Key considerations for designing clinical studies to evaluate digital health solutions

Elaina Bolinger, Benoit Tyl

Submitted to: Journal of Medical Internet Research
on: November 13, 2023

Table of Contents

Key considerations for designing clinical studies to evaluate digital health solutions

Elaina Bolinger[1] PhD; Benoit Tyl[2]

[1]Bayer AG Berlin DE
[2]Bayer HealthCare SAS La Garenne Colombes FR

Corresponding Author:
Benoit Tyl
Bayer HealthCare SAS
10 Place de Belgique
La Garenne Colombes
FR

Abstract

Evidence of clinical impact is critical to unlock the potential of Digital Health Solutions (DHS), yet many solutions are failing to deliver positive clinical results. We argue this failure is linked to current approaches to DHS evaluation design, which neglect numerous key characteristics (KCs) requiring specific scientific and design considerations. We first delineate the KCs of DHS: they are implemented at healthcare system and patient level, can drive multiple clinical outcomes indirectly through a multitude of smaller clinical benefits, their mechanism of action can vary between individuals and change over time based on patient needs, and finally, they develop through short, iterative cycles - optimally within a real-world use context. Finally, we provide research design suggestions that better address these KCs, including tips on mechanism of action mapping, alternative randomization methods, control arm adaptations, and novel endpoint selection, as well as innovative methods utilizing real-world data and platform research.

(JMIR Preprints 13/11/2023:54518)
DOI: https://doi.org/10.2196/preprints.54518

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Title:

Key considerations for designing clinical studies to evaluate digital health solutions

Elaina M. Bolinger PhD[1], Benoit Tyl MD[2]
[1]Integrated Evidence Generation & Business Innovation, Bayer AG, Berlin, Germany
[2]Integrated Evidence Generation & Business Innovation, Bayer Healthcare SAS, La Garenne Colombes, France

Corresponding Author: Benoît Tyl, 10 Place de Belgique, F-92254 La Garenne Colombes, France, email: benoit.tyl@bayer.com, Mobile: +33 6 80 29 07 79

Main text, Word count: 2509, two figures, no table

Abstract

Evidence of clinical impact is critical to unlock the potential of Digital Health Solutions (DHS), yet many solutions are failing to deliver positive clinical results. We argue in this Viewpoint that this failure is linked to current approaches to DHS evaluation design, which neglect numerous key characteristics (KCs) requiring specific scientific and design considerations. We first delineate the KCs of DHS: they are implemented at healthcare system and patient level, can drive multiple clinical outcomes indirectly through a multitude of smaller clinical benefits, their mechanism of action can vary between individuals and change over time based on patient needs, and finally, they develop through short, iterative cycles - optimally within a real-world use context. Following our objective to drive better alignment between clinical evaluation design and the unique traits of DHS, we then provide methodological suggestions that better address these KCs, including tips on mechanism of action mapping, alternative randomization methods, control arm adaptations, and novel endpoint selection, as well as innovative methods utilizing real-world data and platform research.

Introduction

Digital health solutions (DHS) can improve healthcare access, patient equity, operational efficiency, and cost-effectiveness for healthcare organizations – while delivering clinical outcomes for patients. Despite the rapid proliferation of DHS, most lack convincing evidence supporting their clinical impact [1], preventing their uptake within the healthcare system and blocking subsequent development cycles required to realize their full clinical potential. Closing this evidence gap and producing strong evidence in a timely and cost-effective manner is critical for the establishment of trust in DHS and imperative for their adoption and implementation – prerequisites for unlocking digital health's potential to truly transform healthcare [2].

Randomized clinical trials (RCTs) using placebo controls groups, individual randomization, and strict, arguably artificial clinical settings, are the gold standard for evaluating drug efficacy and safety [3]. But contrary to pharmacological treatments, in which efficacy is driven by modulation of biology at the molecular level, DHS belong to a specific interventional class known "complex interventions" [4]. By definition, the impact of this type of intervention is not solely biological, but also encompasses psychological, behavioral, and systems-level effects indicating context-dependency [5]. These attributes may, for example, help explain why effect sizes for pharmaceutical products are generally higher in traditional RCTs compared to research conducted under conditions reflecting real world conditions [6, 7], while DHS effect sizes tend to be higher when using designs that mimic real world conditions [8]. The unique attributes of DHS and the "complex" class of interventions they deliver may therefore necessitate a more tailored approach to their evaluation to measure their true impact.

