Checklist of Evidence Quality Criteria for Digital Health Interventions (DHIs)

This checklist (Supplementary Table 2) is designed to supplement established evidence assessment frameworks.

- Group 1 Criteria are those where adaptations to established criteria are recommended, due to differences between digital and non-digital interventions.
- Group 2 Criteria pertain in both digital and non-digital domains, but increased vigilance is encouraged for DHIs in the current regulatory context.

#	Evidence Assessment Criterion	Evidence Criterion Group	Rationale for Inclusion and Notes	✔ Examples Meeting / ✖ Not Meeting Criterion	Recommended Actionability Level (AL) Change if Not Met	Importance
1	DHI assessment is not based solely on association with eminent individuals or institutions.	Group 2. Increased vigilance recommended for DH.	Stakeholders may overvalue a DH solutions provider's association with eminent individuals or institutions. Though relevant, none of the following is a replacement for evidence: • Expert advisors (may have no meaningful role) • University collaboration (DHSPs can pay for this) • KOL endorsements (they may not evaluate DHIs appropriately and often have conflicts of interest) • Endorsement by compensated third parties • Adoption by reputable clients (DHI assessment standards are highly variable, even across reputable organizations)	✔ **Example meeting criterion** High-quality, peer-reviewed evidence shows a mean A1c reduction of 0.7, relative to no change for randomly assigned control participants. ✖ **Example not meeting criterion** A KOL endorses a DHI's effectiveness based on unreviewed, low-quality evidence.	KOL endorsement should not impact AL rating.	Essential

2	Control condition is consistent with evaluator priorities.	Group 1. Adaptations recommended for DH.	Sham controls are designed to blind participants to trial arm assignment and equalize engagement across arms. This approach may allow unconfounded attribution of benefit to a DHI. However, “sham apps” may mask non-specific risks associated with increased smartphone exposure, because smartphone use is equal across arms in trials employing sham apps. Growing evidence suggests that increased smartphone exposure may harm mental health. Usual care (UC) controls (defined elsewhere) receive no treatment from the study. UC controlled trials cannot distinguish specific effects (eg, impact of app-delivered health education) from non-specific effects (eg, impact of taking time to use an app, which may reduce time exposed to stressors). However, UC control conditions should not mask the aforementioned non-specific harms, where they exist. Advantages and disadvantages of other control condition types, including standard of care controls, are reviewed elsewhere. Sham controls may be appropriate for explanatory trials, where DHI safety has been established. UC controls may be appropriate for pragmatic trials, where the goal is to generate evidence that guides real-world decisions, and where some non-specific mechanism of benefit is acceptable.	✔ **Example meeting criterion** A high-quality trial with UC controls showed clinically and statistically significant benefit, and evaluators are comfortable with the possibility of non-specific mechanisms of benefit. ✖ **Example not meeting criterion** A high-quality trial with UC controls showed clinically and statistically significant benefit. Evaluators have stringent standards and want to know that benefits are mediated through specific mechanisms.	Decrease rating by 1 level if control condition is inconsistent with evaluator priorities	Essential for controlled studies

3	Results are not “cherry picked.”	Group 2. Increased vigilance recommended for DH.	DHSPs may “cherry pick” analyses that show atypical effect sizes. Note that a) for any intervention, different patient samples will show different effectiveness levels and b) some patient samples may show meaningful clinical benefit due to sampling error alone, even when the true effect size is zero. Risk for unrepresentativeness increases if studies are retrospective, unregistered, registered after start of enrollment, or small in sample size.	✔ **Example meeting criterion** A preregistered, high-quality trial shows mean hemoglobin A1c reductions of 0.7, relative to no change observed for controls. ✖ **Example not meeting criterion** An unregistered, retrospective analysis excludes 90% of participants and reports robust clinical improvements among those retained.	Decrease rating by 1-2 levels.	Essential
4	Data missingness is addressed appropriately.	Group 2. Increased vigilance recommended for DH.	Substantial data missingness is common in DH. It is often assumed implicitly that data are missing completely at random or at random (MCAR or MAR), even where these assumptions are implausible. This may cause underappreciated bias. Missingness should be handled per best practices detailed elsewhere. It is often appropriate to compare baseline scores by attrition status; meaningful differences rule out MCAR and MAR assumptions. Sensitivity analyses should assess robustness of findings to “worst case” and other degrees of difference between missing and observed data. Note that poor user experiences may cause attrition of all but the most motivated patients. If motivated patients have better outcomes on average, then this attrition pattern would paradoxically skew poorly designed DH products toward favorable per-protocol results.	✔ **Example meeting criterion** A trial reports 10% attrition, with statistically significant differences at baseline between completers and non-completers. However, sensitivity analyses reveal that study conclusions would hold under “worst case” assumptions. ✖ **Example not meeting criterion** A trial reports 40% attrition. No analyses address risk for biased missingness.	Decrease rating by 1-2 levels.	Essential

