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

Vita Santa Barletta[1] · Danilo Caivano[1] ·Lucio Colizzi[1] · Giovanni Dimauro[1] · Mario Piattini[2]

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 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. For this purpose, we developed a set of quality measures for the three clinical chatbot dimensions: Providing information, Providing prescriptions, Process management. Each dimension was composed of the characteristics of ISO/IEC 25010 “quality in use”, i.e., effectiveness, efficiency, satisfaction, freedom from risk, and context coverage.

Findings. As a case study, a comparison of two versions of chatbots was performed with an analytic hierarchy process (AHP) method. 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.

Conclusions. 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. The proposed method contributes to providing a reference base for performing a quality comparison of clinical chatbots compliant with the ISO/IEC 25010 standard.

Keywords

medical-chatbot quality, clinical pathway, AHP, ISO/IEC 25010

✉ Lucio Colizzi
lucio.colizzi@uniba.it

[1] University of Bari Aldo Moro, Computer Science Department, Via Edoardo Orabona, 4, 70125 Bari – Italy, name.surname@uniba.it

[2] University of Castilla-La Mancha, Alarcos Research Group, Information Systems and Technologies Institute, Ciudad Real, Spain.
mario.piattini@uclm.es

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 role-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. 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 [4], chatbots in the health care domain sector 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. Personalization of services, case management, user-centric dialogue, active and tireless guidance 24 hours 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 (but not the only ones). To lay the foundations for the diffusion of clinical chatbots, it is necessary to control all aspects related to health safety and user responsibility. This objective can only be achieved by applying the "quality comparison" method [5].

The ISO/IEC 25010 standard [6] defines two models for quality measurement: "quality in use" and "product quality". The former 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 and the related measures reported in ISO/IEC 25024 [7] [8].

In this paper, we propose a method for evaluating the "quality in use" of clinical chatbots according to ISO/IEC 25010. The proposed method is based on the analytic hierarchy process (AHP) [9]. 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. Chatbots quality in clinical domain

In this section, the most recent approaches for measuring the quality of clinical chatbots are presented.

In [10], a hospitality index is defined for each specific quality attribute as an indicator of how effective the platform is in achieving that attribute.

In [11], the quality assessment of chatbots is addressed through the integration of AHP and quality function deployment (QFD) methods. In [12], 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 [13], 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 [14] [15]to evaluate the usability of the eMMA chatbot. The evaluation criterion is qualitative, and the characteristics are all considered as a whole; however, in [16], 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. Some papers partially addressed the quality of the chatbot by measuring the quality of the dialogue [17] [18] [19]. In [20], 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 [21]. 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 [22] 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 [23], 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 [24] 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 [25] a chatbot solution based on predefined answer sets is proposed. The “quality in use” was also measured in this case through a 7-point Linkert scale questionnaire. Based on the questionnaire feedback, different types of statistical methods were used for the quality assessment.

Analysing the functionalities made available by the most famous chatbots operating in the clinical domain, the main field of application is diagnosis [26]. 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 [27]. In this case, the bot becomes an intelligent diary ready to remind the patient to take his 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 [28]. 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 [29]. 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.

Some studies have helped to understand what to measure when assessing dialogue quality. In [30] 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. In [31], 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.

From what we have amply reported above, it emerges that the need to measure the quality of chatbot technologies is an open question [22]. 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 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.

As stated in ISO/IEC 25010, **"**quality in use 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." Analysing the characteristics of chatbots in Appendix B, it emerged that the most recurrent functionalities relate to user interactions. In particular, the three clinical chatbot dimensions of "quality in use" are the following:

- **Providing information**. This is perhaps the most widespread interaction in medical chatbots. The user wants to deepen his 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.

It is understandable that this dimension could contain complexities at both the technological and organizational levels. Think, for example, of all those functions that have the objective of making agile the collaboration between different actors of clinical processes (general practitioners, nurses, case managers, medical specialists, hospitals, rehabilitation facilities, public administration).

Moreover, the intersection of these three types of interactions 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 precise prescriptions.

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

The proposed method is based on the procedure defined by [12]. 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. When comparing different chatbots, it is important that they operate in the same dimensions described in section 3 , which means they share one or more classes of functionality. Moreover, as will be shown, when incremental versions of the same chatbot are compared, improvement of quality is not always guaranteed.

