

PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research



Article submitted to journal

Subject Areas:

Digital health, reproducibility, AI

Keywords:

Digital health, reproducibility, AI

Author for correspondence:

Susobhan Ghosh

e-mail:

susobhan_ghosh@g.harvard.edu

Reproducible workflow for online AI in digital health

Susobhan Ghosh^{1*}, Bhanu T. Gullapalli^{1*},
 Daiqi Gao², Asim Gazi¹, Anna Trella¹,
 Ziping Xu², Kelly Zhang³,
 Susan A. Murphy^{1,2}

¹Department of Computer Science, Harvard University, Cambridge, MA, USA, 02135

²Department of Statistics, Harvard University, Cambridge, MA, USA, 02135

³Department of Mathematics, Imperial College London, London, UK

*Equal contribution

Online artificial intelligence (AI) algorithms are an important component of digital health interventions. These online algorithms are designed to continually learn and improve their performance as streaming data is collected on individuals. Deploying online AI presents a key challenge: balancing adaptability of online AI with reproducibility. Online AI in digital interventions is a rapidly evolving area, driven by advances in algorithms, sensors, software, and devices. Digital health intervention development and deployment is a continuous process, where implementation—including the AI decision-making algorithm—is interspersed with cycles of re-development and optimization. Each deployment informs the next, making iterative deployment a defining characteristic of this field. This iterative nature underscores the importance of reproducibility: data collected across deployments must be accurately stored to have scientific utility, algorithm behavior must be auditable, and results must be comparable over time to facilitate scientific discovery and trustworthy refinement. This paper proposes a reproducible scientific workflow for developing, deploying, and analyzing online AI decision-making algorithms in digital health interventions. Grounded in practical experience from multiple real-world deployments, this workflow addresses key challenges to reproducibility across all phases of the online AI algorithm development life-cycle.

© The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, provided the original author and source are credited.

1. Introduction

Digital health interventions – or simply *digital interventions* – present an opportunity to address substantial health challenges. Leveraging ubiquitous technologies such as smartphones and wearable devices, these interventions are delivered directly to individuals as they go about their daily life. For example, individuals may receive reminders to take prescribed medication, prompts encouraging physical activity after long periods of inactivity, or supportive messages during periods of elevated stress. These actions—delivered through mobile notifications, SMS, or app-based interfaces—aim to promote healthy behaviors in their natural environment. Digital interventions are now used across a wide range of health contexts, including depression [1,2], medication adherence [3], substance use [4–7], physical health [8,9], cardiovascular disease [10], cancer recovery [11], and many more.

Because the field of digital interventions is an evolving area, driven by rapid advances in software, sensors and wearable devices, the field is inherently dynamic, driven by a process of *continuous improvement* — that involves designing, testing, refining, and re-deploying interventions. Data collected during each deployment plays a dual role: (1) it supports scientific discovery by enabling the evaluation of intervention efficacy and the analysis of individual behaviors, and (2) it informs the (re)development of the intervention and any associated algorithm(s). Here, we use the term *deployment* to refer to the release and operation of a digital intervention in a real-world environment. This process of continuous improvement—driven by successive deployments and refinements is fundamental to how digital intervention research and practice evolve, enabling interventions to respond to not only changes in operating systems, sensors and wearable devices, but also real-world societal changes and individual diversity .

Challenges in this continuous improvement process are many. Models or insights developed in one deployment may fail to transfer effectively to subsequent ones, due to changes in population characteristics, social dynamics, technological platforms, intervention delivery mechanisms and environmental contexts. Second, even within a single deployment, individuals differ widely in their needs, preferences and circumstances affecting their engagement with and response to interventions. To address these challenges, digital interventions have begun utilizing *online* artificial intelligence (AI) decision-making algorithms [12–14]. Here, the term *online* refer to AI decision-making algorithms that leverage streaming within and between individual data—such as passive sensing and self-reports— to learn parameters in a model for health outcomes and update estimates of these parameters repeatedly as new data becomes available during deployment, in contrast to *offline* algorithms, that learn model parameters on historical data which then remain fixed throughout deployment. Online AI decision making algorithms make use of this online learning to make decisions regarding which and when to deliver interventions during deployment.

Unlike traditional data collection settings, digital interventions employing online AI decision-making algorithms operate within an adaptive feedback loop: past data informs the algorithm's current intervention decisions—such as sending a prompt or a message—which in turn influence individual biobehavioral outcomes and thus the data collected in the future (see Fig 1) [15]. This dynamic interdependence introduces tightly coupled system behaviors, where seemingly minor changes can cascade into substantial shifts in data distributions, individual engagement, and intervention outcomes. As a result, the need for reproducibility becomes especially acute in such systems.

The continuous improvement of digital intervention deployments relies on the ability to learn from past deployments – using collected data and observed outcomes to refine interventions and algorithms. For this process to support cumulative scientific progress and continuous data-driven improvements, the intervention system must produce data that is *reproducible*. This paper follows the National Academies of Science, Engineering, and Medicine Consensus Report definition of reproducibility as the ability to obtain consistent results using the same input data, computational steps, methods, and conditions of analysis [16]. Critically, reproducibility is a prerequisite for

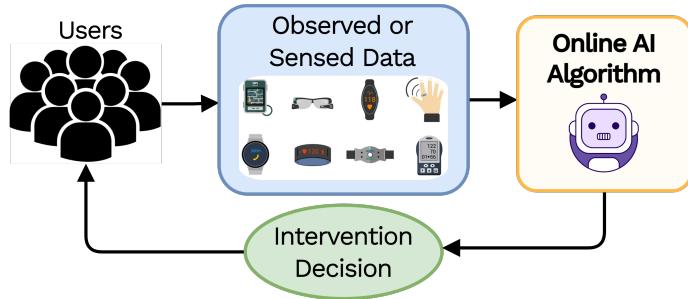


Figure 1. Adaptive feedback loop of an online AI decision-making algorithm. Data collected during deployment informs the algorithm's intervention decisions—such as sending a prompt, a supportive message, or a behavioral recommendation—which in turn influence the individual's future behavior and the data subsequently collected.

replicability, which is the ability to achieve consistent results across deployments aimed at answering the same scientific question but using independent data and potentially in a different implementation setting. While the National Academies definition provides a broad foundation, reproducibility in online AI digital interventions is multifaceted. This paper distinguishes three complementary forms of reproducibility in online AI for digital health: *design reproducibility*, *deployment reproducibility*, and *inference reproducibility*.