5	Intention-to-treat (ITT) analyses are reported and any per-protocol (PP) analyses are described as such.	Group 2. Increased vigilance recommended for DH.	ITT analyses should be reported wherever possible. Per-protocol (PP) analyses should be described as such, with reporting on the proportions of ALL enrolled participants who are included in each PP analysis.	✔ **Example meeting criterion** ITT analyses show mean hemoglobin A1c reductions of 0.7, relative to no change observed for controls. ✖ **Example not meeting criterion** Only PP analyses are reported.	Decrease rating by 1-2 levels.	Essential
6	Trials are preregistered (eg, using clinicaltrials.gov) and results are reported publicly.	Group 2. Increased vigilance recommended for DH.	Trials should be registered prior to start of enrollment. Results should be shared within 12 months of trial completion wherever feasible, and should be published in peer-reviewed journals. Registration is not required for some DHI commercialization paths. This can increase publication and reporting bias, reducing replicability of findings. We therefore cannot predict that future DHI deployments will be as effective as reported in unregistered trials. Note that all interventions show distributions of effect sizes across samples. Due to selective reporting and the inconsistency of trial registration in DH, many published DHI effect sizes may represent only the most favorable sliver of the relevant effect size distributions.	✔ **Example meeting criterion** A preregistered, high-quality trial shows mean hemoglobin A1c reductions of 0.7, relative to no change for controls. ✖ **Example not meeting criterion** An unregistered trial shows robust reductions in hemoglobin A1c.	Decrease rating by 1-2 levels.	Essential

7	Conclusions regarding safety or effectiveness are not based on DHSP attestation alone.	Group 2. Increased vigilance recommended for DH.	Some digital health solutions providers (DHSPs) formally self-attest to following best practices, often in collaboration with a trade organization. This may be helpful, but self-attestation is not a substitute for evidence.	✔ **Example meeting criterion** High-quality, peer-reviewed evidence shows a mean reduction in hemoglobin A1c of 0.7, relative to no change for controls. ✖ **Example not meeting criterion** A DHSP signed a self-attestation stating that they follow best practices.	Self-attestations should not impact ratings.	Essential
8	Marketing claims are consistent with peer-reviewed evidence and are not misleading.	Group 2. Increased vigilance recommended for DH.	In the current regulatory context, misleading and evidence-discordant claims are common.	✔ **Example meeting criterion** Reporting in a peer-reviewed article is consistent with marketing claims. ✖ **Example not meeting criterion** A DHSP changes patient-decades (in a peer-reviewed article) to patient-years (in marketing claims) without moving the decimal point, causing claims to be overstated by an order of magnitude.	Decrease rating by 1-2 levels.	Essential

9	Evidence is reported in peer-reviewed journals rather than white papers or other unreviewed materials.	Group 2. Increased vigilance recommended for DH.	Though peer review is often expected, some DHSPs rely on “white papers.” These marketing documents may show levels of rigor and transparency that are inadequate for appropriate evidence assessment. Evidence published in predatory journals (defined elsewhere) is also inadequate.	✔ **Example meeting criterion** High-quality, peer-reviewed evidence shows a mean reduction in hemoglobin A1c of 0.7, relative to no change for controls. ✖ **Example not meeting criterion** An uncontrolled, retrospective analysis for an unreported number of patients shows robust A1c reductions. Evidence is not peer-reviewed, but rather is reported in a white paper.	Non-peer-reviewed evidence can be considered, but alone does not justify any increase in actionability rating. The same is true for evidence published in predatory journals.	Essential
10	Results are clinically and statistically plausible.	Group 2. Increased vigilance recommended for DH.	Implausible reporting does happen in digital health, even in high-impact, peer-reviewed journals.	✔ **Example meeting criterion** All quantitative findings reported are plausible. ✖ **Example not meeting criterion** Reported confidence intervals imply a standard deviation of 45 for hemoglobin A1c, which is implausible clinically.	Decrease rating by 1-2 levels.	Essential
11	It is not assumed that numerous peer-reviewed publications indicate effectiveness or safety.	Group 2. Increased vigilance recommended for DH.	Published editorials may be relevant, but are not a substitute for evidence. High numbers of published, low-quality studies should not be confused with high-quality evidence.	✔ **Example meeting criterion** High-quality, peer-reviewed evidence shows a mean A1c reduction of 0.7, relative to no change in controls. ✖ **Example not meeting criterion** A DHSP published editorials but not clinical evidence.	Peer-reviewed editorials should not impact evidence ratings. Low-quality evidence should not justify ALs greater than 2, even if multiple peer-reviewed articles are available.	Essential

