An evaluation framework for unconstrained conversational agents in healthcare: a scoping review

Hang Ding[1], Joshua Simmich[1], Atiyeh Vaezipour[1], Nicole Andrews[1,2,3], Trevor Russell[1]

[1] RECOVER Injury Research Centre, Surgical Treatment and Rehabilitation Services, Herston, The University of Queensland, QLD, Australia

[2] Tess Cramond Pain and Research Centre, Metro North Hospital and Health Service, Herston, QLD, Australia

[3] Occupational Therapy Department, The Royal Brisbane and Women's Hospital, Herston. QLD, Australia

Introduction

Conversational agents (CAs), also known as chatbots or virtual assistants, are software programs designed to imitate human conversations to engage with users.

Over the past decade, CAs have been increasingly studied to address care burdens and service needs in healthcare delivery. For example, studies have demonstrated the potential to use CA-enabled care programs to assist in triaging people with unurgent health conditions, supporting inpatient care, providing post-discharge follow-ups, and self-managing mental health and chronic diseases.

To use CAs safely and effectively, rigorous evaluations are essential but challenging to achieve. Although reviews have provided some technical metrics and outcome measures, selection of these variables for individual studies remains challenging. To achieve robust and effective outcomes, evaluation frameworks are often needed to help understand essential design, outcome measures, and evaluation targets at different trial stages. However, such frameworks remain absent.

Our aim: To synthesize existing knowledge and outline a framework for evaluating CAs in healthcare.

Method

We conducted a scoping review according to the PRISMA Extension for Scoping Reviews.

We searched CINAHL, Medline, Scopus, Embase and IEEE Xplore, focusing on CAs which were unconstrained by predefined answers or options. We reviewed study designs, categorised outcome measures, and finally outlined an evaluation framework which can be used to evaluate CAs in healthcare.

To synthesise the framework, we extracted study designs and outcome measures, nested them within well-recognised categories, and mapped the results on an established overarching framework for digital health evaluations.

Results

The search identified 1553 articles, of which 43 studies were included in the review.

We identified 23 quasi-experimental studies, nine randomised controlled trials, four observational studies, and seven research test-based studies

A total of 175 outcome measures which were used in the reviewed studies were nested into seven categories: 1) functionality, 2) clinical / health outcomes, 3) user experience, 4) costs and cost benefits, 5) safety and information quality, 6) usage, adherence and uptake, and 7) user characteristics for implementation research. We consolidated the framework in the figure below, which shows the outcome measures across four evaluation stages: I) feasibility/usability, II) efficacy, III) effectiveness, and IV) implementation science.

Discussion

In this review, we found that many studies evaluated CAs using predefined questions, unable to reflect overall CA performances. New strategies to comprehensively evaluate CAs are needed.

We highlighted that some evaluations involving researchers and clinicians are essential, especially for evaluating information quality and risks of interventions, because normal users or participants often have insufficient knowledge to judge the quality or safety of CA-based interventions.

The outcome measures of functionality, user experience, safety and information quality are often diverse and inconsistent. Validated questionnaires and algorithms are needed to achieve robust and effective evaluations.

We identified several measures to be used in future research, such as stability, consistency, standard compliance, and measures to reduce health inequities.

This study is limited to using a single evaluation framework to map the investigation results.

Conclusion

This systematic review presents a consolidated evaluation framework which can be used to evaluate the performance of CAs in healthcare.

Figure: The consolidated framework for evaluating conversational agents in healthcare. The framework demonstrates the studies (n=43) and their outcome measures at four major evaluation stages of an established practical guide, named Monitoring and Evaluating Digital Health Intervention, the World Health Organization. Note: Essential measures at different stages, which we identified, are mark by a light blue. * denotes the measures which we proposed to be included in future studies.

