, 2018, pp. 1–2.
- [12] B. AbuShawar, E. Atwell, Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems, *Int. J. Speech Technol.* 19 (2) (2016) 373–383.
- [13] K. Kuligowska, Commercial chatbot: performance evaluation, usability metrics and quality standards of embodied conversational agents, *Professionals Center for Business Research* 2.
- [14] R.A. Lionberger, S.L. Lee, L. Lee, A. Raw, L.X. Yu, Quality by design: Concepts for andas, *The AAPS Journal* 10 (2) (2008) 268–276.
- [15] I.O. for Standardization, S. Technical Committee ISO/IEC JTC 1, Information technology. Subcommittee SC 7, systems engineering, Systems and Software Engineering: Systems and Software Quality Requirements and Evaluation (SQuARE): System and Software Quality Models, ISO, 2011.
- [16] I.O. for Standardization/International Electrotechnical Commission, et al., Iso/iec 25012: Software engineering-software product quality requirements and evaluation (square)-data quality model, ISO/IEC.
- [17] I.O. for Standardization/International Electrotechnical Commission, et al., Iso/iec 25024 - systems and software engineering — systems and software quality requirements and evaluation (square) — measurement of data quality, ISO/IEC.
- [18] J. Estdale, E. Georgiadou, Applying the iso/iec 25010 quality models to software product, *European Conference on Software Process Improvement*, Springer (2018) 492–503.
- [19] F. Agustin, H. Kurniawan, Y. Yusfrizal, K. Umami, Comparative analysis of application quality between appserv and xampp webserver using ahp based on iso/iec 25010: 2011, in: 2018 6th International Conference on Cyber and IT Service Management (CITSM), IEEE, 2018, pp. 1–5.
- [20] A. Idri, L. Sardi, J.L.F. Alemán, Quality evaluation of gamified blood donation apps using iso/iec 25010 standard, *HEALTHINF* (2018) 607–614.
- [21] M. Falco, G. Robiolo, Building a catalogue of iso/iec 25010 quality measures applied in an industrial context, in: *Journal of Physics: Conference Series*, Vol. 1828, IOP Publishing, 2021, p. 012077.
- [22] A. Yulianty, A. Kurniawati, Quality analysis of bios portal website at banking companies using iso/iec 25010: 2011 method, *Int. Res. J. Adv. Eng. Sci* 6 (2) (2021) 11–16.
- [23] R.C. Wibawa, S. Rochimah, R. Anggoro, A development of quality model for online games based on iso/iec 25010, in: 2019 12th International Conference on Information & Communication Technology and System (ICTS), IEEE, 2019, pp. 215–218.
- [24] L. Liuliyah, A.P. Subriadi, Performance measurement of academic information systems using performance prism and iso/iec 25010, *The Winners* 21 (2) (2020) 75–83.
- [25] M. Islam, R. Imran, S. Hosain, The evaluation of enterprise resource planning using iso 25010 based quality model. 2021 2nd International Informatics and Software Engineering Conference (IISEC), IEEE, 2021, pp. 1–6.
- [26] T. Saaty, Modeling unstructured decision problems: A theory of analytical hierarchy, in: *Proceedings of the First International Conference on Mathematical Modeling*, Vol. 1, 1997, pp. 59–77.
- [27] R. de FSM Russo, R. Camanho, Criteria in ahp: a systematic review of literature, *Procedia Computer Science* 55 (2015) 1123–1132.
- [28] W. Ho, Integrated analytic hierarchy process and its applications—a literature review, *European Journal of Operational Research* 186 (1) (2008) 211–228.
- [29] S. Mahmudova, Z. Jabrailova, Development of an algorithm using the ahp method for selecting software according to its functionality, *Soft. Comput.* 24 (11) (2020) 8495–8502.
- [30] M. Yazdi, T. Saner, M. Darvishmotevali, Application of an artificial intelligence decision-making method for the selection of maintenance strategy, in: *International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions*, Springer, 2019, pp. 246–253.