12	Patients who declined to participate are not used as comparators.	Group 2. Increased vigilance recommended for DH.	Patients who enroll in health management programs often differ meaningfully from those who decline to participate. For example, enrollees may have stronger motivation to self-manage chronic conditions. Matching on demographics does not resolve this.	✔ **Example meeting criterion** The rate of acute clinical events for DHI users is 15% lower than that of randomly assigned, waitlisted controls. ✖ **Example not meeting criterion** The rate of acute clinical events for DHI users is 15% lower than that of demographics-matched adults who declined to participate.	Decrease rating by 1-2 levels.	Strongly Preferred
13	Observed clinical improvements are not attributable to healthy user effects or other selection biases.	Group 2. Increased vigilance recommended for DH.	Patients who use health management tools like DHIs may differ from those who do not. DHI users may have stronger health-related motivations and exhibit healthier behaviors. Patients who use DHIs may show improved outcomes over time irrespective of intervention. It should not be assumed that clinical status would be static without intervention. For example, a recent RCT showed a meaningful 10.6 mm Hg reduction in systolic blood pressure for DHI users, but a comparable 10.1 mm Hg reduction for controls. Without a control arm, it would have been easy to misinterpret this as evidence of effectiveness. In some cases it is possible to reduce risk of healthy user (and similar) biases. Investigators should follow best practices, summarized elsewhere, to analyze and interpret data where healthy user (and other) biases may inflate effectiveness estimates.	✔ **Example meeting criterion** High-quality, peer-reviewed evidence shows a mean A1c reduction of 0.7, relative to no change for randomly assigned controls. ✖ **Example not meeting criterion** An uncontrolled study shows a mean A1c reduction of 0.7.	Decrease rating by 1-2 levels.	Strongly Preferred where Relevant

14	Frequency and intensity of interaction with human personnel (eg, mental health professionals or health coaches) has not changed following evidence generation.	Group 1. Adaptations recommended for DH.	As DHI deployment scales up, or as business models evolve, intervention components previously implemented by program staff may be automated. Reducing human interaction may reduce effectiveness in some cases.	✔ **Example meeting criterion** High-quality evidence of efficacy was generated for an automated DHI product version. ✖ **Example not meeting criterion** High-quality, peer-reviewed evidence was generated for a DHI version incorporating video chat with a clinical pharmacist. After a pivotal trial, this intervention component was automated. No post-automation evidence is available.	Decrease rating by 1-2 levels, unless evidence shows noninferiority of an automated DHI product version, relative to a non-automated version.	Strongly Preferred
15	Qualifications of personnel delivering the intervention (eg, mental health professionals or health coaches) remain consistent following evidence generation.	Group 1. Adaptations recommended for DH.	As DHI use scales up, if more personnel are needed, minimum qualification requirements may be relaxed.	✔ **Example meeting criterion** High-quality evidence was generated for a DHI product version after relaxing minimum qualifications required of DHI personnel. ✖ **Example not meeting criterion** High-quality evidence was generated for a DHI incorporating video chat with a clinical pharmacist. Subsequently, pharmacists were replaced with “care coordinators” who do not have clinical training. No evidence is available comparing DHI versions.	Decrease rating by 1 level. An exception should be made if evidence shows noninferiority of a DHI version in which minimum qualifications of program personnel were relaxed.	Strongly Preferred

16	If the target population includes underserved patients, then study samples should have included such patients.	Group 1. Adaptations recommended for DH.	DHIs often require adaptations for underserved patient populations. For example, adaptations may be needed to address varying levels of literacy, health literacy, numeracy, digital literacy, and broadband access.	✔ **Example meeting criterion** An organization is assessing a DHI for use in underserved patient communities. The DHI has shown effectiveness among racial minority subgroups as well as subgroups residing in low-SES zip codes. ✖ **Example not meeting criterion** An organization is assessing a DHI for use in underserved patient communities. Relevant studies investigated high-SES patients only.	Decrease rating by 1-2 levels.	Strongly Preferred