	Stage →	1. Feasibility & Usability →	2. Efficacy →	3. Effectiveness →	4. Implementation Science
WHO Digital Health	**Brief description**	Feasibility: The ability to work as intended. Usability: The degree of a system being used to achieve specified goals in a specified context of use.	Efficacy: The ability to achieve the intended results in a research setting or trial.	Effectiveness: The ability to achieve the intended results in a real application (non-research setting).	Implementation science: To assess the uptake, integration and sustainability of evidence-based digital health interventions for a given context, including policies and practices.
WHO Digital Health	**Evaluation targets**	◆ Stability (system uptime/failure rates) ◆ Performance consistency ◆ Standards adherence (terminology, interoperability, security)	◆ User satisfaction ◆ Workflow "fit" ◆ Learning curve (design) ◆ Cognitive performance / errors ◆ Reliability	◆ Changes in care processes (time) ◆ Changes in outcomes (system performance / health)	◆ Changes in process, outcome, coverage, and costs ◆ Total cost of implementation, and health impact ◆ Error rates ◆ Learning curve of users ◆ Changes in policy, practices attributable to system ◆ Adaptability and extendibility to new use-cases
Studies Reviewed and Outcome Measures at Four Major Evaluation Stages Aligned with the WHO Guide	**Studies included in the review**	Almusharraf 2020 [1]; Bibault 2019 [9]; Gabrielli 2020 [2]; Park 2019 [10]; Gaffney 2020 [3]; Suganuma 2018 [11]; Boczar 2020 [4]; Bickmore 2018 [12]; Bonnevie 2020 [5]; Philip 2014 [13]; Denecke 2020 [6]; Yasavur 2014 [14]; Rehman 2020 [7]; Rhee 2014 [15]; Stephens 2019 [8]	Davis 2020 [16]; Auriacombe 2018 [23]; Polignano 2020 [17]; Fulmer 2018 [24]; Caballer 2020 [18]; Fitzpatrick 2017 [25]; Maher 2020 [19]; Friederichs 2014 [26]; Lee 2020 [20]; Harless 2009 [27]; Bennion 2020 [21]; Bian 2020 [22]	Maeda 2020 [28]; Philip 2017 [33]; Dosovitsky 2020 [29]; Ly 2017 [34]; Chaix 2019 [30]; Crutzen 2010 [35]; Bott 2019 [31]; Perski 2019 [32]	Yang 2021 [36]; Nobles 2020 [41]; Fan 2021 [37]; Boyd 2018 [42]; Schindler-Ruwisch 2020 [38]; Miner 2016 [43]; Kocaballi 2020 [39]; Ferrand 2020 [40]
Studies Reviewed and Outcome Measures at Four Major Evaluation Stages Aligned with the WHO Guide	**Study designs**	Single-arm studies[1-8,10,12-15] Randomized controlled trials[9] Two-arm quasi-experimental study[11]	Single-arm studies[16-20,23,27] Case-control study[22] Randomized controlled trials[21,24-26]	Cross-over study [33] Case-control study[31] Cross-sectional study[29,30,35] Randomized controlled trials[28,32,34]	Research test[36,38-43] Cross-sectional study[37]
Studies Reviewed and Outcome Measures at Four Major Evaluation Stages Aligned with the WHO Guide	**User characteristics**	- User characteristics [5]	-	- User characteristics [30,35]	- User characteristics [37] - Users in geographic regions * - Gender, equity and rights – to reduce health inequities *
Studies Reviewed and Outcome Measures at Four Major Evaluation Stages Aligned with the WHO Guide	**Usage, adherence and uptake**	- Usage [3,5,8] - Uptake [5,15]	- Usage [16,19,21,22,24] - Uptake [26] - Adherence [16,22,26]	- Usage [29,30,32,35] - Uptake * - Adherence [34]	- Usage [37] - Uptake * - Adherence [37]
Studies Reviewed and Outcome Measures at Four Major Evaluation Stages Aligned with the WHO Guide	**Costs and cost benefits**	- Costs [5]	- Cost effectiveness [22]	- Costs * - Cost effectiveness *	- Implementation costs *
Studies Reviewed and Outcome Measures at Four Major Evaluation Stages Aligned with the WHO Guide	**Clinical / health outcomes**	- Knowledge and skills[3] - Health wellbeing and issues [11] - Psychological / mental health [3,11] - Clinical assessment performance [13] - Behavioral modification and risk factors [8]	- Knowledge and skills [21,27] - Psychological / mental health [21,24,25] - Clinical assessment performance [18,20,23] - Behavioral modification and risk factors [16,19,26]	- Knowledge and skills [28] - Health wellbeing and issues [31,34] - Psychological / mental health [28,31,34] - Clinical assessment performance [33] - Behavioral modification and risk factors [28,30,32]	- The effectiveness of the approved intervention in less controlled environment *
Studies Reviewed and Outcome Measures at Four Major Evaluation Stages Aligned with the WHO Guide	**User experience**	- Usability [15] - Feasibility [2,8] - Usefulness [3] - Ease of use [4] - Satisfaction [12] - Open comments [10,15] - Overall experience [3,6,7,10,12,14] - Acceptance / preference [4] - Conversational capability [4,6,13,14] - Perceived quality and trust [9] - Suggestions for improvement [1,2,10,15]	- Usability [21] - Feasibility [27] - Usefulness [16] - Ease of use [25] - Satisfaction [17,24-26] - Overall experience [16,17,20,25] - Other open comments [20] - Conversational capability [25] - Perceived quality and trust [20] - Acceptance / preference [20,21,23] - Suggestions for improvement [16,24]	- Usefulness [35] - Ease of use [35] - Satisfaction [30] - Overall experience [28,30,34] - Acceptance / preference [35] - Conversational capability [35] - Suggestions for improvement [30]	- Satisfaction [37] - Overall experience [37]
Studies Reviewed and Outcome Measures at Four Major Evaluation Stages Aligned with the WHO Guide	**Safety and information quality**	- Risk of causing death [12] - Risk of misinformation * - Risk of misunderstanding * - Risk of unintended harms [12]	- Risk of misinformation * - Risk of misunderstanding * - Risk of unintended harms *	-	- CA response capability [36,38,40,43] - Risk of misinformation [40] - Risk of unintended harms [43] - CA response appropriateness [36,38,39,41] - Resources and contents quality [38,43]
Studies Reviewed and Outcome Measures at Four Major Evaluation Stages Aligned with the WHO Guide	**Functionality**	- Response speed [12] - Task achievements [7,12,14] - Engagement functions [8,12,14] - Classification accuracy [1,7] - Understanding and accurate responses [4] - Stability, consistency, and standard *	- Understanding and accurate responses [16]	-	-

h.ding@uq.edu.au
Dr Hang Ding

RECOVER Injury Research Centre

Acknowledgements

We thank Ms. Jackie Devenish from UQ library for her guidance and support on developing the literature search strategy.

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE