Evaluating Personalized Pervasive Health Technology - But How?

Bardram, Jakob E.

Published in:
IEEE Pervasive Computing

Link to article, DOI:
10.1109/MPRV.2020.2989880

Publication date:
2020

Document Version
Peer reviewed version

Evaluating Personalized Pervasive Health Technology – But How?

Jakob E. Bardram
Copenhagen Center for Health Technology
Department of Health Technology, Technical University of Denmark
Email: jakba@dtu.dk

IN the call for this special issue on Personalized Pervasive Health it was stated that it "intends to provide a comprehensive view on innovative pervasive computing methods, ubiquitous technology, data-based inference algorithms, as well as evaluation studies, all related to personalized health. [...] All investigations must include *thorough evaluations of their approaches and methods*." (my emphasis). When reading this, I came to wonder what "thorough evaluations" actually is, in this context? Are authors supposed to show clinical evidence for the health efficacy of their technology? Or, are they supposed to show that the technology is technically sound and working? Or, that it is secure and has appropriate privacy-protection of sensitive personal data (c.f. the focus on this special issue). Or, that the technology is usable and user-friendly for the users? Or, ...?

These questions touch upon a more fundamental question, namely how researchers can evaluate novel ubiquitous computing technology in the health domain in a manner that allows them to make meaningful claims about their utility and argue for their scientific contributions in the broader health technology domain.

Evaluation of health technology has always been difficult and subject to significant scientific disputes. From a *technological perspective*, we are mainly interested in the design of novel technology and understanding how it works under different circumstances, while gradually and iteratively improving on its technical features and capabilities. This calls for formative evaluation methods, which help us understand how the technology works and how it can be improved according to a set of design goals. In this approach, we seek to understand the technology and look 'into' it, i.e. a white-box evaluation strategy.

From a *health perspective*, we are mainly interested in the efficacy of the technology in terms of clinical outcome related to screening, diagnosis, treatment, or care of patients. We care about establishing solid evidence for the clinical claims of a technology and less about how it actually works from a technical point-of-view. This calls for summative evaluation methods, which compare the outcome of using the technology to some control situation. In this approach, we seek to understand the outcome of its use, and care less about what is inside the technology, i.e. a black-box evaluation strategy.

This tension between different methodological approaches in the technical and health sciences has been managed for many years in the more traditional medical device domain. However, with the increasing profiliation of novel technological opportunities – not least coming from the mobile and ubiquitous computing area – this tension has increased. We have seen an explosion of the use of mobile and wearable technology in health. For example, in a recent review in this magazine, we found 46 different mobile and wearable applications that have been introduced in the mental health domain over the last decade [3]. Similarly, a recent consensus report on diabetes digital app technology found, across the USA and Europe, that mobile health (mHealth) apps intended to manage diabetes health and wellness were largely unregulated and lacked any kind of evidence for their safety and efficacy [9].

This challenge of providing appropriate evaluation methods for health technology is a topic that, fortunately, has gained increased attention lately. In this spotlight article, I will first try to outline how different attempts to address this challenge have emerged lately from both the technological as well as the health sciences. Then, I will share some insight and experience on how we have approached evaluation of personal health technology in the Copenhagen Center for Health Technology (CACHET). I hope this can be of use for others who are facing the question of how to evaluate personal health technology, and guide them in the design of an appropriate evaluation strategy.

I. Evaluation Approaches

Methodologically, the design of health technology, including the growing research into 'Personalized Pervasive Health Technology' [2], sits at the intersection of design science and health science.

On the one hand, health technologies need to be designed, developed, and refined in a design process, which often relies on technological and user-centered design methodologies. In the biomedical engineering sciences, a novel technology is often evaluated according to a performance standard. For example, a classic approach to evaluate the accuracy of a novel approach to detect hearth arrhythmia, such as atrial fibrillation, is to compare the proposed approach to labelled data, such as the PhysioNet database (e.g. [16]). In computer science, the design of a novel mHealth application is typically evaluated using a user-centered approach. For example, in a recent study of a recommender system for depressive patients, the Unified Theory of Acceptance and Use of Technology (UTAUT) methodology [19] was used to access the perceived usefulness and usability of the system [17].