- *Design reproducibility* : ability to rerun algorithm development and evaluation under identical conditions—for instance, replicating results in a digital twin testbed given the same input data, computational steps, and random seeds. It can be measured by verifying whether identical simulations and training pipelines yield identical outputs.
- *Deployment reproducibility* : ability to reproduce system behavior during real-world operation—for instance, ensuring that identical inputs and model states yield identical decisions at runtime. It can be measured by verifying that decisions are fully traceable and auditable through logs, version control, and monitoring systems, including replaying logged states to confirm identical outputs or stress-testing monitoring systems to ensure that decision failures are consistently captured
- *Inference reproducibility* : ability to obtain consistent scientific findings from deployment data, including analyses of algorithm performance and behavioral outcomes.

These terms are introduced here for completeness, but the remainder of this paper focuses primarily on design and deployment reproducibility, as inference reproducibility with regards to inference based on analyses of existing data has been addressed in some detail in the literature, for examples, see [17,18]. Reproducibility in the setting of online AI in digital health serves two complementary aims. First, it underpins *scientific trustworthiness*, ensuring that results generated during deployment can be interpreted and verified. Second, it supports *continuous improvement*, providing a stable and auditable foundation on which algorithms and interventions can be iteratively refined across deployments. These aims are aligned but not identical, and distinguishing between them clarifies how reproducibility contributes both to scientific validity and to system evolution.

This paper argues that reliable and stable infrastructure facilitates reproducibility. Reliability ensures that system behaviors remain consistent across contexts and time, thereby enabling scientific findings to be meaningfully interpreted and independently validated. Instability – understood as the breakdown of reliability – can be broadly categorized into two types: (1) engineering instability, stemming from issues such as inconsistent codebases, misaligned execution environments, or deployment mismatches; and (2) algorithmic instability, arising from learning procedures lacking practical convergence guarantees or abrupt changes to model

logic. The adaptive feedback loop present in online AI decision-making algorithms can further exacerbate issues caused by both forms of instability, making the same recorded experimental inputs yield different and often unexpected outcomes at different times — ultimately challenging reproducibility and the iterative development cycle of online AI in digital interventions. Note that instability is not synonymous with uncertainty. In online settings, explicitly accounting for uncertainty—such as drift in the distribution of health outcomes and variability in individual engagement and outcomes —may yield less precise estimates, but conclusions that are ultimately more reproducible, trustworthy, and generalizable.

To address these challenges, this paper proposes a reproducible workflow tailored for researchers and developers designing online AI digital interventions. The workflow provides structured guidance for building systems in which data and outcomes from each deployment can meaningfully contribute to ongoing scientific learning and system refinement. It is designed to help developers anticipate and manage the sources of instability identified above—ensuring that their systems not only function reliably in real time, but also support long-term, cumulative learning. The remainder of the paper is organized as follows: Section 2 provides background on the structure of online AI algorithms and the digital interventions in which they are deployed. Section 3 presents the proposed workflow, organized around key stages in the digital intervention system lifecycle. Section 4 offers practical recommendations drawn from real-world deployments, and Section 5 discusses broader implications, limitations, and future directions.

2. Background

This section describes the underlying components of online AI decision-making algorithms, along with examples of digital intervention involving AI algorithms that were deployed by the authors. The section begins by outlining two core components that are common across many such algorithms, followed by brief descriptions of real-world digital interventions that serve as running examples throughout the paper and ground the workflow in practical deployment contexts.

(a) Components of Online AI algorithms

Online AI decision-making algorithms deployed in digital health interventions are typically composed of two interdependent components:

- (i) **Online learning / Update procedures:** This component governs how the algorithm learns and adapts over time based on new data. It is responsible for updating the parameters of the algorithm's model using information collected during deployment. The updates may occur continuously, at fixed intervals, or in response to specific triggers (e.g., user events or contextual changes), and are often subject to constraints on computation, memory, or communication.
- (ii) **Optimization / Decision-making:** This component determines, in real time, what intervention to deliver given the current context (e.g. data from sensors, self-report data, etc.) and the algorithm's learned model parameters. It aims to optimize a predefined utility function – such as maximizing individual engagement or behavior change – while satisfying practical and ethical constraints on intervention burden and safety.

(b) Reference Interventions

Throughout this paper, insights and examples are drawn from several real-world digital health interventions that the authors have been involved in. Each of these interventions deployed an online AI decision-making algorithm. These interventions will provide context for the workflow and recommendations discussed in later sections. Examples were specifically selected to illustrate both the successes and failures of the authors in designing and deploying an online AI algorithm

as part of a digital intervention evaluated in a clinical trial: HeartSteps (ClinicalTrials.gov ID: NCT03225521), Oralytics (NCT05624489), and MiWaves (NCT05824754). In each example below, the goal of the intervention, the specific decision task assigned to the online AI—implemented as a reinforcement learning (RL) agent—and the contextual factors used to inform that decision are highlighted.

HeartSteps The HeartSteps digital intervention [19–22] is designed to promote physical activity by delivering contextually tailored activity suggestions through smartphones and wearables. HeartSteps integrates a commercially available fitness tracker with a mobile application to sense individuals' behavior and context. HeartSteps uses an online RL [8] to decide, five times daily, whether or not to deliver an activity suggestion based on individuals' current context including prior intervention dose and physical activity.

Oralytics The Oralytics digital intervention [23] is designed to improve oral self-care behaviors in individuals at risk of dental disease. Oralytics integrates a commercially available electric toothbrush with Bluetooth connectivity and a mobile application to sense individuals' brushing behavior. An online RL algorithm [24,25] is utilized to determine, twice-daily, whether or not to deliver an intervention prompt based on recent brushing behavior, date and time, and recent engagement.

MiWaves The MiWaves digital intervention [26] is designed to help emerging at-risk adults reduce their cannabis use. MiWaves uses a mobile application to collect individuals' self-monitoring information. MiWaves uses an online RL algorithm [4,27] to determine, twice daily, whether or not to deliver an intervention prompt containing behavior change strategies based on recent cannabis use, time of day and recent engagement in the behavior change strategies.

3. Workflow for Online AI in Digital Health

This section introduces an end-to-end workflow for developing, testing, deploying, and analyzing online AI-based decision-making algorithms in digital health, with reproducibility as a central design goal. The workflow is intended for researchers and practitioners building digital interventions that incorporate online AI, and who aim to ensure that the data, decisions, and outcomes generated during deployment can support scientific learning and continuous improvement. Illustrated in Figure 2, the workflow is structured into three key phases: (1) Designing the online AI decision-making algorithm, (2) Integration, assurance, and deployment in real-world environments, and (3) Post-deployment data analysis and inference.

Phase 1: Designing the online AI decision-making algorithm. Reproducibility considerations begin during the algorithm design phase, where *past deployment data and inference* inform two key activities: (1) *generating candidate decision-making algorithms*, and (2) *evaluating the candidate decision-making algorithms in digital twin testbeds*.