- [31] T.L. Saaty, L.G. Vargas, The analytic network process, in: *Decision making with the analytic network process*, Springer, 2013, pp. 1–40.
- [32] Y. elikbilek, F. Tüsiyüz, An in-depth review of theory of the topsis method: An experimental analysis, *Journal of Management Analytics* 7 (2) (2020) 281–300.
- [33] A.S. Jadhav, R.M. Sonar, Evaluating and selecting software packages: A review, *Information and software technology* 51 (3) (2009) 555–563.
- [34] S. Laumer, C. Maier, F. Gubler, Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis, in: *Proceedings of the 27th European Conference on Information Systems (ECIS)*, AISel, 2019.
- [35] J.J. Sophia, D.A. Kumar, M. Arutselvan, S.B. Ram, A survey on chatbot implementation in health care using nltk, *Int. J. Comput. Sci. Mob. Comput* 9.
- [36] C. Ardito, D. Caivano, L. Colizzi, G. Dimauro, L. Verardi, Design and execution of integrated clinical pathway: A simplified meta-model and associated methodology, *Information* 11 (7) (2020) 362.
- [37] L. Kinsman, T. Rotter, E. James, P. Snow, J. Willis, What is a clinical pathway? development of a definition to inform the debate, *BMC medicine* 8 (1) (2010) 1–3.
- [38] S. Srivastava, T.V. Prabhakar, Hospitality of chatbot building platforms, in: *Proceedings of the 2nd ACM SIGSOFT International Workshop on Software Qualities and Their Dependencies*, 2019, pp. 12–19.
- [39] M. Edirisooriya, I. Mahakalanda, T. Yapa, Generalised framework for automated conversational agent design via qfd. 2019 Moratuwa Engineering Research Conference (MERCon), IEEE, 2019, pp. 297–302.
- [40] N. Radziwill, M. Benton, Evaluating quality of chatbots and intelligent conversational agents, *Software Quality Professional* 19 (3) (2017) 25.
- [41] A. Atiyah, S. Jusoh, F. Alghanim, Evaluation of the naturalness of chatbot applications, in: *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, IEEE, 2019, pp. 359–365.
- [42] M. Tschanz, T.L. Dorner, J. Holm, K. Denecke, Using emma to manage medication, *Computer* 51 (8) (2018) 18–25.
- [43] G.I. Hess, G. Fricker, K. Denecke, Improving and evaluating emma’s communication skills: a chatbot for managing medication, *Stud Health Technol Inform* 259 (2019) 101–104.
- [44] E. Ruane, T. Faure, R. Smith, D. Bean, J. Carson-Berndsen, A. Ventresque, Botest: a framework to test the quality of conversational agents using divergent input examples, in: *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, 2018, pp. 1–2.
- [45] W. Liu, J. Zhang, S. Feng, An ergonomics evaluation to chatbot equipped with knowledge-rich mind. 2015 3rd International Symposium on Computational and Business Intelligence (ISCBBI), IEEE, 2015, pp. 95–99.
- [46] A. Abd-Alrazaq, Z. Safi, M. Alajlani, J. Warren, M. Househ, K. Denecke, et al., Technical metrics used to evaluate health care chatbots: scoping review, *Journal of medical Internet research* 22 (6) (2020) e18301.
- [47] J. Pereira, . Dıaz, Using health chatbots for behavior change: a mapping study, *Journal of medical systems* 43 (5) (2019) 1–13.
- [48] K. Denecke, S.L. Hochreutener, A. Pöpel, R. May, Self-anamnesis with a conversational user interface: concept and usability study, *Methods of information in medicine* 57 (05/06) (2018) 243–252.
- [49] T. Kowatsch, D. Volland, I. Shih, D. Rieger, F. Künzler, F. Barata, A. Filler, D. Büchter, B. Brogle, K. Heldt, et al., Design and evaluation of a mobile chat app for the open source behavioral health intervention platform mobilecoach, in: *International Conference on Design Science Research in Information System and Technology*, Springer, 2017, pp. 485–489.