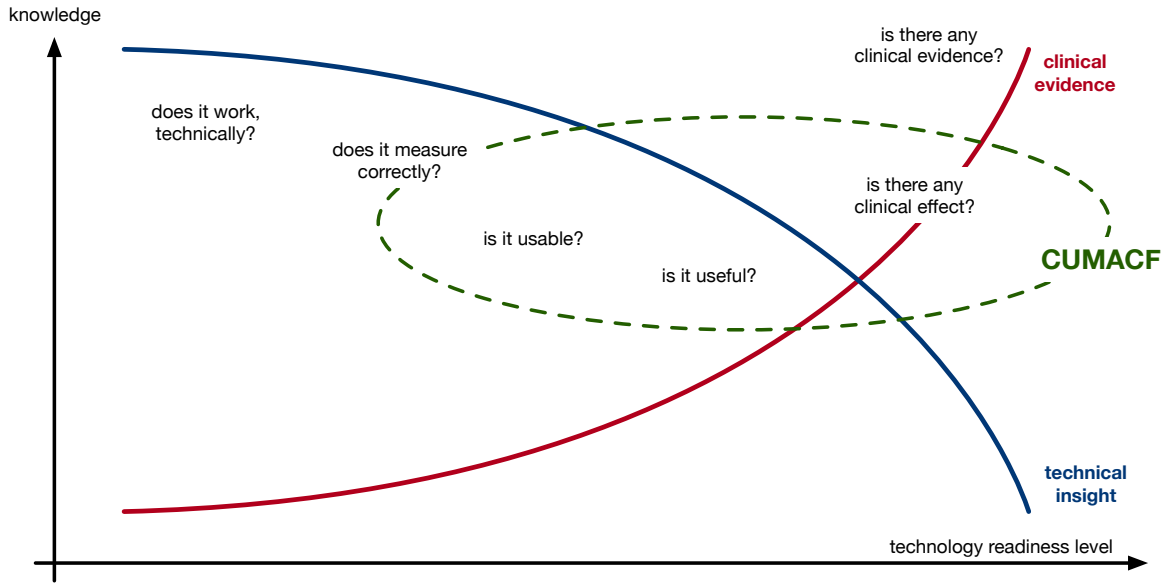


Fig. 1. Addressing different evaluation questions during different phases of the technology development and hence technology readiness levels. CUMACF: CACHET Unified Methodology for Assessment of Clinical Feasibility.

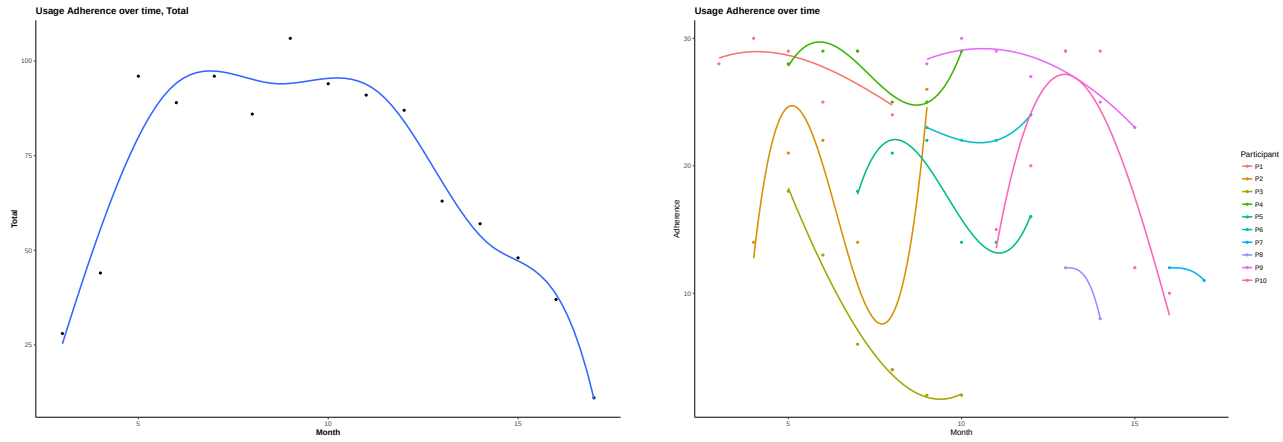
On the other hand, health technologies need to be clinically verified in order to assess clinical safety, efficacy, and effectiveness. From a health-oriented perspective, a carefully designed randomized controlled trial (RCT) represents the highest level of evidence in evidence-based medicine (EBM), and is the 'gold standard' for determining the effectiveness of pharmacological agents. This approach has been transferred to evaluating non-pharmacological interventions, including health technology. The Consolidated Standards of Reporting Trials (CONSORT) statement is a guideline that specifies how to report RCTs and is applied in all papers within EBM [18]. In order to accommodate specific issues related to describing an mHealth intervention, the CONSORT guidelines were extended to the Consolidated Standards of Reporting Trials of Electronic and Mobile HEalth Applications and on-Line TeleHealth (CONSORT-EHEALTH) [7]. This guideline suggests to expand the description of the intervention of a mHealth study to include more detailed information on the technology, including how it was funded, designed, developed, and deployed. It also recommended to provide access to the technology and its source code, by e.g. releasing it as open source, or at least release a video containing a detailed walk-through of the system and its features.