The candidate algorithms must account for real-world deployment constraints such as latency, bandwidth, and computational resources. These constraints influence not only how algorithms perform in deployment, but also what aspects of their behavior can be reliably reproduced. For instance, the interventions discussed in this paper (HeartSteps, Oralytics, and MiWaves) executed most of the online AI algorithm logic in the cloud, enabling greater control over computation, logging, and versioning. In contrast, interventions that rely on on-device computation and storage [28,29] must contend with stricter resource constraints and increased variability in execution environments – such as limited processing power, memory, and storage capacity on wearables and smartphones. These limitations can directly impact reproducibility. For example when storage is limited, devices may be unable to store all intermediate computation outputs, or full decision logs. This may force system designers to incorporate lossy compression or selective

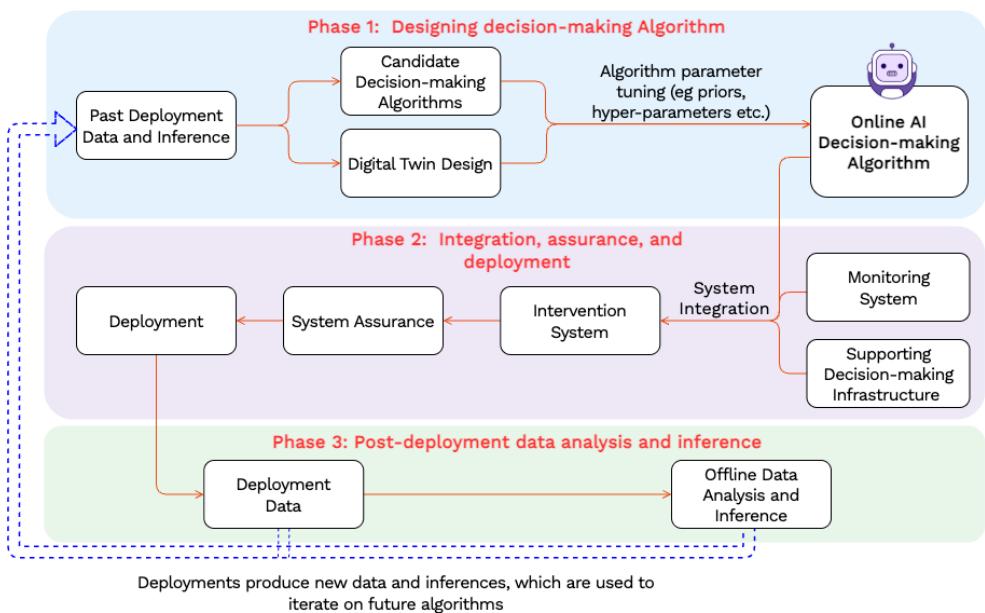


Figure 2. End-to-end reproducible workflow for online AI decision-making algorithms in digital health interventions. The workflow consists of three key phases: (1) algorithm design informed by past deployment data and digital twin(s), (2) system integration, assurance, and deployment with supporting infrastructure and monitoring systems, and (3) post-deployment data analysis and inference to evaluate performance and guide future algorithm and intervention iterations.

storage of only the highest priority outputs for reproducibility, making it impossible to perfectly reconstruct the algorithm's behavior post hoc.

To ensure that candidate algorithms are robust and reproducible under such real-world conditions, their performance must be thoroughly evaluated prior to deployment. The *Digital twin* plays a central role in this process. It is constructed using historical deployment data to replicate key features of individual's behavior, contextual dynamics, and intervention responses [30]. By grounding the digital twin in *empirical data distributions* – such as engagement patterns, outcome trajectories, or contextual variable transitions – researchers can construct a high-fidelity digital approximation of the deployment setting. Within this twin, one can instantiate multiple *simulation environments* by systematically varying key parameters that represent real-world uncertainties. These include changes in the magnitude or variability of treatment effects, levels of individuals' heterogeneity, degrees of contextual drift, and patterns of data missingness or latency. Each combination of these factors defines a distinct environment, reflecting a scenario that domain experts and researchers might reasonably expect to encounter in deployment. These environments serve as a low-cost, low-stakes, and highly controllable sandbox, allowing algorithm designers to rapidly prototype, compare, and refine candidate algorithms without risking harm to individuals or incurring the overhead of live deployments.

Following the construction of the digital twin, the parameters of each candidate decision-making algorithm are tuned to optimize performance across multiple simulation environments in the testbed. This tuning process may include incorporating priors derived from past deployment data, performing hyper-parameter searches, etc. Once tuned, the candidate algorithms undergo rigorous evaluation across multiple simulation environments within the testbed. This evaluation enables the triage and selection of the most promising candidate(s) based on predefined criteria. In many cases, this includes multi-objective optimization, balancing factors like performance, fairness, and stability. The selected algorithm is then advanced to the next phase of the workflow: integration, testing and deployment. Taken together, these activities illustrate how Phase 1

contribute to *design reproducibility*: by enabling algorithm development and evaluation to be repeated under identical conditions, such as rerunning digital twin simulations with the same input data, computational steps, and random seeds.

Phase 2: Integration, Assurance, and Deployment Once an online AI-based decision-making algorithm has been finalized, the focus shifts to embedding the algorithm within an *intervention delivery system*. This integration phase is critical for enabling the online AI algorithm to interact with real-world individuals and data in real time.

The first step in this phase involves connecting the online AI decision-making algorithm to the broader *supporting decision-making infrastructure*. This infrastructure typically includes three major components: data collection components (e.g., sensors and individual inputs), user interfaces, and the intervention backend. *Data collection* components include both data from passive sensors – from wearable devices and smartphones (e.g., GPS, accelerometer) – and active inputs by individuals that capture self-reported mood, stress, cravings, etc. These diverse data streams may provide the contextual data necessary for the online AI-based decision-making algorithm to function effectively. *User interfaces* may include applications through which the individual can interact with the intervention. Specifically in mobile health interventions, well-designed interfaces can be critical for individual engagement, adherence, and ultimately the success of the intervention [31,32]. Finally, the *intervention backend* coordinates all computational and logistical operations. This may include pre-processing raw sensor or self-report data into structured features, managing decision scheduling logic (e.g., determining when the algorithm's decision-making capabilities or update procedures should be triggered), and routing the assigned interventions to the individual's interface for delivery. Additionally, digital intervention systems may also incorporate a dedicated *monitoring system*. A *monitoring system* is a set of tools and processes used to track critical components of the digital intervention. Monitoring is essential in digital health applications, where individual safety and system reliability are paramount. A monitoring system typically logs decisions, system states, and errors, and may include alerting mechanisms to flag abnormal software behavior or intervention delivery failures [33,34].