To obtain a more effective method in the case of clinical chatbots, it has been proposed to use a multicriteria procedure wherein the alternatives coincide with the characteristics identified by ISO/IEC 25010 for "quality in use" assessment, while the criteria are the abovementioned: *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 user-chatbot macrointeractions identified (*Providing information, Providing prescriptions, Process management*). Given a specific quality characteristic and a chatbot interaction type, 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 [30], [31] [12];

3) measures derived from other ISO standards and applicable to this context ([7] [8]. for data quality.

The ISO/IEC 25022standard [32] 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". Moreover, some characteristics of ISO/IEC 25012 [7] (and consequently, the relative measures in ISO/IEC 25024 [8]) have been used as measures transforming them into functionality, whose presence or absence represents lower or higher "quality in use".

Table 1 Clinical chatbot "quality in use" proposed measures

			Measures for three-chatbot dimensions		
ISO25010 "Quality in use" characteristics	**ISO25010 "Quality in use" subcharacteristics**	**Characteristic description**	**Providing information**	**Providing prescriptions**	**Process management**
Effectiveness		*accuracy and completeness with which users achieve specified goals*	- text only - semantic annotation - figure & video - accurate speech synthesis - meets neurodiverse needs	- provide prescription - provide suggestion - formal sending (pdf, email, legalmail)	- indirect process information grasping for better answers and process management
Efficiency		*resources expended in relation to the accuracy and completeness with which users achieve goals*	- real time information - low cost/free information predilection - web service information instead of phisical logistic	- product/service suggestion	- low finalized interaction for information grasping (indirect knowledge building) - patient/medics interaction - patient/PA interaction
Satisfaction	***Usefulness***	*degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the resultsof use and the consequences of use*	- accuracy related to lineguide - accuracy related to territory - completness - consistency respect the EBM - personalized information	- concreteness and practicability - lineguide respondance - personalized prescriptions/suggestions - supplied in time	- tasks alignment - times alignment - costs alignment
	Trust	*degree to which a user or other stakeholder has confidence that a product or system will behave as intended*	- certified by third medical-parties (credibility) - mediated by doctors - linked to the sources	- certified by doctors - linked to scientific, lineguide, EBM sources	- real pathway state corrispondance - completeness in the examination of the patient's datalog
	Pleasure	*degree to which a user obtains pleasure from fulfilling their personal needs*	- Personalized information - supported by feedback from others - psychological support - Gracefgul degradation - Effective function allocation -gramatical fit - meaning fit - visual look	- effective function allocation	- 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
	Comfort	*degree to which the user is satisfied with physical comfort*	- multichanneling - human like interaction - Linguistic accuracy of output - multimedia interaction - on demand and real time information retreaving	- direct virtual interaction with clinical stakeholder	- 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*	- information accompanied by economic and financial rights	- consider whether an insurance policy has been taken out - providing price benchmark	- case management
	Health and Safety Risk Mitigation	*degree to which a product or system mitigates the potential risk to people in the intended contexts of use*	- Robustness to manipolation - certified information - care giver involvement - medics involvement - provide mechanisms to avoid interaction during travel	- lineguide compliant - EBM compliant - validated by medics - care giver involvment	- avoid tasks/token error - protecd and respect privacy - care giver involvement - avoid inappropriate utterrances and be able to perform damage control - off line personal health datalog access - ranking process state grasping - history of execution tracking
	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 mechanisms to avoid interaction during travel	- provide mechanisms to avoid interaction during travel	- provide mechanisms to avoid interaction during travel
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 - linking information to other similar user feedback - providing mechanism to ranking information depending user objectives	- 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	*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 - medical stakeholder language - care giver involvment	- robustness to anespected input	- robusteness to unclearness and enoughness infomration in the datalog patient - robustness to anespected input

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..., n} expresses an ordinal categorical measure [scarce, insufficient, sufficient, discrete, good]. We refer to the degree of [Digitare qui]

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..9.

To address this scoring issue, and according to the procedure defined in [12], 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 by applying two levels of criteria. 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. 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) [33]. 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, pairwise): P1 = (0,1); P2= (n-1,9). We assume that any P = (d, 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:

$$\text{pairwise} = 8 * \text{d} + 1 \qquad (1)$$

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 2.

