Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.DOI

Reliable data collection in participatory trials to assess digital healthcare applications

JUNSEOK PARK[1,2], SEONGKUK PARK[3], GWANGMIN KIM[1,2], KWANGMIN KIM[1,2], JAEGYUN JUNG[2,4], SUNYONG YOO[5], GWAN-SU YI[1] AND DOHEON LEE[1,2]

[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141 Republic of Korea
[2]Bio-Synergy Research Centre, KAIST, Daejeon 34141, Republic of Korea
[3]Information Electronics Research Institute, KAIST, Daejeon, 34141, Republic of Korea
[4]Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141 Republic of Korea
[5]School of Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, Republic of Korea

Corresponding author: Doheon Lee (dhlee@kaist.ac.kr)

“This work was supported by the Bio-Synergy Research Project (NRF-2012M3A9C4048758) of the Ministry of Science, ICT and Future Planning through the National Research Foundation of the Republic of Korea.”

ABSTRACT The number of digital healthcare mobile applications in the market is exponentially increasing owing to the development of mobile networks and widespread usage of smartphones. However, only few of these applications have been adequately validated. Like many mobile applications, in general, the use of healthcare applications is considered safe; thus, developers and end users can easily exchange them in the marketplace. However, existing platforms are unsuitable for collecting reliable data for evaluating the effectiveness of the applications. Moreover, these platforms reflect only the perspectives of developers and experts, and not of end users. For instance, typical clinical trial data collection methods are not appropriate for participant-driven assessment of healthcare applications because of their complexity and high cost. Thus, we identified the need for a participant-driven data collection platform for end users that is interpretable, systematic, and sustainable, as a first step to validate the effectiveness of the applications. To collect reliable data in the participatory trial format, we defined distinct stages for data preparation, storage, and sharing. The interpretable data preparation consists of a protocol database system and semantic feature retrieval method that allow a person without professional knowledge to create a protocol. The systematic data storage stage includes calculation of the collected data reliability weight. For sustainable data collection, we integrated a weight method and a future reward distribution function. We validated the methods through statistical tests involving 718 human participants. The results of a validation experiment demonstrate that the compared methods differ significantly and prove that the choice of an appropriate method is essential for reliable data collection, to facilitate effectiveness validation of digital healthcare applications. Furthermore, we created a Web-based system for our pilot platform to collect reliable data in an integrated pipeline. We compared the platform features using existing clinical and pragmatic trial data collection platforms.

INDEX TERMS Digital health, Digital healthcare app data collection platform, Crowdsourcing, Participatory trial, Biomedical informatics

I. INTRODUCTION

With the widespread popularity of wireless devices, such as smartphones, healthcare has become one of the most promising fields in the applications industry. Approximately 200 mobile health applications are newly added in mobile application stores every day, and investment in digital healthcare is booming [1], [2]. While many of these applications are general wellness applications that help users manage their activities, some healthcare applications claim to directly ameliorate a symptom or disease. These applications offer information about the treatments they propose, which may take the form of visual stimuli that are effectively delivered to users as digital therapy. reSET-O, an 84-day prescription digital therapeutic application for the treatment of opioid use disorders, has been approved by the U.S. Food and Drug Administration (FDA) [3]. Akili Interactive has clinically demonstrated that interactive digital treatment through a video game, which is under review by the FDA, may reduce the symptoms of attention deficit hyperactivity disorder (ADHD) and sensory processing disorder [4], [5]. Omada Health showed that a Web-based diabetes prevention program can significantly lower the hemoglobin A1c levels of a user. Moreover, the Web-based program showed a higher rate of patient engagement than the traditional in-person diabetes prevention plan [6]. Researchers and the public have shown considerable interest in this product as an example of digital medicine [7].

The increased interest in digital health has led researchers to study the effectiveness of current applications. As a practical example, Pokémon GO, which is a game application based on augmented reality, positively affects social interaction: in a study, 43.2% of users spent more time with their family [8]. In contrast, a six-week online intervention experiment revealed that Headspace, which is a healthcare application that is claimed to achieve mindfulness by guided meditation, has a relatively small effect on mindfulness [9]. This outcome contradicts the results of previous randomized controlled trials of the application [10]. Because of such contradictions, a reliable mechanism to validate digital health applications is required [11].