Once the integration is complete, the intervention system undergoes extensive *quality assurance* testing to ensure safe and reliable operation under realistic conditions. Testing may begin with automated simulations, where artificial inputs are streamed into the system to verify data flow, decision logic, and intervention delivery. This allows developers to identify and fix issues in a controlled environment before involving real individuals. The system may then progress to beta testing or pilot deployments, where a small group of individuals interacts with the intervention under supervised conditions. These tests help verify correct data processing, validate algorithmic responses to real-time contexts, and reveal potential issues such as message delays, sensor dropout, or latency. Following successful quality assurance testing, the intervention system incorporating the online AI decision-making algorithm is ready for real-world deployment. Taken together, these activities contribute to *deployment reproducibility*: ensuring that identical inputs and model states lead to the same intervention decisions, and that all system behaviors can be reconstructed and audited through comprehensive logging, monitoring, and version control.

Phase 3: Post-Deployment Data Analysis and Inference Subsequent sections do not address the reproducibility of phase 3 – from deployment data pre-processing to offline analysis and inference – as this topic has been extensively explored in prior work [17,18,35] and falls outside the scope of this paper. However for completeness a brief overview of this phase is provided.

After deployment is completed, the intervention system (incorporating the online AI decision-making algorithm) generates a stream of data—including contextual variables, algorithmic decisions, and observed outcomes—which is processed and analyzed offline. This analysis plays a critical role in interpreting system behavior and guiding future development. Specifically, this phase serves three primary purposes: (1) to evaluate algorithm performance and behavior, and (2) to guide future iterations of the algorithm and the intervention system , and (3) to

support scientific discovery, including investigations into mechanisms of behavior change. Offline analysis leverages the deployment data—including contextual variables, algorithmic decisions, and observed outcomes—to answer a range of scientific and engineering questions. Depending on the evaluation goals, this analysis may involve statistical, causal inference, or machine learning methods to assess whether the algorithm achieved its intended outcomes [25,36], identify unintended consequences, or examine variation in effectiveness across subpopulations [37]. Techniques such as counterfactual modeling, heterogeneous treatment effect estimation, and mediation analysis are commonly used to support these objectives [38,39]. In addition, algorithmic stability is often evaluated to determine how sensitive the system’s decisions are to small changes in input data or modeling assumptions [40], which is particularly important in high-dimensional or noisy settings.

Beyond evaluation, the collected data is used to refine future decision-making algorithms and intervention systems. This can involve retraining on new or more diverse datasets, updating model parameters to reflect the newly observed data distributions, or generating new candidate algorithms informed by recent deployment outcomes. Insights from offline analysis may also be used to construct more representative simulation environments, thereby improving the fidelity and robustness of subsequent simulation testbeds (Phase 1).

Overall, this phase closes the loop in the online AI algorithm development workflow by feeding empirical evidence from real-world deployment back into the design phase for subsequent deployments. This iterative process enables data-driven refinement of digital health interventions, supporting both immediate enhancements and long-term system evolution. Although detailed discussion is beyond the scope of this paper, Phase 3 directly involves *inference reproducibility*, to ensure that analyses of deployment data can be interpreted and contribute to cumulative scientific progress.

Finally, the above workflow can also be understood within the PCS framework for veridical data science [17,18], which emphasizes Predictability, Computability, and Stability as foundations for trustworthy analysis. In the digital intervention setting careful construction of digital twin environments, enactment of pilot studies, and use of monitoring systems can contribute to *P* (*Predictability/Reality check*), while practices such as version control, logging, and stability assessments can contribute to *S* (*Stability*) by making decision-making reproducible and robust to perturbations in data and context. Considerations of infrastructure, compute resource management, and system integration can contribute to *C* (*Computability*). Here, recent work on simulation design under the MERITS framework [41] provides additional practical guidance, particularly for the construction of digital twins that are modular, efficient, realistic, intuitive, transparent, and stable. Framing the above workflow through PCS highlights its alignment with broader principles of veridical data science. An illustrative case study of a sleep intervention with wearable sensors, showing how each phase of the workflow can be implemented in practice, is provided in the Supplementary material.

4. Recommendations

While the workflow in Section 3 outlines how online AI decision-making algorithms can be developed and deployed in digital health settings, ensuring reproducibility in practice requires careful attention to system design, logging, and deployment procedures. This section offers concrete recommendations for improving reproducibility and reliability across each phase of the workflow. These recommendations are grounded in lessons from real-world deployments – including HeartSteps, Oralytics, and MiWaves, described in Section 2(b) – and address common failure modes and engineering challenges encountered in practice.

(a) Software Environments and Isolation

Ensure workflow and system component isolation

Ensure strict separation between development, testing, and deployment environments, and isolate system components (e.g., using containers such as Docker) to avoid interference and ensure stable operation.

A *software environment* refers to the complete configuration – software libraries, operating system, dependencies, and runtime settings – within which an AI algorithm or digital health intervention system operates. In the context of online AI in digital interventions, this includes both the environment in which algorithmic decisions are made and the broader system that supports data collection, processing, and intervention delivery. *To ensure reliability in the intervention system, it is essential to maintain a clear separation between development, testing, and deployment software environments.* This separation ensures that modifications or instability during the development process do not accidentally propagate into deployment, and that testing reflects deployment conditions as closely as possible. For instance, suppose developers are working on a new version of the decision-making algorithm in the development environment while an ongoing deployment is actively running. If the environments are not clearly separated—for example, if both versions share a codebase or infrastructure—a partially tested update from development environment might inadvertently be pushed into the current deployment. Imagine that this update introduces a bug that causes the algorithm to skip generating interventions under certain conditions (e.g., when recent sensor data is missing). This could lead to reduced intervention delivery, compromising both individual outcomes and the integrity of the collected data. Moreover, if the deployment logs do not reflect that a new version was unintentionally introduced, researchers analyzing the data post hoc may be unable to reconcile the algorithm’s behavior with the recorded system state, ultimately undermining reproducibility.

Beyond this vertical separation across workflow stages, *a second form of separation – component-level isolation within the deployment software environment – is crucial for reliable functioning of online AI systems* during Phase 2: Integration, Testing, and Deployment. The digital intervention system typically comprises of multiple components running concurrently: the online AI decision-making algorithm, backend servers, databases, sensors, and user interfaces. Some of these components may be co-located on shared infrastructure (e.g., a cloud machine or mobile device), leading to potential interference through shared memory or processing resources. *Isolation*, in this context, refers to the runtime separation of these components to prevent one from negatively impacting another’s performance. Without isolation, failures or delays in one component can cascade into system-wide instability. For example, in the pilot study for MiWaves [4,26,27], both the algorithm and the intervention backend were deployed on the same machine without isolation. Because they shared memory and CPU resources, the system experienced memory management issues that eventually caused both components to crash – negatively impacting the intervention delivery. While the developers involved in MiWaves systematically tracked these crashes with the help of a monitoring system (Section 4(f)), such instability can easily go unnoticed and untracked. If not recorded, these issues can silently alter the data being collected during deployment, impacting the scientific validity and reproducibility of data. Ensuring such environment and component isolation is central to *deployment reproducibility*, as it guarantees that system behavior can be replicated across environments, and any decisions made in deployment can be traced to a consistent and auditable execution context. This type of instability can be effectively avoided by utilizing containerization and orchestration tools—such as containers managed with Docker, and orchestration frameworks like Kubernetes [42,43]. These tools are widely used to isolate critical system components, by allocating dedicated resources to each component, and enforcing strict software dependency versions independently.

(b) Autonomous Algorithm Design

Design algorithms for autonomous operation

Design algorithms to run autonomously with pre-specified update rules and stability guarantees, avoiding manual modifications during deployment to preserve scientific integrity.

In clinical trials for digital interventions, online AI decision-making algorithms are considered part of the intervention itself. As such, they are subject to constraints on modifiability – particularly in clinical trials, where the intervention logic must be pre-specified in the study protocol. Once a trial has been launched, any changes to the algorithm’s behavior or logic are typically prohibited to maintain scientific integrity [44]. To support this requirement, algorithms must be *autonomous*, meaning they operate independently once deployed, without requiring manual intervention or modifications during the study period. In practice, this autonomy necessitates the use of computational procedures that provide *stability guarantees*—ensuring that the algorithm’s behavior remains controlled, interpretable, and robust to new data throughout deployment. Algorithms must be designed (in Phase 1) such that they are capable of learning online using well-defined update rules, without requiring mid-deployment calibration. This autonomy also directly supports *deployment reproducibility*: by ensuring that the algorithm runs uninterrupted and unmodified throughout the trial, the system’s behavior remains consistent with what was pre-specified in the study protocol. Any observed outcomes can thus be reliably attributed to a fixed, well-documented process, rather than to ad hoc or undocumented adjustments.

For example, during the MiWaves pilot study, the algorithm designers utilized a simple Bayesian learning setup to derive closed-form posterior updates for model parameters, enabling stable online learning and avoiding convergence issues associated with more complex Bayesian inference techniques. Future work could use approximate inference methods like approximate Bayesian computation, for settings where closed-form posteriors are unavailable [45,46]. These approaches offer a promising direction for autonomous online learning with tractable approximation guarantees.

(c) Data Storage

Log all data for full algorithm traceability

Log all inputs, outputs, model states, update steps, and sources of randomness during deployment to enable exact post-hoc reconstruction of algorithm behavior.

To ensure reproducibility in online AI decision-making algorithms for digital interventions, it is essential to systematically store all data used in the decision-making and update process. This includes not only self-reported data or raw sensor streams, but also any processed features or summary statistics if those are the actual inputs to the algorithm. In addition, intermediate outputs – such as decisions made, observed outcomes, model states, and update parameters should be logged. Each data point, particularly those from smartphones or wearables, should be logged with its local device timestamp. Relying solely on backend timestamps can introduce inaccuracies due to communication delays, which may compromise reproducibility. Minor clock drift can typically be corrected post hoc if local timestamps are available.

When storing the entirety of reproducibility-critical raw data in its original form may not be feasible due to local storage constraints (e.g., sensor data sampled on the order of 10 Hz to 100 kHz), it is critical to thoughtfully employ the combination of lossless compression techniques, local or on-device storage (e.g., micro secure digital (SD) card), and asynchronous data synchronization with cloud storage. Without local storage, extended connectivity outages can lead to irreversible data loss, posing a serious risk to reproducibility. Compression can help

reduce storage needs, but decompression procedures must be retained to allow accurate data reconstruction.

When raw data cannot be transmitted in real time due to bandwidth or power limitations, parts of the online AI algorithm – such as decision logic or model updates – may need to run on local devices capable of computation near the data source, often referred to as edge devices (e.g., a smartwatch or smartphone). These devices should locally store input summaries, decisions made, and model state, then transmit them asynchronously to the cloud when possible. To manage local storage, data should only be overwritten after successful cloud upload. In practice, uploads from such devices are best scheduled during charging periods or when the device is connected to Wi-Fi, to minimize power consumption and avoid transmission failures. This ensures that critical data are retained locally until reliable upload is possible, supporting reproducibility even in settings with intermittent connectivity.

One example that illustrates many of these challenges is the photoplethysmogram (PPG) signal, a common high-frequency input used in digital health interventions to estimate heart rate variability. The raw PPG signal reflects changes in peripheral blood volume, but it must be filtered, quality-checked, and segmented to extract inter-beat intervals. These pre-processing steps – such as filter parameters, quality thresholds, and artifact removal procedures – must be well-documented and consistently applied. Without this, variability in processing pipelines can produce inconsistent results that reflect software differences rather than underlying physiology.

Accurate timestamping is also critical. If local timestamps are not recorded at the device level, transmission delays can cause sampling intervals to appear irregular, compromising the integrity of the signal and violating reconstruction assumptions (e.g., Nyquist criteria). Inconsistent sampling or missing data due to connectivity loss may result in gaps or distortions that alter algorithm behavior. To avoid this, devices should retain data locally with high-resolution timestamps and only overwrite after confirming successful cloud upload. When connectivity is unavailable and the device cannot temporarily store data (i.e., lacks buffering support), permanent data loss can occur. In such cases, even recreating the same input context may yield divergent outcomes due to inconsistent data availability.

If storing raw data is infeasible, then the summaries or features used in decision-making must be logged, along with the code used to generate, compress, and decompress them. For probabilistic algorithms, additional elements – such as random seeds, model parameters, and update timestamps – must be recorded to enable exact reproduction of algorithm behavior and support post-deployment analysis.

Finally, it is essential to validate that algorithm updates are reproducible from stored data. Before deployment, researchers should confirm that the online learning pipeline produces consistent outputs from logged inputs. After deployment, these updates should be recreated offline to verify alignment. For example, in the Oralytics [24,25] and MiWaves [4,26] study, model parameters were reconstructed after each update and matched against deployment logs to confirm reproducibility throughout the trial. These validation procedures reinforce *deployment reproducibility*, by ensuring that online decisions can be exactly replayed from logged states.

(d) Delayed and Missing Data for Decision Making

Preserve real-time decision data

Preserve the exact data version used in real-time decisions, including imputed values, and never overwrite with delayed ground-truth data to ensure reproducibility.

Online AI decision-making algorithms deployed in digital interventions often utilize streaming data from sensors and individual inputs for decision-making. However, these data streams can be subjected to delays, dropouts or corruption during real-world deployments, which can impact algorithmic performance and compromise reproducibility if not properly accounted for. For example, in the Oralytics study, the decision-making algorithm used real-time toothbrush

sensor data to deliver personalized interventions. However, delays in receiving the data occurred due to issues such as improper docking of the toothbrush and Bluetooth or Wi-Fi connectivity problems. Data was permanently lost in some cases because the toothbrush lacked onboard storage and could not transmit data to the cloud. Unfortunately, the toothbrush was a commercial product and thus, the team had no control over whether local storage was available onboard the toothbrush, as discussed in the prior subsection on data storage. Despite these challenges, the algorithm had to continue functioning to make decisions with delayed or missing inputs. These lapses underscore the need for robust strategies to handle delayed and missing data during deployment, both to maintain algorithm performance and to ensure reproducibility.

When data are delayed or unavailable, it is critical to clearly document and archive the exact version of the data used for online learning or decision-making. If imputation is used to substitute missing inputs, the imputed values must be stored separately and preserved for reproducibility. These datasets should not be overwritten when the actual data becomes available later. Failure to do so can make it impossible to faithfully reconstruct how the algorithm operated in real time. For instance, in one deployment of HeartSteps [36], communication delays resulted in recent step-count information not being accessible to the algorithm at the time the algorithm needed to make decisions. In these scenarios, the online AI algorithm used imputed step-counts to make decisions. However, these imputed step count values were not logged, and later when the real step count information became available, it overwrote these imputed values, erasing the operational conditions under which real-time decisions had been made. As discussed in the limitations of a post-hoc analysis utilizing HeartSteps [36] deployment data, the lack of imputation records made it impossible to reconstruct the decision parameters used by the algorithm at specific decision-times – ultimately undermining the reproducibility of downstream scientific analyses. This example highlights why preserving the exact version of data used at decision-time is central to *deployment reproducibility*. Without storing imputed or provisional inputs alongside later ground-truth values, it becomes impossible to replay the system’s behavior under actual operating conditions, breaking the audit trail needed for reliable analysis and interpretation.

(e) Algorithm Version Control

Track algorithm logic with version control

Use rigorous version control to document every logic or code change to the algorithm, and record the deployed version at all times for traceable decision reconstruction.

During Phase 1, the online AI decision-making algorithm may undergo multiple revisions or modifications. Further changes can occur in Phase 2 (even during deployment) to address emerging issues, such as unforeseen technical constraints or bug fixes. These changes differ significantly from the online parameter updates performed during algorithm optimization; rather, they reflect changes in the algorithm’s core logic or structure. To ensure reproducibility, it is crucial to systematically track all such logic and code-level modifications using rigorous version control practices. It is equally important to document which specific software version of the algorithm was deployed at any given time. This typically involves maintaining a detailed, version-controlled repository – using tools such as Git – that records each change, including the specific code modifications, the rationale behind them, and corresponding timestamps. This applies also to code used for preprocessing input data from sensors and data collection tools, as well as data compression algorithms used to minimize storage requirements onboard edge devices and in the cloud.

For instance, during the MiWaves pilot study, the algorithm developers identified a bug in the online AI algorithm that caused it to incorrectly determine the time of day for each decision point. The bug was subsequently fixed, and the corrected algorithm was thoroughly tested to ensure the issue no longer persisted. The corrected algorithm was assigned a new version number and was deployed while the study was still ongoing. This version change

was appropriately recorded, ensuring that all subsequent decisions could be attributed to the correct algorithm version – an essential step for maintaining reproducibility of the algorithm’s decision-making. This example underscores the importance of having algorithmic version control to enable accurate reconstruction of decisions, verification of intervention outcomes, and overall transparency throughout the study. Rigorous version control therefore serves as a cornerstone of *deployment reproducibility*, since it guarantees that every decision can be traced back to the exact algorithmic logic that produced it.

(f) Monitoring System and Fallback methods

Use monitoring and fallbacks for safe delivery

Deploy a monitoring system to track algorithm behavior and support failure diagnosis, and define fallback methods to ensure consistent intervention delivery even when errors occur.

While monitoring systems have long been used in clinical trials [47–49] – to ensure intervention fidelity, assess participant engagement, and catch operational failures – online AI decision-making algorithms introduce novel and under-addressed monitoring challenges. Here, *intervention fidelity* refers to the extent to which an intervention is delivered as intended. In the context of online AI systems, this includes whether the algorithm generated decisions at the correct times, used the appropriate inputs, and behaved according to the deployed logic. The presence of an adaptive feedback loop significantly amplifies the potential consequences of system failures, data inconsistencies, or undetected bugs. Without a dedicated monitoring system for the online AI algorithm, failures in intervention systems powered by online AI algorithms can propagate rapidly and silently, compromising both intervention fidelity, scientific validity, and reproducibility. These failures can result in corrupted deployment data, missed or inappropriate interventions, and ultimately, invalid post-deployment analyses. In the worst-case scenario, such failures can lead to unintended participant harm, particularly if the AI algorithm is delivering biobehaviorally significant interventions. To address these challenges, recent frameworks have proposed systematic approaches to designing monitoring systems specifically for online AI decision-making algorithms in digital health [50]. By recording key metrics (e.g., model inputs, outputs, system states, errors, and exceptions) in a time-synchronized and queryable format, monitoring systems enable researchers to reconstruct algorithm behavior and validate whether the system functioned as intended throughout deployment.

However, monitoring alone is insufficient. *Fallback methods* are essential to ensuring that intervention delivery remains functional even when the online AI system encounters operational failures. These are predefined safety nets that activate in the presence of critical errors, missing data, or system crashes. Without them, system failures might be handled inconsistently or arbitrarily – often through ad hoc human decisions that are undocumented and difficult to reproduce. This makes it nearly impossible to reproduce the algorithm’s behavior or interpret the effects of interventions. By providing a consistent and well-documented response to failure conditions, fallback mechanisms help maintain the traceability and integrity of decision-making, even when the system is running in real-world conditions where things can go wrong.

For instance, in the Oralystics study [24,25], a fallback intervention decision was executed if the online AI algorithm failed to retrieve the necessary data to form a decision. The fallback intervention strategy randomly selected among the available intervention options with equal probability. Although these interventions were not personalized using the online AI algorithm, this ensured that individuals would still receive an intervention decision, preserving both the reliability of the system and the completeness of collected data required for scientific discovery and iterative deployment. Together, monitoring systems and fallback methods are essential for *deployment reproducibility*, by ensuring that system behavior and decision-making can be reconstructed even in the presence of failures.

5. Discussion

Reproducibility is not a purely academic concern. It is essential for both researchers and practitioners. For researchers, reproducibility enables re-analysis of past deployments, refinement of models, and valid causal inference — all of which depend on high-fidelity, reliable system execution. For industry stakeholders, it supports dependable product behavior, streamlined algorithm updates, and reduced operational risk. In digital health research, where interventions and associated algorithms are iteratively revised to accommodate evolving technologies and shifting real-world conditions, reproducibility and fidelity are indispensable.

This paper proposed a workflow aimed at addressing critical challenges related to the reproducibility and reliability of online AI decision-making algorithms in digital health interventions. These challenges arise both across deployments—where algorithms are repeatedly designed, tested, and refined—and within deployments, where algorithms must respond in real time to dynamic individual contexts, data irregularities, and system constraints. The workflow is structured around three key phases: (1) online AI algorithm design, (2) integration, assurance, and deployment, and (3) post-deployment analysis. Each of these phases maps onto a distinct reproducibility objective introduced in Section 1: *design reproducibility* in Phase 1, *deployment reproducibility* in Phase 2, and *inference reproducibility* in Phase 3. This framing highlights how our workflow operationalizes reproducibility across the full lifecycle of an online AI system for digital-interventions. Section 3 introduces these phases and Section 4 offers concrete, field-tested recommendations based on real-world deployments of HeartSteps, Oralytics, and MiWaves. These include isolating software environments and system components, designing autonomous algorithms with stability guarantees, logging inputs and outputs for traceability, preserving real-time decision context despite data latency or dropouts, implementing version control for algorithm logic, and deploying monitoring systems and fallback methods. Together, these recommendations translate the abstract goal of reproducibility into a set of operational practices that can be implemented across diverse deployment settings.

At the same time, achieving reproducibility in practice is inherently constrained by real-world limitations. In many cases, digital interventions rely on commercial systems—such as wearable sensors, smartphones, or cloud platforms—whose internal software or data pipelines are not transparent or controllable. These systems may update their firmware or processing logic without notice, making it more difficult to achieve reproducible results even when the same inputs are presented. For example, a commercial wearable may alter its step-count algorithm or apply new filters without informing developers, breaking assumptions of consistency across deployments. These constraints must be acknowledged as inherent limits on reproducibility, particularly in intervention systems built atop proprietary platforms.

In short, reproducibility must be treated as a central design goal in online AI for digital health—pragmatically pursued, transparently documented, and continuously evaluated. While perfect data reproducibility may not always be achievable, systems should be engineered to support it to the greatest extent possible. Doing so is essential not only for scientific validity and individual safety, but also for enabling the kind of continual learning and iterative refinement that underpins adaptive digital health interventions. As the field continues to evolve, centering reproducibility in the design and deployment of online AI systems will be key to building digital health interventions that are not only effective, but trustworthy, sustainable, and scientifically robust.

Acknowledgements. Research reported in this paper was supported by NIH/NIDA P50DA054039, NIH/NIDCR UH3DE028723, NIH/NHLBI R01HL125440, NIH/NIBIB K99EB037411, NIH/NIBIB and OD P41EB028242, and 5P30AG073107-03 GY3 Pilots. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Asim Gazi is supported by Schmidt Science Fellows, in partnership with the Rhodes Trust. Susan Murphy holds concurrent appointments at Harvard University and as an Amazon Scholar. This paper describes work performed at Harvard University and is not associated with Amazon.

References

- Himle JA, Weaver A, Zhang A, Xiang X. 2022 Digital mental health interventions for depression. *Cognitive and Behavioral Practice* **29**, 50–59.
- Gunning FM, Anguera JA, Victoria LW, Areán PA. 2021 A digital intervention targeting cognitive control network dysfunction in middle age and older adults with major depression. *Translational psychiatry* **11**, 269.
- Killian JA, Wilder B, Sharma A, Choudhary V, Dilkina B, Tambe M. 2019 Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* pp. 2430–2438.
- Ghosh S, Guo Y, Hung PY, Coughlin L, Bonar E, Nahum-Shani I, Walton M, Murphy S. 2024 MiWaves Reinforcement Learning Algorithm. *arXiv preprint arXiv:2408.15076*.
- Nesvåg S, McKay JR. 2018 Feasibility and effects of digital interventions to support people in recovery from substance use disorders: systematic review. *Journal of medical Internet research* **20**, e255.
- Garland EL, Gullapalli BT, Prince KC, Hanley AW, Sanyer M, Tuomenoksa M, Rahman T. 2023 Zoom-based mindfulness-oriented recovery enhancement plus just-in-time mindfulness practice triggered by wearable sensors for opioid craving and chronic pain. *Mindfulness* **14**, 1329–1345.
- Gazi AH, Harrison AB, Lambert TP, Obideen M, Alavi P, Murrah N, Shallenberger L, Driggers EG, Ortega RA, Washington BP, Walton KM, Welsh JW, Vaccarino V, Shah AJ, Tang YL, Gupta R, Back SE, Inan OT, Bremner JD. 2022 Transcutaneous Cervical Vagus Nerve Stimulation Reduces Behavioral and Physiological Manifestations of Withdrawal in Patients with Opioid Use Disorder: A Double-Blind, Randomized, Sham-Controlled Pilot Study. *Brain Stimulation* **15**, 1206–1214.
- Liao P, Greenewald K, Klasnja P, Murphy S. 2020 Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* **4**, 1–22.
- Stieger M, Allemand M, Lachman ME. 2023 Effects of a digital self-control intervention to increase physical activity in middle-aged adults. *Journal of health psychology* **28**, 984–996.
- Widmer RJ, Allison TG, Lennon R, Lopez-Jimenez F, Lerman LO, Lerman A. 2017 Digital health intervention during cardiac rehabilitation: a randomized controlled trial. *American heart journal* **188**, 65–72.
- Roberts AL, Fisher A, Smith L, Heinrich M, Potts HW. 2017 Digital health behaviour change interventions targeting physical activity and diet in cancer survivors: a systematic review and meta-analysis. *Journal of Cancer Survivorship* **11**, 704–719.
- El Hassouni A, Hoogendoorn M, Eiben AE, Van Otterlo M, Muhonen V. 2019 End-to-end personalization of digital health interventions using raw sensor data with deep reinforcement learning. In *IEEE/WIC/ACM International Conference on Web Intelligence* pp. 258–264.
- Hwang M, Jiang Y. 2023 Personalization in digital health interventions for older adults with cancer: A scoping review. *Journal of geriatric oncology* **14**, 101652.
- Cancela J, Charlafti I, Colloud S, Wu C. 2021 Digital health in the era of personalized healthcare: opportunities and challenges for bringing research and patient care to a new level. *Digital health* pp. 7–31.
- Zhang KW, Closser N, Trella AL, Murphy SA. 2024 Replicable Bandits for Digital Health Interventions. *arXiv preprint arXiv:2407.15377*.
- of Sciences NA, Medicine, Policy, Affairs G, on Research Data B, Information, on Engineering D, Sciences P, on Applied C, Statistics T et al.. 2019 *Reproducibility and replicability in science*. National Academies Press.
- Yu B. 2020 Veridical data science. In *Proceedings of the 13th international conference on web search and data mining* pp. 4–5.
- Yu B, Barter RL. 2024 *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press.
- Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, Murphy SA. 2015 Microrandomized trials: An experimental design for developing just-in-time adaptive interventions.. *Health Psychology* **34**, 1220.
- Klasnja P, Smith S, Seewald NJ, Lee A, Hall K, Luers B, Hekler EB, Murphy SA. 2019 Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. *Annals of Behavioral Medicine* **53**, 573–582.

21. Dempsey W, Liao P, Klasnja P, Nahum-Shani I, Murphy SA. 2015 Randomised trials for the Fitbit generation. *Significance* **12**, 20–23.
22. Liao P, Klasnja P, Tewari A, Murphy SA. 2016 Sample size calculations for micro-randomized trials in mHealth. *Statistics in medicine* **35**, 1944–1971.
23. Nahum-Shani I, Greer ZM, Trella AL, Zhang KW, Carpenter SM, Ruenger D, Elashoff D, Murphy SA, Shetty V. 2024 Optimizing an adaptive digital oral health intervention for promoting oral self-care behaviors: Micro-randomized trial protocol. *Contemporary clinical trials* **139**, 107464.
24. Trella AL, Zhang KW, Carpenter SM, Elashoff D, Greer ZM, Nahum-Shani I, Ruenger D, Shetty V, Murphy SA. 2024 Oralytics Reinforcement Learning Algorithm. *arXiv preprint arXiv:2406.13127*.
25. Trella AL, Zhang KW, Jajal H, Nahum-Shani I, Shetty V, Doshi-Velez F, Murphy SA. 2025 A Deployed Online Reinforcement Learning Algorithm In An Oral Health Clinical Trial. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 39 pp. 28792–28800.
26. Coughlin LN, Campbell M, Wheeler T, Rodriguez C, Florimbio AR, Ghosh S, Guo Y, Hung PY, Newman MW, Pan H et al.. 2024 A mobile health intervention for emerging adults with regular cannabis use: A micro-randomized pilot trial design protocol. *Contemporary Clinical Trials* **145**, 107667.
27. Ghosh S, Guo Y, Hung PY, Coughlin L, Bonar E, Nahum-Shani I, Walton M, Murphy S. 2024 rebandit: Random effects based online rl algorithm for reducing cannabis use. *arXiv preprint arXiv:2402.17739*.
28. Roy A, Hossain M, Muromachi Y. 2022 A deep reinforcement learning-based intelligent intervention framework for real-time proactive road safety management. *Accident Analysis & Prevention* **165**, 106512.
29. MacLean D, Roseway A, Czerwinski M. 2013 MoodWings: a wearable biofeedback device for real-time stress intervention. In *Proceedings of the 6th international conference on PErvasive Technologies Related to Assistive Environments* pp. 1–8.
30. Gazi AH, Gao D, Ghosh S, Xu Z, Trella A, Klasnja P, Murphy SA. 2025 Digital twins for just-in-time adaptive interventions (JITAI-twins): A framework for optimizing and continually improving JITAIIs (preprint). *JMIR Preprints*. (<https://doi.org/10.2196/preprints.72830>)
31. Hentati A, Forsell E, Ljótsson B, Kaldo V, Lindefors N, Kraepelien M. 2021 The effect of user interface on treatment engagement in a self-guided digital problem-solving intervention: A randomized controlled trial. *Internet interventions* **26**, 100448.
32. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. 2020 Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. *Journal of medical Internet research* **22**, e21687.
33. of Health NI et al.. 2018 Data and safety monitoring board (DSMB) guidelines. *National Institutes of Health*.
34. Proschan MA, Lan KG, Wittes JT. 2006 *Statistical monitoring of clinical trials: a unified approach*. Springer Science & Business Media.
35. Duncan J, Kapoor R, Agarwal A, Singh C, Yu B. 2022 VeridicalFlow: a Python package for building trustworthy data science pipelines with PCS. *Journal of Open Source Software* **7**, 3895.
36. Ghosh S, Kim R, Chhabria P, Dwivedi R, Klasnja P, Liao P, Zhang K, Murphy S. 2024 Did we personalize? assessing personalization by an online reinforcement learning algorithm using resampling. *Machine learning* **113**, 3961–3997.
37. Jones RB, Thapar A, Rice F, Mars B, Agha SS, Smith D, Merry S, Stallard P, Thapar AK, Jones I et al.. 2020 A digital intervention for adolescent depression (MoodHwb): mixed methods feasibility evaluation. *JMIR mental health* **7**, e14536.
38. Bendtsen M, Åsberg K, McCambridge J. 2023 Mediators of effects of a digital alcohol intervention for online help-seekers: findings from an effectiveness trial. *Drug and alcohol dependence* **251**, 110957.
39. Domhardt M, Lutsch A, Sander LB, Paganini S, Spanhel K, Ebert DD, Terhorst Y, Baumeister H. 2023 Mediators of digital depression prevention in patients with chronic back pain: Findings from a multicenter randomized clinical trial.. *Journal of Consulting and Clinical Psychology* **91**, 462.
40. Yu B. 2013 Stability. .
41. Elliott CF, Duncan JP, Tang TM, Behr M, Kumbier K, Yu B. 2024 Designing a data science simulation with MERITS: A primer. *arXiv preprint arXiv:2403.08971*.

42. Brewer EA. 2015 Kubernetes and the path to cloud native. In *Proceedings of the sixth ACM symposium on cloud computing* pp. 167–167.
43. Merkel D et al.. 2014 Docker: lightweight linux containers for consistent development and deployment. *Linux j* **239**, 2.
44. Trella AL, Zhang KW, Nahum-Shani I, Shetty V, Doshi-Velez F, Murphy SA. 2022 Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms* **15**, 255.
45. Stephan M, Hoffman MD, Blei DM et al.. 2017 Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research* **18**, 1–35.
46. Hoffman MD, Blei DM, Wang C, Paisley J. 2013 Stochastic variational inference. *the Journal of machine Learning research* **14**, 1303–1347.
47. Wiltsey Stirman S, Baumann AA, Miller CJ. 2019 The FRAME: an expanded framework for reporting adaptations and modifications to evidence-based interventions. *Implementation Science* **14**, 1–10.
48. Glasgow RE, Vogt TM, Boles SM. 1999 Evaluating the public health impact of health promotion interventions: the RE-AIM framework.. *American journal of public health* **89**, 1322–1327.
49. McCreight MS, Rabin BA, Glasgow RE, Ayele RA, Leonard CA, Gilmartin HM, Frank JW, Hess PL, Burke RE, Battaglia CT. 2019 Using the Practical, Robust Implementation and Sustainability Model (PRISM) to qualitatively assess multilevel contextual factors to help plan, implement, evaluate, and disseminate health services programs. *Translational behavioral medicine* **9**, 1002–1011.
50. Trella AL, Ghosh S, Bonar EE, Coughlin L, Doshi-Velez F, Guo Y, Hung PY, Nahum-Shani I, Shetty V, Walton M et al.. 2024 Effective Monitoring of Online Decision-Making Algorithms in Digital Intervention Implementation. *arXiv preprint arXiv:2409.10526*.