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								

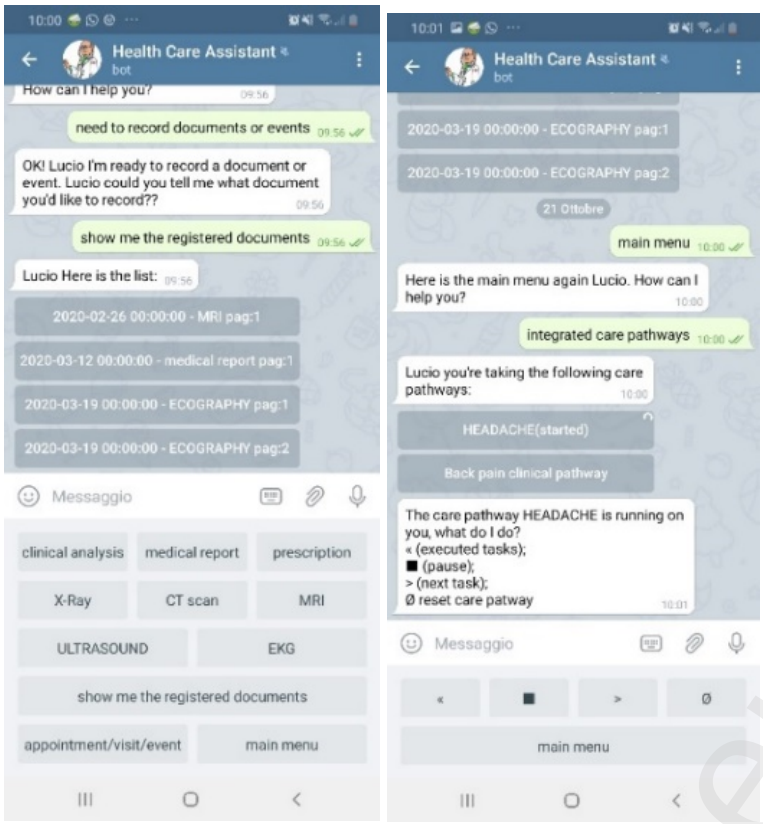
The proposed method is general and includes the possibility of using a perception-based approach in pairwise definition. The generalization lies precisely in the fact that the granularity n of a measure is not bound to a predefined value but can vary according to the particular measure defined for a 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 [34].

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* [35].

5. Example of chatbot “quality in use” comparison

Assistente Sanitario (“Health care Assistant”) [28]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 his 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 [36].



(a) *(b)*

Fig. 1 Assistente Sanitario GUI. V. 1 (a); V. 2 (b)

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*.

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 ≤ Q_Score ≤ 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.

Table 3 Pairwise value applied to each measure defined for *effectiveness* characteristic

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 granularity (n)	*Q_Score*	pairwise	*Q_Score*
Effectiveness	*accuracy and completeness with which users achieve specified goals*	provide prescription	2	0	1	0
		provide suggestion	2	1	1	1
		formal 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 granularity (n)	*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

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 [37] implementing the AHP method was 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*.

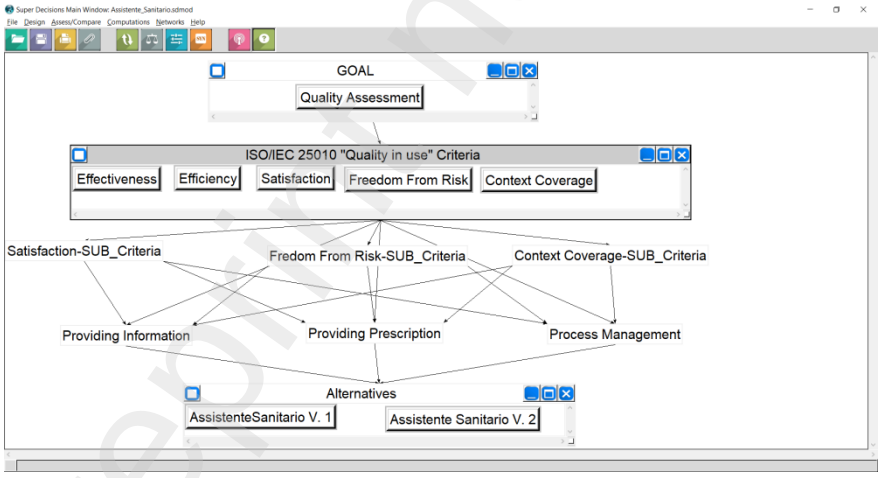


***Fig.** 2 Superdecision software:* ISO/IEC 25010 "Quality in use" (goal)

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.