Currently, no appropriate platform that evaluates the effectiveness of an application using a systematic and objective validation method, exists in the digital healthcare field [11]. In particular, the development of mobile health applications is very fast and they present few safety issues, which tends to reduce the burden of regulation. As a result, direct trade in these applications between developers and end users is facilitated. However, the few existing platforms are not suitable for evaluating the effectiveness of such applications. For example, the review platforms and guidelines that are being developed to resolve this issue, provide only limited technical information obtained from the perspectives of developers, experts, and regulatory authorities [12]–[16]. Thus, we identified the need for a participant-driven data collection platform for end users as an interpretable, systematic, and sustainable tool to validate the effectiveness of these applications.

In an appropriate validation method, first, reliable data are collected; then, these data are analyzed for verification [17]. Hence, data collection is the basis of data analysis. The utilization conditions of healthcare applications differ from an ideal research-intensive environment, and the collected data are related to the term "real-world data" [18]. An incomplete collection of real-world data can cause inaccurate data analysis as a result of inconsistency, poor data quality, and noisy or missing data [19]. Thus, data collection is an essential stage of the verification workflow. In the following, we compare previous expert-oriented data collection platforms in existing research fields that are relevant to the development of a participant-driven data collection platform.

A general study that offers a highly effective and efficacious measurement method is clinical trial. A clinical trial includes a reliable data collection procedure based on a clinical protocol in a controlled environment [20]. The clinical protocol is a translational guide for data collection, and advanced digital technology has been developed to provide an electronic data capture (EDC) platform to prepare, store, share, examine, and analyze data from electronic case report forms (eCRFs) within the protocol. Therefore, the platform is complex and the general public cannot easily use it. For instance, Rave, an EDC platform created by Medidata, showed the highest level of popularity and satisfaction in G2 Crowd, a software comparison Website [21]. Rave follows a typical EDC design approach. In addition, it has a more advanced feature for designing the basic document structure of a protocol and provides optimized protocols using a PICAS database containing trial cost information [22]. Despite this functionality, Rave is still appropriate only for expert use. Transparency Life Sciences (TLS) is a leader in digital drug development services offering virtual and hybrid clinical trials [23]. TLS uses the crowdsourcing method for protocol design. Here, experts create a survey project for a disease and release the project to all TLS users. After the release, the users interested in the project start to participate and provide their identifying characteristic, such as patient family member, researcher, or patient. In the last step of the project, the TLS clinical trial team designs the protocol using the collected data of the participants. However, in this case also, the creation of the protocol is driven by an expert team, rather than general users. Accordingly, because of their complexity and high cost, the data collection methods in clinical trials are not appropriate for participant-driven assessment of healthcare applications, the number of which is growing exponentially, [24], [25].

A pragmatic trial is a study of the real-world measure of the effectiveness of intervention in broad groups [26]. A data collection platform of a pragmatic trial includes not only EDC but also specialized research tools and general surveys. These data collection platforms can be Web-based survey applications, mailed questionnaires, or specific healthcare applications developed from research kits [27]–[29]. Therefore, while the platforms still suffer from the issue of complexity, the possibility of collecting less reliable data also exists. For instance, PatientsLikeMe is a platform that shares experience data to help participants understand the possible treatments of particular disease conditions based on the experiences of others [30]. However, it does not provide an environment in which members of the public can lead the study preparation, and the platform has no feature for evaluating the reliability of the collected data. An additional example is Amazon Mechanical Turk (MTurk). MTurk is a crowdsourcing marketplace for the recruitment of research participants and for platforms for conducting surveys [31]. However, the platforms do not provide a standardized method for the data preparation stage. In other words, the platforms require clinical experts to prepare data collection procedures and systems based on their knowledge. MTurk provides a feature to approve or reject an individual participant, but the feature relies on a subjective judgment and suffers obvious objectivity limitations. We found that this platform offers no suitable method to measure the reliability of the data collected from the participants.

A participatory trial platform allows a public user to create and conduct a participant-driven study to measure the effectiveness of products in daily life. The core factors of the platform are simplicity, reliability, and sustainability. According to this description, we identified comparable platforms that have alternative, similar, or essential features. For example, Google Play and Apple App Store are alternative platforms, because they maintain user ratings and reviews in the public domain [32]–[34]. Both platforms provide free text reviews and scaled rating functions as a data storage feature. However, the free text reviews have an issue with natural language processing, which does not allow structural data collection [35]. In other words, the platforms offer no simple data preparation method to create a systematic data collection protocol. In addition, because they reveal previously collected data to new participants, a possible risk factor of the platforms is transfer biases that could affect new reviews and ratings [36]. An additional limitation of the platforms is that they offer no features for evaluating data reliability. RankedHealth and NHS Apps Library also represent platforms that are similar [37]–[39]. RankedHealth includes a procedure to minimize the influence of individual reviewers to ensure data reliability. NHS App Library publishes safe and secure applications by utilizing questions designed by experts from technical and policy backgrounds, in their evaluation procedure. However, these platforms suffer from a limitation: the consensus of the experts is focused on evaluating the technical performance of an application and the expert assessment does not validate the effectiveness or the user experience of the application. Thus, they are not appropriate for participant-driven studies.

Finally, all the platforms mentioned above are limited in that they address neither the prevention of participant dropout nor the software characteristics of digital healthcare applications. A study that included daily collection of pain data using digital measures demonstrated that the average self-report completion rate was 71.5% (261/365) days. In the most recent research studies, attempts were made to develop an incentivized program or in-game rewards system to increase the self-report rates, because sustainable data sharing for collecting large amounts of data is a crucial function of participatory trial platforms [40]–[42].

Furthermore, unlike drugs, functional foods, and mechanical devices that are difficult to modify after the market launch, the software of healthcare applications can potentially be updated [43]. The application features may require iterative evaluation following the upgrade. In summary, a new platform for digital healthcare applications should have a sustainable data collection function.

In this paper, we propose a participant-driven reliable data collection method for participatory trial platforms as a first stage in a system for understanding the effectiveness of healthcare applications. The method consists of three steps: interpretable data preparation, systematic data storage, and sustainable data sharing. We utilized a clinical trial protocol database and a semantic relatedness method for the participatory trial protocol to prepare data collection. We developed a data reliability weight (DRW) formula that collects data systematically. We propose a future reward distribution function related to the DRW to enable sustainable data collection. In the results section, we describe the experiments conducted to validate the reliability of the data collection method. The experiments included a comparison of the simplicity of the data preparation method, validation of the data reliability, and observations of the effect of future reward distribution. We report the results of experiments that involved a total of 718 human participants. The Institutional Review Board (IRB) of KAIST approved an IRB exemption for the experiments.

Moreover, we developed a Web-based pilot platform that is accessible to the public with real-world data as a crowdsourcing tool based on the citizen science concept. The pilot platform systematically integrates all the proposed methods. We conducted case studies on the pilot platform to validate its participant recruitment efficiency. To demonstrate the advantages of the proposed platform, we compared its functionality with that of an existing platform.

II. DEFINITION OF PARTICIPATORY TRIAL

A participatory trial is an expanded form of a trial involving humans, in which the public is utilized to test the effectiveness of products in daily life. The concept follows crowdsourcing and citizen science in the aspect of data-driven science [44], [45]. A participatory trial includes voluntary participation, in contrast to the selective inclusion of clinical trials and the broad inclusion of pragmatic trials. The participants who operate the trials, are members of the general public. The objective of a participatory trial is to inform consumers about the effectiveness of a product, and the method consists of a protocol that reflects daily life to maximize the reliability of the clinically relevant results. We compare clinical, pragmatic, and participatory trials in Table 1 [46]. We define a participatory trial so that the difference between this and other current types of trials conducted in modern society is clear.

The definitions of the terms and phrases used in this manuscript are as follows.

- Platform: Digital platform that is an environmental software system associated with each functional part of application programs.
- Data preparation: The process of determining the data collection structure. The term has the same meaning as the process of designing the protocol.
- Data storage: Evaluation of collected data to allow reliable storage and to determine the value of the data.
- Data sharing: Integration of collected data to allow sharing with others and to provide benefits to the data provider.

III. RELATED WORK

TABLE 1. Comparison of clinical, pragmatic, and participatory trials

Trial type	Clinical trial	Pragmatic trial	Participatory trial
Outcome	Efficacy	Effectiveness	Effectiveness
Objective	Assessment	Decision making	Information delivery
Protocol	Rigid protocol	Quasi-explanatory protocol	Interpretable protocol
Enrolment	Selective inclusion	Broad inclusion	Voluntary inclusion
Location	Controlled environment	Daily environment	Daily environment
Operation	Expert		Public

A. CLINICAL TRIALS AND PLATFORMS

Clinical trials have been the gold standard for the evaluation and development of medical interventions for over 70 years [47]. Focused on evaluating the efficacy of specific interventions, clinical trials frequently use controlled settings and apply strict criteria for the inclusion of participants and practitioners. The EDC system is increasingly recognized as a suitable method that has advantages in terms of real-time data analysis, management, and privacy protection [48]. The system is most useful in trials with complicated contexts, such as international, multi-centered, and cluster-randomized settings [49]. However, because of their strict and complex nature, clinical trials still have certain limitations in the in-depth evaluation of effectiveness or external validity [50].

B. PRAGMATIC TRIALS AND PLATFORMS

Clinical trials are frequently time consuming and challenging, because they utilize rigorous methods. These characteristics of clinical trials have increased the need for pragmatic trials that seek to understand real-world evidence, assessing the effectiveness of an intervention in actual clinical practice settings [26]. As introduced in the “PRECIS’’ assessment by Thorpe et al., these trials are designed to maximize the external validity of the research by including a heterogeneous population and setting patient-centered endpoints [51]. Various platforms, including mailed questionnaires, Web-based forms, and research kits, are used in these types of trials [27]–[29]. Consequently, pragmatic trials are an emerging source of clinical evidence pertaining to issues ranging from pediatric asthma and cardiovascular diseases to monetary incentives for smoking cessation [37], [52], [53]. However, they are subject to several challenges, such as low patient recruitment, insufficient data collection, and treatment variability [26], [54]. Further, inconsistency in data platforms is an additional major limitation of pragmatic trials [55].

IV. METHOD

We divided the data collection method for evaluating the effectiveness of the healthcare applications, into the stages of interpretable data preparation, systematic storage, and sustainable sharing. We used the essential methods at each stage to continuously collect highly reliable data. In addition, from a participatory trial perspective, we organized each method so that the public can easily participate and organize their own research, provide reliable information and share it to induce continued interest (Figure 1).

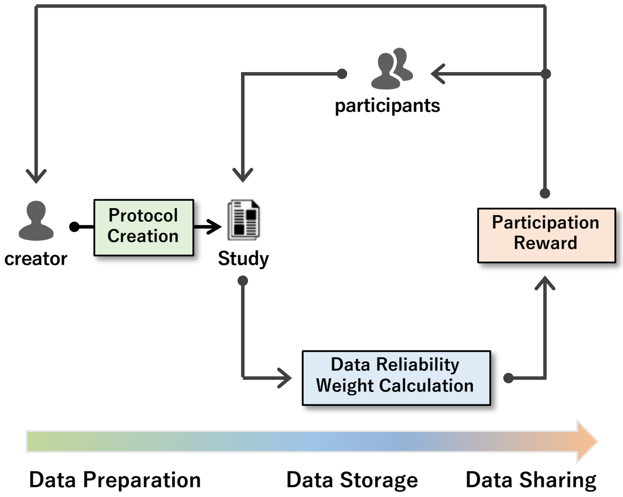


FIGURE 1. Method overview. The data collection method consists of three parts: data preparation, data storage, and data sharing. In the data preparation part, a creator can devise a new study for the evaluation of an application by searching in the protocol creation methods. Next, in data storage, the participants conduct a study, and the platform stores all the participants’ responses. At the same time, it collects the statistics of the responses and calculates the data reliability weight (DRW) of each. After the study is complete, in data sharing, the platform calculates future rewards according to the reliability of both the participants and the creator and distributes the rewards to each of them.

A. INTERPRETABLE DATA PREPARATION

At the data preparation stage, we determine the types of data that should be collected, the manner in which the data should be collected, and the preferred source of the data. Our objective is to consider the data as an interpretable resource for determining the effectiveness of a healthcare application. In the fields of clinical and pragmatic trials, researchers typically develop a protocol to facilitate the assessment of a changing symptom or disease [51], [56]. Healthcare applications are also focused on ameliorating a symptom or treating a disease. A phenotype is an observable physical characteristic of an organism, including a symptom and a disease. Therefore, we can effectively measure the change in a phenotype using an existing protocol that is semantically related to it.

We developed a protocol database system and a concept embedding model to calculate semantic relatedness to realize a protocol retrieval system for accomplishing interpretable data preparation. The database contains 184,634 clinical trial protocols to assist context-dependent selection of protocol elements [57]. We created a model to find a vector on a latent space based on the distributed representations of a documented method. Furthermore, we calculated the semantic relatedness between the input phenotype and protocols using cosine-similarity [58], [59]. Finally, we integrated the methods into the proposed platform for participatory trials, so that a member of the general public can create a study to easily determine the effectiveness of a healthcare application. Figure 2 shows the overall procedure of the method.

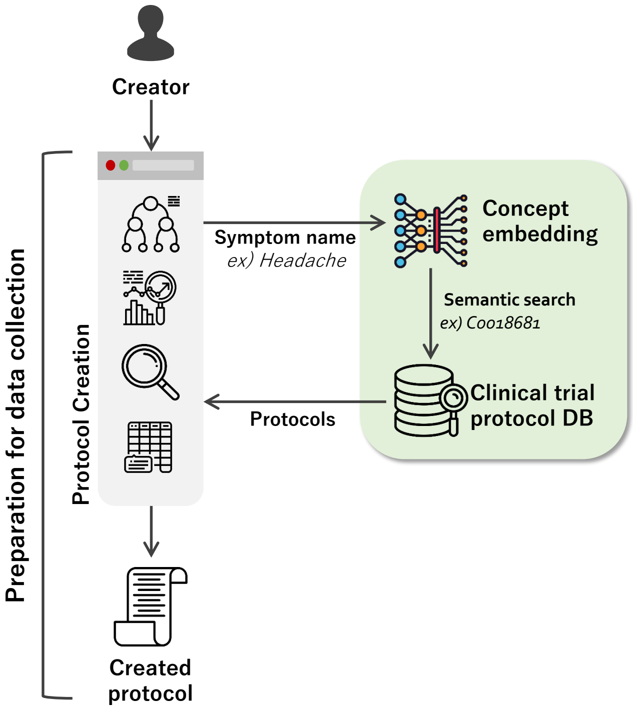


FIGURE 2. **Preparation to collect data.** The creator retrieves a previous protocol based on the symptom name of the application to verify effectiveness. Then, the creator prepares to collect data on the retrieved protocol and creates the participatory trial protocol for data storing.

B. SYSTEMATIC DATA STORAGE

The participatory trial platform uses crowdsourcing methods to collect data. Crowdsourcing refers to obtaining data for scientific research through the voluntary participation of the public. Data collection based on crowdsourcing is applied in many research fields, because it is a method that can quickly obtain a large amount of data from the public at a low cost [60], [61]. However, it is frequently noted that data collected by means of crowdsourcing are less reliable than those obtained by systematic approaches [62], [63]. Therefore, the main challenge with regard to data reliability is to devise a method to measure data credibility.

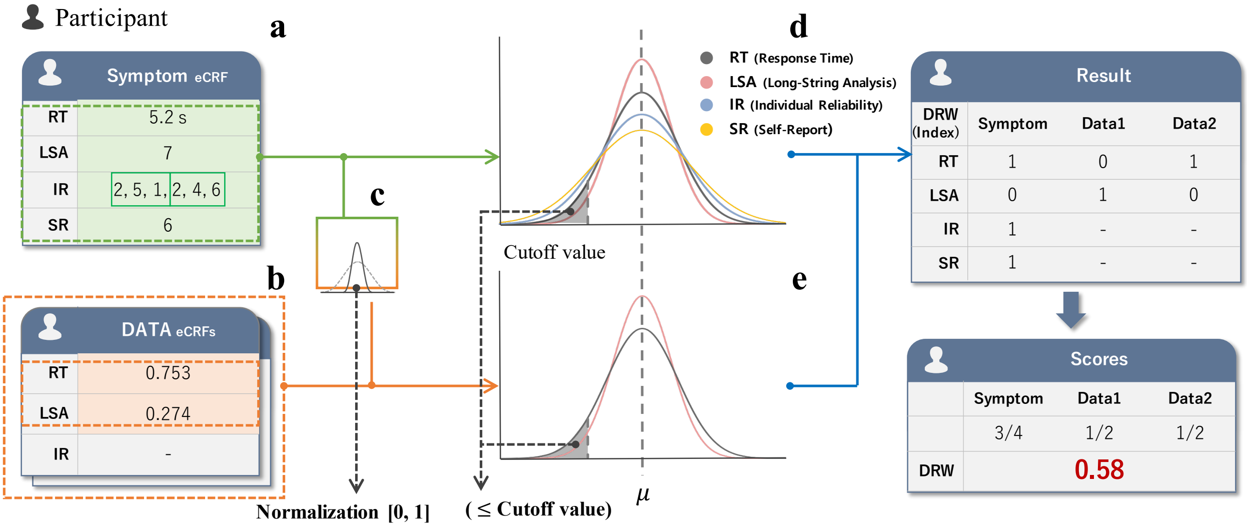
The primary purpose of this method is reliable data collection and delivery for evaluating effectiveness, which is the subsequent level on the path to validation. To achieve this purpose, a method to guarantee the reliability of input data is required. Therefore, we propose a formula to calculate the DRW through combined measures of the effort expended on inputting data.

Careful, relevant, and purposive data input by the participant can improve the quality of data in systematic data storage [64]. Huang et al. developed an insufficient effort responding (IER) detection method to determine the input data needed for providing accurate responses and correctly interpreting the data [65]. The recommended composition of IER is response time (RT), long-string analysis (LSA), and individual reliability (IR). RT is used to determine the effort expended based on execution time [66]. LSA uses intentional long-string patterns to assess the effort expended by a participant [67]. To calculate IR, a pair of data items are divided into two halves and the correlation between them is used to determine the normal response [68]. In addition, Huang et al. also presented a single self-report (SR) IER item, which has empirical usefulness [69]. The SR is a single item for detecting IER and consists of a 7-point Likert scale [70]. We developed the DRW calculation method based on the IER methods.

We calculated $DRW_{idx}(X)$ according to the type of eCRF for obtaining the DRW (1). f_{RT}, f_{LSA}, f_{SR}, and f_{IR} are the indexes of the DRW. The created protocol contains eCRFs, where an eCRF is a tool used to collect data from each participant. We provided four types of eCRF, as described in Supplementary Table 1. We calculated the DRW from the symptom and data type eCRF. The symptom type eCRF measures the presence and the degree of targeted symptoms only once, whereas the data type eCRF measures the effectiveness of the targeted applications repeatedly. $X = \{x_1, x_2, ..., x_n\}$ is the set of items in an eCRF. The platform receives 11 types of data items in an eCRF, as shown in Supplementary Table 2. Type codes ED, TEL, SET, DATE, PDATE, and FILE were not used to calculate the DRW.

$$f(X) = \{f_{RT}(X), f_{LSA}(X), f_{IR}(X), f_{SR}(X)\} \qquad (1)$$

We calculated f_{RT} and f_{LSA} for the symptom and data type of X. f_{RT} is calculated as the difference between the start and end time when the participant inputs data to X. The calculation of f_{LSA} uses the length of the input strings. Conversely, we calculated f_{SR} and f_{IR} only for the symptom type of X. We did not include the SR and IR item in the data type of X to avoid distortion of the iterative procedures by the response of the participant. To measure whether a participant is attentive while entering data, we placed an SR item at the end of the symptom type of X. The measured pattern of participants should be similar to the symptom type. To reflect the pattern for f_{IR}, we calculated a Pearson correlation coefficient [71]. r_{IR} is the correlation value of Z and $Z^{'}$ (2). $Z_k = \{z_{1_k}, z_{2_k}, ..., z_{m_k}\}$ and $Z^{'}_k = \{z^{'}_{1_k}, z^{'}_{2_k}, ..., z^{'}_{m_k}\}$ are sets to measure r_{IR} among items belonging to X. The number of items in Z and $Z^{'}$ is the same. We generated Z and $Z^{'}$ such that the correlation is close to one. z_{m_k} and $z^{'}_{m_k}$ are the value of each item of X_k. K is the number of participants, and k is the kth participant of K.



Result DRW (Index)	Symptom	Data1	Data2
RT	1	0	1
LSA	0	1	0
IR	1	-	-
SR	1	-	-

Scores	Symptom	Data1	Data2
	3/4	1/2	1/2
DRW	0.58		

FIGURE 3. Example of calculating the data reliability weight (DRW) of one participant in the study (a) The platform calculates the response time (RT), long-string analysis (LSA), individual reliability (IR), and self-report (SR) based on the data entered to calculate the DRW of the symptom electronic case report form (eCRF). Each participant has one symptom eCRF. (b) The platform stores the collected data in data eCRF. A data eCRF, unlike a symptom eCRF, does not display a specific pattern, and therefore, RT and LSA excluding IR and SR are calculated. (c) The platform calculates the cumulative distribution function for the normalization of the input values. (d) The platform calculates the cutoff values for RT, LSA, SR, and IR using values collected from all symptom type eCRFs entered by all participants in a study. (e) The platform groups all participants in the study into an eCRF consisting of items that are the same as the data type entered. (f) The platform then uses the values collected from each group's eCRF to calculate the RT, LSA, SR, and IR. (g) The platform calculates the DRW index values of the participant using the cutoff values obtained from each eCRF type. The platform uses the obtained cutoff values to calculate the DRW.

$$r_{IR_k} = \frac{m(\sum z_{m_k} z'_{m_k}) - (\sum z_{m_k})(\sum z'_{m_k})}{\sqrt{[m\sum {z_{m_k}}^2 - (\sum z_{m_k})^2][m\sum {z'_{m_k}}^2 - (\sum z'_{m_k})^2]}} \tag{2}$$

Each index has a cutoff value and the cutoff values for each prepared eCRF in a study is calculated. In detail, we calculated the mean (μ) value and the standard deviation (σ) to remove outliers. We removed outliers having a value greater than $\mu + 3\sigma$ or smaller than $\mu - 3\sigma$. After removing the outliers, we calculated μ' and σ'. Then, we calculated the cumulative distribution function (CDF) for the values of the DRW indexes (3). One purpose of CDF is to find a cutoff value that has the lowest area under the probability density function (PDF) among the input values. The second is to represent random variables of real values, the distribution of which is unknown, using the collected values [72]. Accordingly, we used the CDF values to find the μ'' and σ'' again to obtain a normal distribution. For the distribution, we obtained the cutoff value using the z-score for the p-value (default is 0.05). $h(X')$ returns 0 if the normalized input value is smaller than the cutoff value and 1 if the value is larger than the cutoff value (4). We calculated the DRW indexes using the binary decision function.

$$X' = g(f(X);\mu',\sigma') = \int_{-\infty}^{f(X)} \frac{1}{\sqrt{2\pi}\sigma'} e^{-\left(\frac{(x-\mu')^2}{2\sigma'^2}\right)} dx \tag{3}$$

$$h(X') = \begin{cases} 1 & \text{if } X' >= \text{cutoff value} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