In this Viewpoint, we discuss our perspective on what must be considered when designing clinical evaluations of DHS. We specifically aimed to (1) delineate a series of key characteristics (**KCs**) of DHS that make them different from pharmacological agents and should be considered while designing their evaluation, and (2) offer methodological solutions to adapt research to the requirements of DHS, addressing randomization, control arm design, endpoint selection, and non-traditional, innovative adaptations.

DHS Key Characteristics

KC1: DHS are often implemented at healthcare system and patient levels. While patients may

generate most of the data, DHS frequently involve other individuals such as health care professionals (HCPs) and caregivers, who must integrate the solutions into their workflows for the DHS to reach full effectiveness. The benefits of using DHS frequently expand beyond patients to caregivers, HCPs, and to the whole healthcare organization. For instance, implementing a remote patient monitoring DHS in a single healthcare center can have permeating systemic effects, with the product improving outcomes of the patients monitored, and those of the other patients either by decreasing re-admission and then shifting clinical staff resources to preventive care, or by improving healthcare efficiency (streamlined workflow, optimized patient engagement strategies…). Within an evaluation setting, HCPs in the control group may adapt their behavior via learning from those included in the intervention group and change their clinical strategies (i.e. leading to contamination). Therefore, it is important to not just control at the patient level, but also at the healthcare site level [9].

KC2: DHS are "complex" interventions. Owing to their multi-user designs and dependencies on real-world healthcare systems, the mechanism of action (MoA) of DHS depend on the historical, situational, environmental, and psychological factors in which treatments are delivered. Contextual dependencies include but are not limited to: HCP characteristics, patient-HCP relationships, perceived intervention credibility, delivery modality, psychological state of the individual, societal, economic, and cultural factors. Even seemingly small nuances, like a patient having previously been asked about their health, can have a significant impact on their health-related behavior and consequently also on health outcomes [10-12]. Many of these contextual factors are unintentionally artificially modified in pharmaceutical RCTs to maximize internal validity. For instance, the Hawthorn effect, which is an increase in engagement when people are observed [13], and other research participation effects (RPEs) [14] are controlled between groups by using placebos. Yet this engagement is in fact a designed benefit of several DHS, therefore adding measurement points which would not occur in real-world situations to a DHS control group can render the control non-inert. For example, measuring and regularly reporting blood pressure improves engagement and hypertension control [15] and similarly, reporting symptoms improves cancer self-management practices [16]. The artificial contexts and measurement conditions traditionally used in pharmaceutical RCTs can therefore limit ecological validity of how the DHS is used. Consequently, conclusions drawn from such designs may lack external validity.

KC3: DHS may drive multiple clinical outcomes <u>indirectly</u> through a multitude of smaller clinical benefits. DHS are frequently "holistic" interventions that provide a broad spectrum of solutions to induce diverse changes, mainly behavioral, which then work as levers to deliver clinically meaningful outcomes. For instance, a remote patient monitoring platform for diabetics that reduces time to treatment adjustment and thereby improves pharmaceutical adherence should consequently improve blood glucose regulation. Similarly, any improvement of sleep behavior by a sleep coaching DHS may also improve cardiovascular risk through better exercise, decreased weight, better eating habits, decreased cholesterol and improvement in glucose regulation. While the magnitude of effect of the DHS on each of these endpoints may be limited, the cumulative effects may lead to a clinically significant benefit. Choosing a single primary clinical endpoint during evaluation therefore may not capture the complexity of the DHS intended purpose nor its true performance.

KC4: The MoA driving clinical outcomes in DHS can a) vary between individuals, and b) change over time based on patient needs. Many DHS interventions include several separate, situationally-activated components, or features (HCP-facing interface, symptom monitoring, activity tracking etc.), that can bring unique benefits. For example, one patient might use Feature A which indirectly improves cholesterol levels while another might prefer Feature B and experience an improvement in blood pressure, with smaller changes observed in cholesterol – while some patients might benefit from both. The effect size of the intervention for a single endpoint may therefore depend on how the product is used. Moreover, the second patient may begin to prefer Feature A, but only later during use. The effect size of the intervention for a single endpoint can therefore be weakened when the DHS as a whole rather than specific feature use is considered as the intervention. What is more,

many DHS employ adaptive algorithms that are specifically designed to adjust the intervention over time. Depending on the DHS being considered, researchers might therefore anticipate that due to user-dependent variability in MoA and emerging synergistic influences (i.e. better product engagement following use-based personalization) clinical impact may require a longer treatment period to be detected.