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
	AssistenteSanitario_V. 1	0.1063	0.1680	0.2019	2

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

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
	AssistenteSanitario_V. 1	0.1402	0.5280	1.0000	1

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.

6. Conclusions and future work

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 devise three different quality criteria referred to as clinical chatbot dimensions: 1) *Providing information*, 2) *Providing prescriptions* and 3) *Process management*. Second, for each of these areas, we proposed a set of measures for quality-in-use evaluation according to ISO 25010 characteristics. Then, we propose a quantitative method for pairwise homogenization when AHP analysis is used for quality 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.

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. 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 standardization of the set of proposed measures represents the first issue that should be addressed in future research work. 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.

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.

Finally, a recent review [38] 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.

Declarations

Funding and/or Conflicts of interests/Competing interests 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).
The authors declare that they have no conflicts of interests and Competing interests.

References

[1] S. Pérez-Soler, S. Juárez-Puerta, E. Guerra and J. de Lara, "Choosing a chatbot development tool," *IEEE Software,* 2021.

[2] J. Pereira and Ó. Díaz, "Chatbot Dimensions that Matter: Lessons from the Trenches," in *Mikkonen T., Klamma R., Hernández J. (eds) Web Engineering. ICWE 2018. Lecture Notes in Computer Science, vol 10845, (pp. 129 -135), Springer, Cham*, 2018.

[3] T. Makasi, A. Nili, K. C. Desouza and M. Tate, "A Typology of Chatbots in Public Service Delivery," *IEEE Software,* 2021.

[4] J. E. Bibault, B. Chaix, P. Nectoux, A. Pienkowski, A. Guillemasé and B. Brouard, "Healthcare ex Machina: Are conversational agents ready for prime time in oncology?," *Clinical and translational radiation oncology,* no. 16, pp. (pp. 55-59), 2019.

[5] M. Yan, X. Xia, X. Zhang, L. Xu, D. Yang and S. Li, "Software quality assessment model: a systematic mapping study," in *Science China Information Sciences*, 2019.

[6] ISO, "25010, ISO/IEC. Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models: International Organization for Standarization," 2011. [Online]. Available: https://www.iso.org/standard/35733.html. [Accessed 14 01 2022].

[7] ISO, "ISO/IEC 25012. Software engineering. Software product quality requirements and evaluation (SQuaRE). Data quality model Ginebra: International Organization for Standarization.," 2008. [Online]. Available: https://www.iso.org/standard/35736.html. [Accessed 14 01 2022].

[8] ISO, "ISO/IEC 25024 - Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality," 2015. [Online]. Available: https://www.iso.org/standard/35749.html. [Accessed 14 01 2022].

[9] T. L. Saaty, "Modeling unstructured decision problems: a theory of analytical hierarchies," in *Proceedings of the first international conference on mathematical modeling*, 1980.

[10] S. Srivastava and T. V. Prabhakar, "Hospitality of chatbot building platforms," in *Proceedings of the 2nd ACM SIGSOFT International Workshop on Software Qualities and Their Dependencies*, 2019.

[11] M. Edirisooriya, I. Mahakalanda and T. Yapa, "Generalised Framework for Automated Conversational Agent Design via QFD," in *Moratuwa Engineering Research Conference (MERCon)*, Moratuwa, Sri Lanka, 2019.

[12] N. M. Radziwill and M. C. Benton, "Evaluating Quality of Chatbots and Intelligent Conversational Agents.," *Software Quality Professional,* vol. 19(3), 2017.

[13] A. Atiyah, S. Jusoh and F. Alghanim, "Evaluation of the Naturalness of Chatbot Applications," in *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 2019.

[14] M. Tschanz, T. L. Dorner, J. Holm and K. Denecke, “Using eMMA to manage medication,” *Computer,* no. 51(8), pp. (pp. 18-25), 2018.

[15] G. I. Hess, G. Fricker and K. Denecke, “Improving and evaluating eMMA's communication skills: a chatbot for managing medication,” *Stud Heal Technol Inf,* no. 259, pp. (pp. 101-4), 2019.