There is, however, a growing awareness that these two methodological positions represent two opposite 'poles' that have quite different scientific standards for the different stages in research, including how to evaluate and measure the applicability of health technology [4]. The traditional RCT has limitations when used to evaluate health technology; for example, it does not permit iterative improvements to the design and the technology may be outdated by the time the trial is completed [13]. Moreover, since the RCT methodology is summative – i.e., measures outcomes before and after an intervention – it treats the intervention as a 'black-box' and is not suited to helping researchers understand which parts of the intervention (i.e., features of the technology) are actually causing an effect. For example, even though it seems like mHealth interventions can reduce depressive symptoms [8], it is unclear which features of these systems actually account for this effect. Is it, for example, important to provide support for self-assessment? Or in-person feedback? Or cognitive training? Such detailed information cannot be collected from a traditional RCT, since it treats the technology as an all-or-nothing intervention.

For this reason, a number of alternative evaluation methodologies have been proposed. From a health science perspective, the CONSORT guidelines have been extended to also encompass pilot and feasibility trials conducted in advance of a future definitive RCT [6]. The primary aim of a pilot or feasibility trial is to assess feasibility of conducting the future definitive RCT. Such a feasibility study can be useful to test the applicability of a novel technology before moving into an actual RCT. The study can test the feasibility of using the technology for a specific group of patients, the recruitment and on-boarding procedure, and the technical stability and scalability of the technology in real use.

It is, however, important to bear in mind that a clinical study – both a feasibility study and a RCT – is costly and lengthy. It often requires the enrollment of hundreds of patients to ensure statistically significant results, it runs over several months, and it needs a staff of skilled doctors and healthcare professionals to run them. It has been estimated that an RCT costs $41,000 per patient [14]. Therefore, while RCTs are important for evaluation of clinical effectiveness, they are best undertaken only when: i) the intervention and its delivery package are stable, ii) the intervention can be implemented with high fidelity, and iii) there is a reasonable likelihood that the overall benefits will be clinically meaningful [15].

From a design science perspective, it has been argued that health technology should be evaluated from a more formative



(a) (b)

Fig. 2. Example of usage adoption over the duration of a study (smoothed). (a) Total for all participants, and (b) per participant.

and qualitative approach before moving into costly clinical trials. Klasnja et al. [11] provide a set of compelling arguments for why it is important to evaluate behavior change and health technology from a human-computer interaction (HCI) perspective and apply user-centered methods. This helps 'open the black box' to uncover potential problems in the technology and understand the details of its use. It is, for example, a fundamental problem if the user experiences difficulties entering food items in a food tracking app, and a (small) usability problem like this might jeopardize a large and expensive clinical trial. Murray et al. [15] propose a heuristic evaluation method for digital health interventions (DHIs), which consists of 13 questions assessing whether there is a need and population for the technology, if it might be of benefit and impact, and other core aspects. This simple, question-driven approach to the evaluation of a DHI can lead to an accumulating knowledge base around an intervention in a timely and cost-efficient manner.

In summary, there are different questions that are relevant to address in different stages of the development of health technology. Fig. 1 illustrates this; in the early stage of technology development, it is mostly relevant to investigate the technical feasibility of the technology, including its technical capability and accuracy. Once this is in place, it becomes relevant to assess the usefulness and usability of the technology. When the technology is more mature and stable, assessment of potential health effects can be commenced, and in the end clinical evidence can be established. When reporting an 'evaluation' of a novel health technology it is essential to be explicit about where in this space the evaluation is positioned and, subsequently, which kind of 'claim' or conclusion can be drawn from such an evaluation. It is, for example, impossible to drawn any conclusions about clinical efficacy or utility based on technical or usability studies.