KC5: DHS develop through short, iterative cycles - often within a real-world use context. Contrary to active pharmacological ingredients that cannot change, DHS continuously evolve in terms of their functionality with each subsequent release. Because DHS must meet their users' needs in the real world, an environment where many factors remain unknown [17], developers often perform accelerated, iterative tests to both discover product performance factors and adapt it before the next release. Due to the number of factors that must be discovered, how they can influence each other, and how influential certain factors can be (i.e. a seemingly small adjustment of written content to an audio format might completely change engagement and therefore clinical outcomes), the solution often must stay moderately malleable as product design hypotheses are validated over a continuum of non-clinical tests. Clinical researchers often opt to evaluate a "frozen" version of the product to optimize homogeneity in the intervention arm, but these approaches may limit external validity, as the marketed solution may drastically evolve over the time and versions tested by the RCT may be outdated at the time of study completion.

Recalibrating research design to meet the scientific canvas of DHS

Considering the KCs described above, alternative assessment strategies are needed to address the unique challenges of DHS evaluation while maintaining internal validity (unbiased study design), external validity (applicability to different contexts), and ecological validity (generalizability to real-world settings). The solutions discussed below are summarized in the Figure 1.

The first step to design DHS evaluation is to take the time to delineate the MoA of the solution and the context. This includes a map of behavioral changes required from patients and all other individuals who are affected by the product (KC1), as well as the contextual environment of use (KC2). Given the limitations of an active control arm described above (KC2), the use of usual care usually named as treatment as usual (TAU), may serve as an appropriate control. TAU corresponds to the actual routine care and may differ from the state-of-the-art guideline-adherent clinical care. Therefore, a prerequisite for this approach would be a detailed understanding of the procedural standard of care which includes, but is not limited to, *local* (i.e. site-specific) procedures, treatment recommendations, information delivered (e.g. health literacy counseling) and frequency of contacts with HCPs. This task is non-trivial, as evidenced by standard of care descriptions being generally underreported, [18] and can pose exceptional challenges due to the number of factors in a real-world environment. Combining patient journeys, DHS' user journeys, and HCP patient management protocols is a helpful first step to identify how the product induces change within its context of use. Process evaluation techniques [19] can also be employed to extract MoAs for complex interventions during preliminary implementation research 5. With the MoA delineated in detail, one can begin designing an appropriate evaluation employing a fit-for-purpose scientific approach.

As many have suggested before [9, 20], cluster randomization by healthcare site, cities or region is a methodologically sound approach to control for contamination between arms while taking into account DHS benefits which depend on factors extending beyond the treated patient (KC1). This design, however, can necessitate larger sample sizes and site matching can be challenging. Pragmatic designs [9], which anchor RCTs in real-world settings and minimize research-related contact outside of TAU, can help maximize ecological validity of both the experimental (+TAU) and TAU arms

(KC2). Keeping the trial de-centralized can minimize research-related contact while increasing trial accessibility and make the research population more representative of the target population (KC2). While DHS, which collect by nature data digitally are particularly suitable for such designs, limitations secondary to the reduced investigators oversight including data quality and outcome assessment persist.

Even with these tools in mind, designing a control arm with high internal validity remains challenging when evaluating DHS. A reference DHS is an appropriate control, but rarely exists. While some have argued a “sham” app (i.e. as defined by the FDA, an app without the “active” features anticipated to drive the MoA [21] offers optimal internal validity in that it is some form of “placebo” technology, this design can pose critical threats to external and ecological validity, particularly ,when the MoA itself is not fully understood. Providing even minimalistic digital support that is nevertheless seemingly endorsed by HCPs (as would be the case in a clinical trial) might still produce a similar, albeit muted, effect as the DHS. For example, a sham app delivering some health literature is not inert as it can improve health literacy and consequently health-related behaviors. The use of sham as a control is therefore challenging unless it is possible to exclude all active ingredients of the MoA including engagement and interaction with the real world from the sham app. On the other hand, the use of TAU as a control can potentially inflate effect size with some interventions (KC2).]The traditional “waitlist” design used in medical device evaluation may also be appropriate, but can inflate effect sizes by decreasing behavioral activation in the control arm (in essence, inducing the opposite of the Hawthorne effect) [22, 23]. Researchers considering having a waitlist TAU as the comparator might use an appropriate **run-in period** before randomization during which both study arms receive TAU to minimize the influence of treatment expectancy effects [24] within the comparator arm, (KC2). Finally, it's important to consider how the nature of the intervention and the control arm may impact the risk of attrition in the study. Beyond minimizing attrition in general, it is crucial to ensure that the attrition rates between the active and control arms is balanced. Significant differences in dropout rates between both arms may lead to variations in the clinical characteristics of remaining patients, potentially biasing the overall analysis.