- [50] D. Coniam, The linguistic accuracy of chatbots: usability from an esl perspective, *Text & Talk* 34 (5) (2014) 545–567.
- [51] A. Dhoubi, A. Trabelsi, C. Kolski, M. Neji, An approach for the selection of evaluation methods for interactive adaptive systems using analytic hierarchy process, in: 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS), IEEE, 2016, pp. 1–10.
- [52] L. eponienė, V. Drungilas, M. Jurgelaitis, J. eponis, Method for reverse engineering uml use case model for websites, *Information Technology and Control* 47 (4) (2018) 623–638.
- [53] I.O. for Standardization/International Electrotechnical Commission, et al., Iso/iec 25022 - systems and software engineering — systems and software quality requirements and evaluation (square) — systems and software engineering, ISO/IEC.
- [54] P.-N. Tan, M. Steinbach, V. Kumar, *Introduction to data mining*, Pearson Education India, 2016.
- [55] K.D. Goepel, Comparison of judgment scales of the analytical hierarchy process—a new approach, *International Journal of Information Technology & Decision Making* 18 (02) (2019) 445–463.
- [56] T. De Jonge, R. Veenhoven, L. Arends, Homogenizing responses to different survey questions on the same topic: Proposal of a scale homogenization method using a reference distribution, *Soc. Indic. Res.* 117 (1) (2014) 275–300.
- [57] Tagme api, <https://sobigdata.d4science.org/web/tagme/tagme-help>, accessed: 21-10-2020.
- [58] Super decisions cdf, <https://www.superdecisions.com>, accessed: 08-10-2020.
- [59] W. Maroengsit, T. Piyakulpinoy, K. Phonyiam, S. Pongnumkul, P. Chaovalit, T. Theeramunkong, A survey on evaluation methods for chatbots, in: *Proceedings of the 2019 7th International conference on information and education technology*, 2019, pp. 111–119.
- [60] A virtual assistant to help doctors in their daily work, <https://www.safeinbreastfeeding.com/safedrugbot-chatbot-medical-assistant/>, accessed: 15-01-2022.
- [61] Florence your health assistant, <https://www.florence.chat/>, accessed: 14-01-2022.
- [62] superizy, <https://www.facebook.com/superizy/about/>, accessed: 14-01-2022.
- [63] Automate nutrition coaching, <https://getforks.com>, accessed: 14-01-2022.
- [64] babylon, <https://www.babylonhealth.com>, accessed: 14-01-2022.
- [65] D. Chittamuru, S. Ramondt, R. Kravitz, S. Ramirez, Who uses an online intelligent medical information system and what do they do with that information? results from a pilot study of users of buoy health, in: *APHA’s 2019 Annual Meeting and Expo* (Nov. 2-Nov. 6), APHA, 2019.
- [66] buoy health, <https://www.buoyhealth.com>, accessed: 14-01-2022.
- [67] Cancer chatbot, <https://www.facebook.com/CancerChatbot>, accessed: 14-01-2022.
- [68] Sensely chatbot, <http://www.sensely.com>, accessed: 14-01-2022.
- [69] Gyant, <https://gyant.com>, accessed: 14-01-2022.
- [70] Woebot health, <https://woeobot.io>, accessed: 14-01-2022.

- [71] K.K. Fitzpatrick, A. Darcy, M. Vierhile, Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial, *JMIR Mental Health* 4 (2) (2017) e7785.
- [72] healthtap, <https://www.healthtap.com>, accessed: 14-01-2022.
- [73] healthily, <https://www.your.md>, accessed: 14-01-2022.
- [74] Health powered by ada, <https://ada.com>, accessed: 14-01-2022.
- [75] Symptomate, <https://symptomate.com/>, accessed: 31-07-2022.
- [76] bot4health, <https://www.facebook.com/bots4health/>, accessed: 31-07-2022.