II. Assessing Clinical Feasibility

There is a growing need to be able to design and develop health technologies while being able to point to health benefits – in particular in the early stages of technology development and evaluation. For this purpose, technical and health scientists at the Copenhagen Center for Health Technology (CACHET) worked together to create the CACHET Unified Methodology for Assessment of Clinical Feasibility (CUMACF) methodology. The goal of CUMACF is to help researchers in the process of designing and developing health technology to run what we call '*clinical feasibility studies*', i.e., studies that help researchers understand whether the technology under design would be feasible to use in future health interventions, if implemented in clinical use. The position of CUMACF in the evaluation space is illustrated in Fig. 1.

The purpose of CUMACF is twofold. First, borrowing from a design science perspective, CUMACF seeks to support an iterative design process, with frequent design and evaluation sessions involving end-users. The idea is to investigate the feasibility of the technology under design as early as possible – this saves time, effort, and money. Moreover, in contrast to a traditional RCT in which the intervention is treated as a black-box, CUMACF seeks to provide an understanding of the intervention (i.e., the technology under design) by providing insights into which parts of the technology (i.e., which features) help achieve a health outcome. Second, borrowing from a health science perspective, CUMACF seeks to investigate health efficacy, i.e., the extent to which an intervention does more good than harm under ideal circumstances [10]. The goal is to gather early evidence on potential efficacy during design. CUMACF will not establish a high level of evidence since the study typically does not involve a control group and has insufficient statistical power. But a clinical feasibility study will help researchers understand the potential of the technology for health efficacy at an early stage and help researchers understand which other parameters, besides the technology itself, need to be (re)designed in order to obtain the desired health outcome.

A practical guide on how to use CUMACF is described in a technical report [1]. Below we will provide an overview of the method and provide insights into its use.

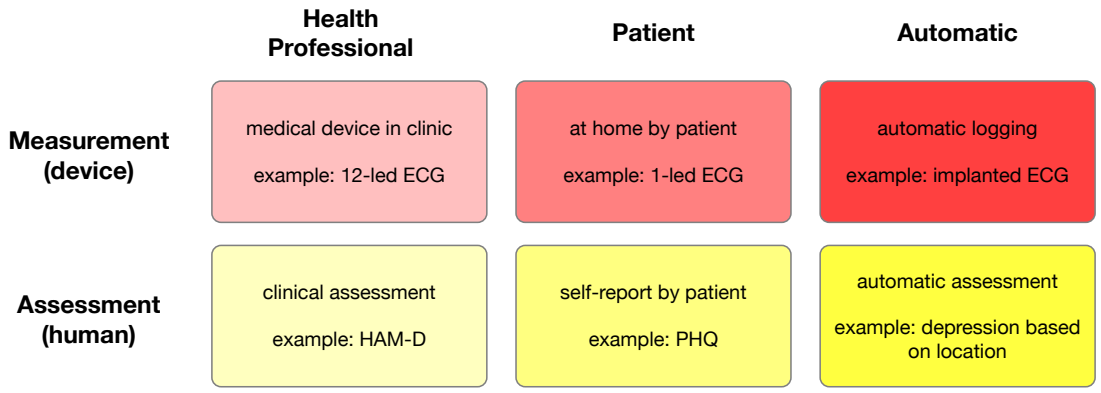


Fig. 3. Taxonomy of health outcome measures.
ECG: Electrocardiography; HAM-D: Hamilton Depression Rating Scale; PHQ: Patient Health Questionnaire.

III. Method

CUMACF targets the evaluation of 'Personal Health Technology' [2] designed for personal use by a patient. It is divided into three parts, assessing; (i) usage adoption, i.e., the degree to which the patient uses the technology; (ii) perceived usefulness and usability, i.e., the likelihood of successful adoption of the technology and acceptance by users; and (iii) health efficacy, i.e., the capacity for beneficial change or therapeutic effect of the intervention provided by the technology. All three things are assessed simultaneously during a clinical deployment of the technology.

A. Usage Adoption

A core prerequisite for assessing the feasibility of a health technology is to know whether the patient; (i) uses the system in the first place, and (ii) uses the technology as instructed and prescribed. To verify this, assessment of *usage adoption*[1] is beneficial. Usage adoption is a relative measure; it assesses to what degree the user uses the technology as compared to what is expected. For example, if a patient is asked to assess depression level as a daily mood score, the adoption rate is the percentage of days of self-reported mood compared to the number of days the patient was enrolled in the study.