[16] A. P. S. Alves, D. O. G. de Alencar, A. M. Gonçalo Filho, S. C. Paiva and D. B. F. Carvalho, “Development and evaluation of a chatbot for the Regional Museum of Sao Joao del-Rei.,” in *XLIV Latin American Computer Conference (CLEI)*, 2018.

[17] S. K. Yuwono, B. Wu and L. F. D’Haro, “Automated scoring of chatbot responses in conversational dialogue,” in *9th International Workshop on Spoken Dialogue System Technology*, Singapore, 2019.

[18] M. Shmueli-Scheuer, T. Sandbank, D. Konopnicki and O. P. Nakash, “Exploring the universe of egregious conversations in chatbots,” in *23rd International Conference on Intelligent User Interfaces Companion*, 2018.

[19] B. AbuShawar and E. Atwell, “Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems,” *International Journal of Speech Technology,* no. 19(2), pp. (pp. 373-383), 2016.

[20] E. Ruane, T. Faure, R. Smith, D. Bean, J. Carson-Berndsen and 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.

[21] W. Liu, J. Zhang and S. Feng, “An Ergonomics Evaluation to Chatbot Equipped with Knowledge-Rich Mind,” in *3rd International Symposium on Computational and Business Intelligence (ISCBI)*, 2015.

[22] A. Abd-Alrazaq, Z. Safi, M. Alajlani, J. Warren, M. Househ and K. Denecke, “Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review,” *Journal of Medical Internet Research,* no. 22(6), e18301, 2020.

[23] J. Pereira and Ó. Díaz, “Using health chatbots for behavior change: a mapping study,” *Journal of medical systems,* vol. 43(5), p. 135, 2019.

[24] K. Denecke, S. L. Hochreutener, A. Pöpel and R. May, “Self-anamnesis with a conversational user interface: concept and usability study.,” *Methods of information in medicine,* vol. 57(05/06), pp. (pp. 243-252), 2018.

[25] T. Kowatsch, D. Volland, I. Shih, D. Rüegger, F. Künzler, F. Barata and P. Gindrat, “Design and evaluation of a mobile chat app for the open source behavioral health intervention platform MobileCoach,” in *nternational Conference on Design Science Research in Information System and Technology*, 2017.

[26] S. Laumer, C. Maier and F. T. Gubler, “Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis,” in *27th European Conference on Information Systems: Information Systems for a Sharing Society, ECIS 2019*, Stockholm and Uppsala; Sweden, 2019.

[27] J. J. Sophia, D. A. Kumar, M. Arutselvan and S. B. A. Ram, “Survey on Chatbot Implementation in Health Care using NLTK.,” *International Journal of Computer Science and Mobile Computing,* vol. 9(3), pp. (pp. 206-210), 2020.

[28] C. Ardito, D. Caivano, L. Colizzi, G. Dimauro and L. Verardi, “Design and Execution of Integrated Clinical Pathway: A Simplified Meta-Model and Associated Methodology,” *Information,* no. 11, issue 7, 2020.

[29] L. Kinsman, T. Rotter, E. James, P. Snow and J. Willis, “What is a clinical pathway? Development of a definition to inform the debate,” *BMC medicine,* no. 8(1), pp. (pp. 1-3), 2010.

[30] D. Coniam, “The linguistic accuracy of chatbots: usability from an ESL perspective.,” *Text & Talk,* vol. 34(5), pp. 545-567, 2014.

[31] K. Kuligowska, “Commercial chatbot: performance evaluation, usability metrics and quality standards of embodied conversational agents,” *Professionals Center for Business Research,* no. 2, 2015.

[32] ISO, “ISO/IEC 25022 - Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Systems and software engineering,” 2016. [Online]. Available: https://www.iso.org/standard/35746.html. [Accessed 14 01 2022].

[33] P. N. Tan, M. Steinbach and V. Kumar, Introduction to data mining, Pearson Education India, 2016.

[34] K. D. Goepel, “Comparison of judgment scales of the analytical hierarchy process-A new approach.,” *International Journal of Information Technology & Decision Making,* vol. 18(02), pp. 445-463., 2019.

[35] T. De Jonge, R. Veenhoven and L. Arends, “Homogenizing responses to different survey questions on the same topic: Proposal of a scale homogenization method using a reference distribution.,” *Social Indicators Research,* vol. 117(1), pp. 275-300, 2014.

[36] “TAGme API,” 2020. [Online]. Available: https://sobigdata.d4science.org/web/tagme/tagme-help. [Accessed 21 10 2020].

[37] “Super Decisions CDF,” 2020. [Online]. Available: https://www.superdecisions.com/. [Accessed 08 10 2020].

[38] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit and T. Theeramunkong, “A survey on evaluation methods for chatbots,” in *Proceedings of the 2019 7th International Conference on Information and Education Technology*, 2019.

[39] “A virtual assistant to help doctors in their daily work,” 2021. [Online]. Available: https://www.safeinbreastfeeding.com/safedrugbot-chatbot-medical-assistant/. [Accessed 15 01 2022].

[40] “Florence your health assistant,” 2021. [Online]. Available: https://www.florence.chat/. [Accessed 14 01 2022].

[41] superizzy, 2021. [Online]. Available: https://www.facebook.com/superizzyai/about/.

[42] “Automate Nutrition Coaching with AI,” 2020. [Online]. Available: https://getforksy.com. [Accessed 14 01 2022].

[43] “babylon,” 2021. [Online]. Available: https://www.babylonhealth.com. [Accessed 14 01 2022].

[44] D. Chittamuru, S. Ramondt, R. Kravitz and 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*, 2019.

[45] “buoy,” 2021. [Online]. Available: https://www.buoyhealth.com. [Accessed 14 01 2022].

[46] “CancerChatbot,” 2021. [Online]. Available: https://www.facebook.com/CancerChatbot/. [Accessed 14 01 2022].

[47] “sensely,” 2021. [Online]. Available: http://www.sensely.com. [Accessed 14 01 2022].

[48] “Gyant,” 2021. [Online]. Available: https://gyant.com. [Accessed 14 01 2022].

[49] “Woebot health,” 2021. [Online]. Available: https://woebot.io. [Accessed 14 01 2022].

[50] K. K. Fitzpatrick, A. Darcy and 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,* 2017.

[51] “healthtap,” 2021. [Online]. Available: https://www.healthtap.com/. [Accessed 14 01 2022].

[52] “healthily,” 2021. [Online]. Available: https://www.your.md. [Accessed 14 01 2022].

[53] “Health powered by ADA,” 2021. [Online]. Available: https://ada.com. [Accessed 14 01 2022].

[54] A. Zagorecki, P. Orzechowski and K. Hołownia, “Online diagnostic system based on Bayesian networks,” in *Conference on Artificial Intelligence in Medicine in Europe*, Berlin, Heidelberg, 2013.

[55] “bots4health,” 2021. [Online]. Available: https://cristinasantamarina.com/work/bots4health/. [Accessed 14 01 2022].

[56] N. Siangchin and T. Samanchuen, “Chatbot Implementation for ICD-10 Recommendation System,” in *International Conference on Engineering, Science, and Industrial Applications (ICESI)*, Tokyo, Japan, 2019.

[57] M. A. Walker, D. J. Litman, C. A. Kamm and A. Abella, “PARADISE: A framework for evaluating spoken dialogue agents,” *arXiv preprint cmp-lg/9704004.,* 1997.

[58] D. Peras, “Chatbot evaluation metrics.,” in *Economic and Social Development: Book of Proceedings*, 2018.

[59] J. Pereira and O. Díaz, “A quality analisys of Facebook Messanger's most popular chatbot,” in *Proceedings of ACM SAC Conference,*, Pau, France, 9-13 April, 2018.

[60] P. B. Brandtzaeg and A. Følstad, “Why People Use Chatbots,” in *International Conference on Internet Science, (pp. 377-392)*, 2017.

[61] M. S. K. Caley, “Estimating the future healthcare costs of an aging population in the UK: expansion of morbidity and the need for preventative care,” *Journal of Public Health,* vol. 33(1), pp. 117-122, 2011.

[62] “ELIZA Terminal,” 1966. [Online]. Available: http://www.masswerk.at/elizabot/eliza.html. [Accessed 14 01 2022].

Appendix (A) AHP pairwise calculated for all measures

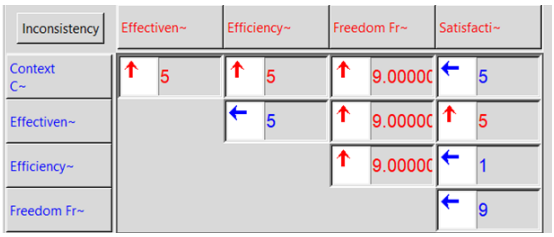
[Digitare qui]

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	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
Efficiency		resources expended in relation to the accuracy and completeness with which users achieve goals	real time information	5	5	-5	3
			low cost/free information predilection	3	1	1	1
			web service information instead of phisical logistic	3	1	1	1
Satisfaction	Usefulness	degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the resultsof use and the consequences of use	accuracy related to lineguide	5	1	5	3
			accuracy related to territory	5	3	1	3
			completness	5	3	1	3
			consistency respect the EBM	3	2	1	2
			personalized information	2	1	1	1
	Trust	degree to which a user or other stakeholder has confidence that a product or system will behave as intended	certified by third medical-parties (credibility)	2	0	1	0
			mediated by doctors	2		9	1
			linked to the sources	5	3	1	3
	Pleasure	degree to which a user obtains pleasure from fulfilling their personal needs	Personalized information	2	1	1	1
			supported by feedback from others	2		1	
			psychological support	2		1	
			Gracefgul degradation	2		1	
			Effective function allocation	5	2	1	2
			gramatical fit	5	5	1	5
			meaning fit	5	5	1	5
			visual look	5	2	1	2
	Comfort	degree to which the user is satisfied with physical comfort	multichanneling	2		1	
			human like interaction	5	3	1	3
			Linguistic accuracy of output	5	4	1	4
			multimedia interaction	2		1	
			on demand and real time information retreaving	5	5	-9	1
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	information accompanied by economic and financial rights	2		1	
	Health and Safety Risk Mitigation	degree to which a product or system mitigates the potential risk to people in the intended contexts of use	Robustness to manipolation	5	5	1	5
			certified information	5	3	1	3
			care giver involvment	2	0	1	
			medics involvement	2	0	9	1
			provide mechanisms to avoid interaction during travel	2		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 mechanisms to avoid interaction during travel	2		1	
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	2	1	1	1
			linking information to other similar user feedback	2		1	
			providing mechanism to ranking information depending user objectives	2		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	patient centered language	2	1	1	1
			medical stakeholder language	2		1	
			care giver involvement	2		1	

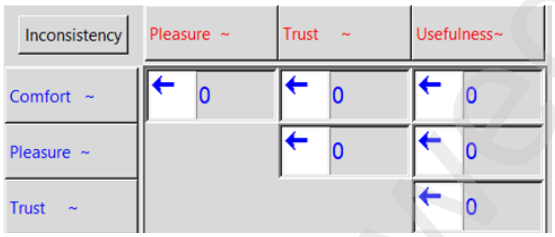
ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	Providing prescriptions (Measures)	Assistente Sanitario V. 1		Assistente Sanitario V. 2	
				Score granularity (n)	*Q_Score*	pairwise	*Q_Score*
Effectiveness		*accuracy and completeness with which users achieve specified goals*	provide prescription	2	0	1	0
			provide suggestion	2	1	1	1
			formal sending (pdf, email, legalmail)	2	0	1	0
Efficiency		*resources expended in relation to the accuracy and completeness with which users achieve goals*	product/service suggestion	2	1	1	1
Satisfaction	Usefulness	*degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the resultsof use and the consequences of use*	concreteness and practicability	5	4	1	4
			lineguide respondance	2	1	1	1
			personalized	2	1	1	1
			supplied in time	2	1	1	1
	Trust	*degree to which a user or other stakeholder has confidence that a product or system will behave as intended*	certified by doctors	2	0	1	0
			linked to scientific, lineguide, EBM sources	2		1	
	Pleasure	*degree to which a user obtains pleasure from fulfilling their personal needs*	take into account the user's inclinations or ethical choices	2	0	1	0
			effective function allocation	5	2	1	2
	Comfort	*degree to which the user is satisfied with physical comfort*	direct virtual interaction with clinical stakeholder	2	0	1	0
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*	consider whether an insurance policy has been taken out	2	0	1	0
			providing price benchmark	2	0	1	0
	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 compliant	2	1	1	1
			EBM compliant	2	1	1	1
			validated by medics	2	1	1	1
			care giver involvment	2	0	1	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 mechanisms to avoid interaction during travel	2	0	1	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	2	1	1	1
			provide mechanisms for formulating different hypotheses (ex. diagnosis) on which to give prescriptions or recommendations	2		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 anespected input	5	2	1	2

ISO25010 "Quality in use" characteristics	ISO25010 "Quality in use" subcharacteristics	characteristic description	Process management (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*	indirect process information grasping for better answers and process management	2	0	9	1
Efficiency		*resources expended in relation to the accuracy and completeness with which users achieve goals*	low finalized interaction for information grasping (indirect	2	1	1	1
			patient/medics interaction	2		9	1
			patient/PA interaction	2		1	
Satisfaction	Usefulness	*degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the resultsof use and the consequences of use*	tasks alignment	5	1	5	3
			times alignment	5	1	5	3
			costs alignment	5	1	5	3
	Trust	*degree to which a user or other stakeholder has confidence that a product or system will behave as intended*	real pathway state corrispondance	5	1	5	3
			completeness in the examination of the patient's datalog	5	1	5	3
	Pleasure	*degree to which a user obtains pleasure from fulfilling their personal needs*	predict in advance the next tasks to be performed	5	1	7	4
			connect all the stakeholder in the clinical pathway	5	2	3	3
			performing tasks privileging solutions, open, low cost, public heath based, obtaining the same autcome	5	3	1	3
			effective function allocation	5	3	1	3
	Comfort	*degree to which the user is satisfied with physical comfort*	use of IoT for health parameter measuring	2		1	
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*	case management	5		7	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	7	4
			protecd and respect privacy	5	5	1	5
			care giver involvement	2		9	1
			Avoid inappropriate utterrances and be able to perform damage control	5	4	1	4
			off line personal health datalog access	2		1	
			ranking process state grasping	3	1	9	3
			history of execution tracking	2	0	1	0
			off road detection	3	1	9	3
			patient/medics interaction	2	0	9	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 mechanisms to avoid interaction during travel	2		1	
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	2		9	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*	robusteness to unclearness and enoughness infomation in the datalog patient	5	1	5	3
			Robustness to anespected input	5	2	7	5

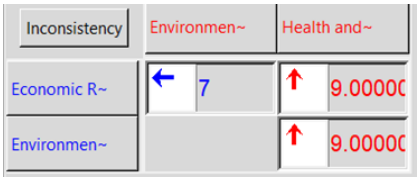
Weights of characteristics and subcharacteristics



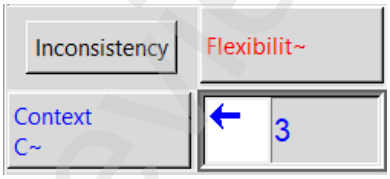
ISO/IEC 25010 "Quality in use" Criteria



Satisfaction-SUB_Criteria



Freedom From Risk-SUB_Criteria



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 on the go so that they can guide the patient in the path they must follow in case of breastfeeding	[39]
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	[40])
Izzy	Messenger	helps women track their period provides information on users' sexual issues and menstrual health reminds them when to take birth control pills	[41])
Forksy	Messenger	assists in tracking calories promotes healthy eating habits food diary	[42])
Babylon Health	mobile App	remote consultation with health care professionals and doctors patient's medical history database symptom checker	[43])
Buoy Health	website	assist patients in diagnosing	[44] [45](buoy,)
CancerChatbot	Messenger	offers detailed information on cancer and related topics	[46]
Sensely	mobile App	tracks health symptoms using both text and speech communication diagnosis formulation tries to understand the level of emergency	[47])
GYANT	Messenger Alexa	symptom checker	[48]
Woebot	mobile App	studies patient mood, personality and suggests remedies as a therapist for depression	[49] [50]
HealthTap	Messenger	physician-patients communication channel via bot	[51]

		make its vast repository of knowledge available to patients using the app	
Your.Md	Messenger Slack KIK Telegram	symptom checker	[52]
Ada Health	mobile App alexa	symptom checker	[53]
Infermedica	mobile App	symptom checker	[54]
Bots4Health	mobile App	sexual and reproductive health chat about a wide range of health issues	[55]
Assistente Sanitario	Telegram	Clinical data repository Provides semantically annotated information	[28]