DHS researchers may also explore alternate evaluation designs. The use of synthetic control arms built using real-world data sources like electronic health records, claims databases, and disease registries may address certain limitations inherent to the conventional control group. However, the creation of these synthetic arms needs to be considered carefully as it poses its own set of challenges, and they are not devoid of their own limitations. In case of products already on the market, epidemiological methods could be used to examine causal inference in product registry data – again minimizing RPEs. These approaches may in some cases be both more cost effective and provide higher quality data on product performance.

Designing a trial which accounts for the holistic impact of DHS is also critical for demonstrating performance and efficacy. Contrary to pharmacological agents that usually improve markedly a single biomarker (HBA1c, CRP…), DHS frequently improve modestly several (physical activity and lipids and weight, etc.) (KC3). Here, rather than examining the effect on a single endpoint, researchers might employ a risk score which better reflects the holistic nature of the intervention. Win ratios are also an elegant way to analyze multiple equally valuable endpoints [25]. Adaptive designs, such as those incorporating sample size re-estimation, hypothesis-adaptive, or response-adaptive designs [26], could also be both powerful and appropriate given the novel nature of many DHS interventions and the different use patterns and emergent interventions that can arise (KC4). Given the dynamic and personalized nature of DHS, evaluation design should also carefully consider anticipated effect size at each time point during planning.

The above can greatly improve the evaluation of DHS clinical performance, but do not address the practical issue of evaluating rapidly and continuously evolving DHS (KC5). While there is no definitive way to address this, a few options exist beyond freezing the product. During evaluation planning, planned future modifications of the product can be scrutinized to determine whether they

constitute major changes to the MoA or intended use and would therefore require protocol modification during study execution. By combining strict product design control practices with a Trials of Intervention Principles approach [27], researchers can aggregate groups of users who may have used product versions with minor design variations not anticipated to change the MoA – allowing the product to continue to organically develop (KC5). Platform trials that incorporate key product design and clinical value hypotheses in master protocols can be leveraged to have the successive versions of the DHS tested and the results of the sub-studies combined to maximize sample sizes, rather than performing several insufficiently powered successive trials, while keeping analyses anchored in clearly documented starting hypotheses and protocols (Figure 2) [20, 28]. Likewise, factorial research designs can be used to investigate efficacy of new features as they are added to the product [29, 30].

Conclusion

DHS need to be rigorously evaluated to fulfil the need of the various stakeholders including HCPs, patients, regulators, and payers. In this Viewpoint we proposed that their unique characteristics and MoA create a critical need for tailored, innovative approaches that move beyond the traditional pharmaceutical RCTs in order to create fit-for-purpose, robust, and scientifically optimized evaluations. While no one-size-fits-all design exists, both researcher and stakeholders should embrace non-traditional methodologies that match the key characteristics of this emerging type of healthcare intervention.

Acknowledgement

Both authors are full time employees of Bayer and equally contributed to the manuscript. We thank Jason Lott for his feedback and comments.