17	Effect sizes are comparable for registered and non-registered trials, if relevant.	Group 1. Adaptations recommended for DH.	This criterion pertains only to DHIs for which evidence has been generated in both registered and unregistered trials. Trial registration (eg, through clinicaltrials.gov) is not required for some commercialization paths. This may increase publication bias and reduce the likelihood of replicating reported effect sizes. We do not expect any two studies to show identical effect sizes. But if effect sizes for registered and unregistered trials differ to a clinically meaningful degree, this may raise concern for publication bias. Consider investigating differences across studies that may explain any effect size inconsistencies. Such differences may relate to DHI versions, implementation protocols, sample characteristics, or sample sizes (small samples increase risk for outlying effect sizes).	✔ **Example meeting criterion** High-quality, peer-reviewed evidence shows mean reductions in hemoglobin A1c of 0.7 and 0.5, both relative to no change observed for controls, in registered and unregistered trials, respectively. ✖ **Example not meeting criterion** High-quality, peer-reviewed evidence shows mean reductions in hemoglobin A1c of 0.7 and 0.1, both relative to no change observed for controls, in registered and unregistered trials, respectively.	Decrease rating by 1 level.	Strongly Preferred where Relevant
18	Distribution of effect sizes does not suggest meaningful uncertainty in average level of benefit.	Group 1. Adaptations recommended for DH.	For some non-digital treatment modalities (eg, drugs), effect size inconsistency may suggest uncertainty in average level of benefit. However, for DHIs, effect size inconsistency may be due to improvements implemented over time. Iterative improvement of DHIs is common and should not cause downgrading of evidence actionability. Some evidence assessment frameworks designed for non-digital interventions recommend reducing ratings if effect sizes differ across studies of similar patient samples. However, if effect sizes for a DHI improve over time, this may reflect product improvements.	✔ **Example meeting criterion** Mean A1c reductions of 0.4 and 0.8 were observed in high-quality studies conducted 4 and 2 years ago, respectively. ✖ **Example not meeting criterion** Mean A1c reductions of 0.8 and 0.4 were observed in high-quality studies conducted 4 and 2 years ago, respectively.	Do not reduce rating if effect size improves over time. Decrease rating by 1-2 levels if unexplained, unfavorable changes in effect size are observed.	Strongly Preferred

19	DHI modifications implemented during and after trials are documented.	Group 1. Adaptations recommended for DH.	DHIs are often improved iteratively, through software updates. Current versions may have clinically meaningful differences from trialed versions. DHSPs should report a) the product version in use at the start of a trial, b) the dates of product updates, and c) the product changes implemented with each update.	✔ **Example meeting criterion** Software versions used during and after a trial are reported in a public website. A summary of each update is provided. ✖ **Example not meeting criterion** Software versioning information is not reported.	Evaluators should be aware of this criterion, though AL adjustment may not be needed.	Preferred
20	Onboarding for trial participants is comparable to onboarding in real-world deployments.	Group 2. Increased vigilance recommended for DH.	Lengthy onboarding assessments are common in digital health trials. This may select for more motivated participants, on average. After DHI deployment, when enrollment may take precedence over evaluation, onboarding burden may be reduced substantially. On average, this may select for less motivated patients. Thus, average clinical benefit may be lower in real-world use.	✔ **Example meeting criterion** Onboarding assessments in a previous trial are the same as those used following real-world deployment. ✖ **Example not meeting criterion** Ten baseline PROs were administered in a pivotal trial, while two are administered following real-world deployment.	Evaluators should be aware of this criterion, though AL adjustment may not be needed.	Preferred
21	Participant incentives are comparable in trials and real-world deployments.	Group 2. Increased vigilance recommended for DH.	Trial participants may receive payment or other incentives for enrolling and meeting engagement targets. Changes in incentives between trials and subsequent real-world deployments may change who enrolls and how much they engage. Stakeholders should be aware that this common scenario may decrease external validity of trial evidence.	✔ **Example meeting criterion** The same incentives are provided in a pivotal trial and a real-world deployment. ✖ **Example not meeting criterion** Trial participants are paid for high engagement; this payment is reduced in real-world DHI deployment.	Evaluators should be aware of this criterion, though AL adjustment may not be needed.	Preferred

Abbreviations: DH, Digital Health; DHIs, Digital Health Interventions; DHSPs, Digital Health Solutions Providers; AL, Actionability Level; KOL, Key Opinion Leader; MAR, Missing at Random; MCAR, Missing Completely at Random; PROs, Patient-Reported Outcomes; RCT, Randomized Controlled Trial; SES, Socioeconomic Status; UC, Usual Care