Calculating usage adoption depends on knowing the baseline, i.e., how often the user is supposed to use the system. This depends on what the user has been instructed to do (e.g., filling in a daily mood score), the duration of the study per participant, and the availability of the system. Given this, adoption for a participant i can be calculated as:

$$adoption_i = \frac{usage_i}{duration_i - downtime_i}$$

If the above data is collected, detailed statistics on usage adoption can be reported; (i) overall, (ii) over time, and (iii) per participant, all of which take system availability into account. Fig. 2 shows an example of how this analysis can be done.

[1]Some call this *'adherence'* to technology use, borrowing the term from clinical studies and treatment. However, we prefer not to use the term 'adherence' since it carries a connotation that the technology is 'prescribed' and 'should' be used like a medical drug. This is often not the case with technology which is seldom 'prescribed'.

B. Perceived Usefulness and Usability

The second part of CUMACF is to measure *perceived usefulness and usability*. According to research into psychometric assessment of technology acceptance, there is a strong correlation between users' perceived usefulness and usability of a system and the likelihood of future successful adoption and acceptance of the technology [5]. For example, a study showed that 80% of all activity tracking devices were abandoned after two months because "participants perceived the data collected as not useful" [12]. Hence, perceived usefulness is key for technology adoption.

CUMACF follows the Unified Theory of Acceptance and Use of Technology (UTAUT) methodology [19] combined with a few usability and behavior change questions, and is designed to assess the user's intention for *future* acceptance of the technology.

CUMACF applies a questionnaire, which is specifically designed to collects data on the likelihood of successful adoption of technology, its acceptance by the users, and their intentions to use it for the intended health outcome. This is done according to the following five dimensions, adopted from UTAUT: (i) *health expectancy*, assessing the degree to which an individual believes that using the system will help the individual to attain *gains in health*; (ii) *effort expectancy*, assessing the degree to which an individual believes that *ease* is associated with use of system; (iii) *social influence*, assessing the degree to which an individual perceives that *important others* believe the individual should use the system; (iv) *facilitating conditions*, assessing the degree to which an individual believes that an *organizational and technical infrastructure* exists to support use of the system; and (v) *behavioural intention*, investigating the degree to which an individual *intends to use the system*.

The CUMACF technical report [1] contains the entire questionnaire and detailed instructions on how to deploy it, and how to analyze and present statistics on perceived usefulness and usability.

C. Health Efficacy

Since CUMACF focuses on 'feasibility', the methodology focuses on establishing *health efficacy*, i.e., involving patients

who are carefully diagnosed, have significant symptoms from the disease in question, lack other serious illnesses, and are likely to follow and respond to the treatment based on the technology [10].

Based on our experience in running several clinical feasibility studies, we have established the following general guidelines: (i) The number of participants (*N*) should be circa 20. (ii) Patients should be *recruited* who are carefully diagnosed and who potentially can benefit from the intervention, while at the same time are early adopters, i.e. have the skills, motivation, and ability to use the technology. (iii) Even though it needs to be adapted to the specific technology, the *duration* of intervention per patient should seldom be longer than six months. (iv) *Compensation* should be tailored to local ethics guidelines, but the technology and the infrastructure should be provided free of charge, including mobile and wearable devices. (v) The study protocol should allow for *adaptation* during the study. However, this should be restricted to adaptation which only has a limited effect on the outcome measure of the study, and should primarily be addressing technical enhancements of non-functional software qualities, such as robustness, security, usability, and scalability.

Defining clinical outcome measures is clearly dependent on the health topic in question and the type of disease being addressed. For example, depressive symptoms may be measured using the Patient Health Questionnaire (PHQ) scale, whereas cardio-vascular disease symptoms are measured using an electrocardiography (ECG) device. In order to inspire researchers to come up with different outcome measures for a study, CUMACF provides a taxonomy that makes a distinction according to; (i) *how* health outcome measures are obtained versus (ii) *who* measures it. This taxonomy, with some examples, is illustrated in Fig. 3.

THESE years, we see an an increasing number of mobile and wearable health technologies being designed and put into use. Most of these are designed to meet the healthcare challenges we are facing in terms of a growing demand with reduced availability of clinical resources. It is important to establish the clinical evidence of such technologies, i.e., do they actually address the clinical need they claim to? For this evaluation, the current golden standard is a randomized controlled trial (RCT). However, an RCT is expensive, time consuming, and treats the technology as a black-box. For this reason, this method is less useful during the early design, development, and feasibility testing of novel technology.

There is a need to strike a balance in order to assess the clinical feasibility of a new technology before planning a full RCT. For this purpose, we have been crafting the CACHET Unified Methodology for Assessment of Clinical Feasibility (CUMACF). The overall objective of CUMACF is to provide a standardized way to assess the 'feasibility' of a health technology during design and development. Such a standardized method will help to compare test results both *within* the iterative design of one specific technology as well as *between* different technologies. The former implies that a design team can assess the progression of its design across multiple iterations of the technology, whereas the latter implies that different technologies – maybe targeting the same health outcome – can be evaluated and compared in a more standardized manner. We hope that the reader finds inspiration in the CUMACF approach to evaluating clinical feasibility.

REFERENCES

[1] J. E. Bardram. CACHET Unified Methodology for Assessment of Clinical Feasibility. Technical report, Copenhagen Center for Health Technology, Copenhagen, Denmark, 2018. Available from http://www.cachet.dk/research/cumacf.

[2] J. E. Bardram and M. Frost. The Personal Health Technology Design Space. *IEEE Pervasive Computing*, 15(2):70–78, 2016.

[3] J. E. Bardram and A. Matic. A Decade of Ubiquitous Computing Research in Mental Health. *IEEE Pervasive Computing*, 19(1):62–72, 2020.

[4] A. Blandford, J. Gibbs, N. Newhouse, O. Perski, A. Singh, and E. Murray. Seven lessons for interdisciplinary research on interactive digital health interventions. *DIGITAL HEALTH*, 4, jan 2018.

[5] F. D. Davis. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3):319–339, sep 1989.

[6] S. M. Eldridge, C. L. Chan, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *Pilot and Feasibility Studies*, 2(1):64, oct 2016.

[7] G. Eysenbach. CONSORT-EHEALTH: Improving and Standardizing Evaluation Reports of Web-based and Mobile Health Interventions. *J Med Internet Res*, 13(4):e126, dec 2011.

[8] J. Firth, J. Torous, J. Nicholas, R. Carney, A. Pratap, S. Rosenbaum, and J. Sarris. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry*, 16(3):287–298, 2017.

[9] G. A. Fleming, J. R. Petrie, R. M. Bergenstal, R. W. Holl, A. L. Peters, and L. Heinemann. Diabetes digital app technology: benefits, challenges, and recommendations. A consensus report by the European Association for the Study of Diabetes (EASD) and the American Diabetes Association (ADA) Diabetes Technology Working Group. *Diabetes care*, 43(1):250–260, 2020.

[10] B. Haynes. Can it work? Does it work? Is it worth it?: The testing of healthcare interventions is evolving. *BMJ: British Medical Journal*, 319(7211):652, 1999.

[11] P. Klasnja, S. Consolvo, and W. Pratt. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3063–3072. ACM, 2011.

[12] A. Lazar, C. Koehler, J. Tanenbaum, and D. H. Nguyen. Why we use and abandon smart devices. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 635–646. ACM, 2015.

[13] D. C. Mohr, S. M. Schueller, W. T. Riley, C. H. Brown, P. Cuijpers, N. Duan, M. J. Kwasny, C. Stiles-Shields, and K. Cheung. Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies. *Journal of medical Internet research*, 17(7), 2015.

[14] T. J. Moore, H. Zhang, G. Anderson, and G. C. Alexander. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. *JAMA Internal Medicine*, 178(11):1451–1457, 11 2018.

[15] E. Murray, E. B. Hekler, G. Andersson, L. M. Collins, A. Doherty, C. Hollis, D. E. Rivera, R. West, and J. C. Wyatt. Evaluating digital health interventions: key questions and approaches. *Am J Prev Med*, 51(5):843–851, Nov 2016.

[16] A. Peimankar and S. Puthusserypady. An ensemble of deep recurrent neural networks for p-wave detection in electrocardiogram. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1284–1288. IEEE, 2019.

[17] D. A. Rohani, A. Q. Lopategui, N. Tuxen, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram. MUBS: A Personalized Recommender System for Behavioral Activation in Mental Health. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2020.

[18] K. F. Schulz, D. G. Altman, and D. Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, 8(1):18, 2010.

[19] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003.