Declaration of interest

None declared

1. Bibliography

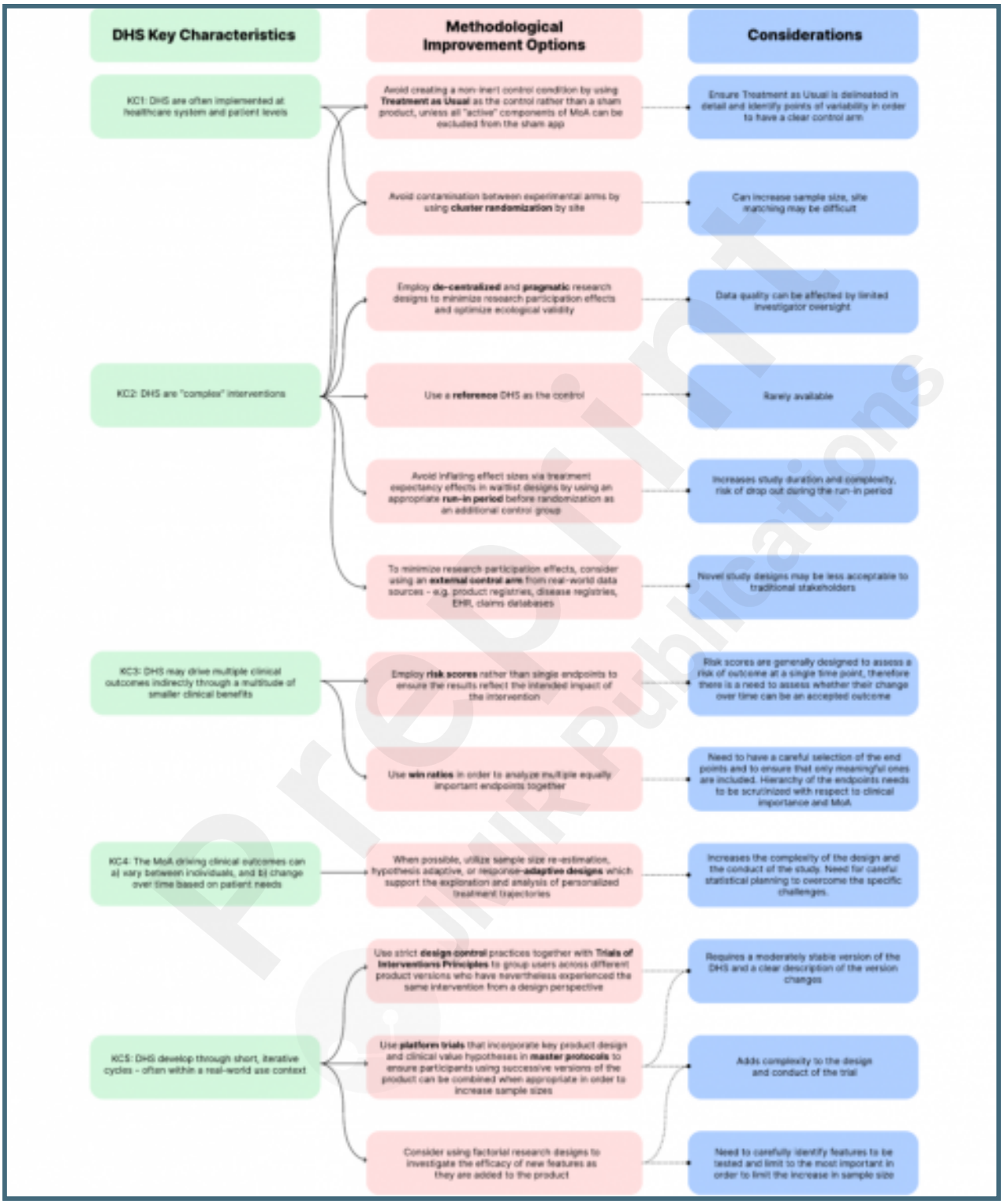
1. Day S, Shah V, Kaganoff S, Powelson S, Mathews SC. Assessing the clinical robustness of digital health startups: cross-sectional observational analysis. Journal of medical Internet research. 2022;24(6e37677.
2. Egermark M, Blasiak A, Remus A, Sapanel Y, Ho D. Overcoming pilotitis in digital medicine at the intersection of data, clinical evidence, and adoption. Advanced Intelligent Systems. 2022;4(92200056.
3. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1986;89(22S-3S.
4. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Elsevier; 2016. p. 843-51.
5. Minary L, Trompette J, Kivits J, Cambon L, Tarquinio C, Alla F. Which design to evaluate complex interventions? Toward a methodological framework through a systematic review. BMC Medical Research Methodology. 2019;19(11-9.
6. Vonbank A, Drexel H, Agewall S, Lewis BS, Dopheide JF, Kjeldsen K, et al. Reasons for disparity in statin adherence rates between clinical trials and real-world observations: a review. European Heart Journal–Cardiovascular Pharmacotherapy. 2018;4(4230-6.
7. Kilcher G, Hummel N, Didden EM, Egger M, Reichenbach S, 4 GWP. Rheumatoid arthritis patients treated in trial and real world settings: comparison of randomized trials with registries. Rheumatology. 2018;57(2354-69.
8. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry. 2017;16(3287-98.
9. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions—A call for innovative evidence generation approaches. NPJ digital medicine. 2020;3(1110.
10. de Bruin M. Risk of bias in randomised controlled trials of health behaviour change interventions: Evidence, practices and challenges. Taylor & Francis; 2015. p. 1-7.
11. Bishop FL, Fenge-Davies AL, Kirby S, Geraghty AW. Context effects and behaviour change techniques in randomised trials: a systematic review using the example of trials to increase adherence to physical activity in musculoskeletal pain. Psychology & health. 2015;30(1104-21.
12. Conner M, Godin G, Norman P, Sheeran P. Using the question-behavior effect to promote disease prevention behaviors: two randomized controlled trials. Health psychology. 2011;30(3300.
13. Berkhout C, Berbra O, Favre J, Collins C, Calafiore M, Peremans L, et al. Defining and evaluating the Hawthorne effect in primary care, a systematic review and meta-analysis. Frontiers in Medicine. 2022;9:1033486.
14. McCambridge J, Kypri K, Elbourne D. Research participation effects: a skeleton in the methodological cupboard. Journal of clinical epidemiology. 2014;67(8845-9.
15. Milani RV, Lavie CJ, Bober RM, Milani AR, Ventura HO. Improving hypertension control and patient engagement using digital tools. The American journal of medicine. 2017;130(114-20.
16. Darley A, Coughlan B, Maguire R, McCann L, Furlong E. A bridge from uncertainty to understanding: The meaning of symptom management digital health technology during cancer treatment. Digital Health. 2023;9:20552076231152163.

17. Smits M, Ludden GD, Verbeek P-P, van Goor H. How Digital Therapeutics Are Urging the Need for a Paradigm Shift: From Evidence-Based Health Care to Evidence-Based Well-being. Interactive Journal of Medical Research. 2022;11(2e39323.
18. Ayling K, Brierley S, Johnson B, Heller S, Eiser C. How standard is standard care? Exploring control group outcomes in behaviour change interventions for young people with type 1 diabetes. Psychology & Health. 2015;30(185-103.
19. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. bmj. 2015;350.
20. Hemkens LG. Nutzenbewertung digitaler Gesundheitsanwendungen–Herausforderungen und Möglichkeiten. Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz. 2021;64(101269.
21. Lutz J, Offidani E, Taraboanta L, Lakhan SE, Campellone TR. Appropriate controls for digital therapeutic clinical trials: a narrative review of control conditions in clinical trials of digital therapeutics (DTx deploying psychosocial, cognitive, or behavioral content. Frontiers in Digital Health. 2022;4:823977.
22. Cunningham JA, Kypri K, McCambridge J. Exploratory randomized controlled trial evaluating the impact of a waiting list control design. BMC medical research methodology. 2013;13(11-7.
23. Bendtsen M, Gunnarsson KU, McCambridge J. Effects of a waiting list control design on alcohol consumption among online help-seekers: protocol for a randomised controlled trial. BMJ open. 2021;11(8e049810.
24. Rutherford BR, Wager TD, Roose SP. Expectancy and the treatment of depression: a review of experimental methodology and effects on patient outcome. Current psychiatry reviews. 2010;6(11-10.
25. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European heart journal. 2012;33(2176-82.
26. Mahajan R, Gupta K. Adaptive design clinical trials: Methodology, challenges and prospect. Indian journal of pharmacology. 2010;42(4201.
27. Mohr DC, Schueller SM, Riley WT, Brown CH, Cuijpers P, Duan N, et al. Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies. Journal of medical Internet research. 2015;17(7e4391.
28. Subbiah V. The next generation of evidence-based medicine. Nature medicine. 2023;29(149-58.
29. Tarricone R, Petracca F, Ciani O, Cucciniello M. Distinguishing features in the assessment of mHealth apps. Expert review of pharmacoeconomics & outcomes research. 2021;21(4521-6.
30. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. Journal of medical Internet research. 2017;19(6e232.

Supplementary Files

Figures

Overview of Key Characteristics (KCs) of the Digital Health Solutions (DHS) and potential methodological solutions to improve their clinical assessment.



Example of design of a Platform trial including 300 patients (150 per arm) enabling the evaluation of multiple versions of a DHS within one study. The figure illustrates the recruitment flow of the patients in the trial. Multiple versions of a DHS can be included in a single study when guidelines for identifying significant changes to study procedures and DHS solutions are delineated in advance. A matched control arm with similar recruitment procedures ensures comparability in the control group.