We calculated the DRW of a participant when we computed all the cutoff values of the DRW indexes for each eCRF (5,6). N is the number of eCRFs assigned to a participant. C_{DRW_i} is the number of DRW calculation item counts for each eCRF type. Figure 3 displays an example calculation procedure for the DRW.

$$h_{DRW}(X') = h_{RT}(X') + h_{LSA}(X') + h_{IR}(X') + h_{SR}(X') \tag{5}$$

$$DRW_{participant_k} = \frac{1}{N}\sum_{i}^{N} \frac{1}{C_{DRW_i}} h_{DRW}(X'_i) \tag{6}$$

C. SUSTAINABLE DATA SHARING

It is intuitive that a reward can drive active participation. Recent scientific findings support the fact that rewards affect data acquisition [41], [73]. In our method, we provide financial future rewards in the form of exchangeable cryptocurrency to achieve sustainable data collection [74], [75]. To deliver the reward, we developed a study results transfer function that sends statistical information on the completed study to an external cryptocurrency system (Figure 4). The study participants periodically receive cryptocurrency as a reward for providing their data. The total amount of the reward depends on the external cryptocurrency system. However, the reward is a compensation that the participants expect to receive after the end of the study and does not constitute an immediate profit. Therefore, we attempted to induce sustainable data sharing by creating an expectation that active participation will earn high rewards [76]. A future reward distribution method that induces this expectation drives the creator's continuous management motivation and encourages the active involvement of participants.

We collect rewards for each completed study. R_{total} is $R_{creator} + R_{participants} + R_{system}$. R denotes rewards. $R_{creator}$ calculates the reward amount based on the proportion of the participants' compensation and the μ_{DRW} value of the study (7). A low μ_{DRW} value acts as a type of penalty. We introduced this factor to encourage the study creators to create a better plan and stimulate a study.

$$R_{creator} = \mu_{DRW} \times (1 - \alpha) \times R_{total} \tag{7}$$

Here, $R_{participant}$ and $R_{creator}$ are substrates of R_{total} (8). K is the total number of participants. DRW_k is the calculated DRW for the kth participant. The platform rewards participants for their efforts to input data. A participant receives a greater reward if he/she invests more than average effort; otherwise, the participant receives a lower reward. R_{system} acquires high compensation when the μ_{DRW} value is low. We added R_{system} to recover the cost incurred when careless participants waste platform resources by generating unreliable data. We utilize R_{system} as a platform maintenance cost.

$$\begin{aligned} R_{participant} &= \alpha \times R_{total} \\ &= \sum_{k=1}^{K} \frac{DRW_k}{\mu_{DRW}} \times \frac{\alpha \times R_{total}}{K} \end{aligned} \tag{8}$$

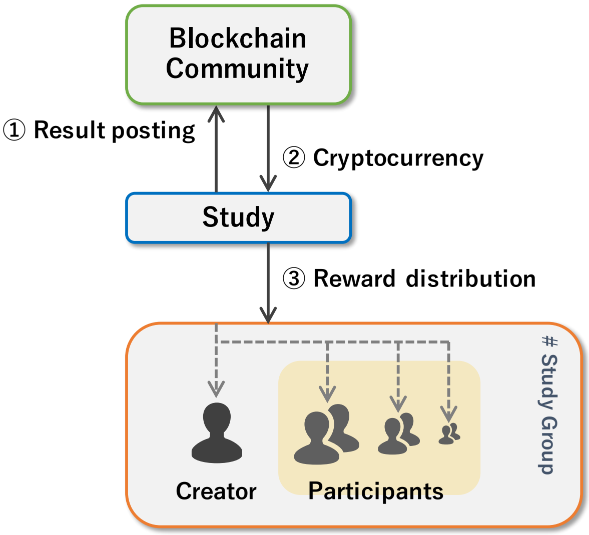


FIGURE 4. Participation and future rewards of the study. When participants take part in the study, the results are posted on the blockchain community (①) and the community provides cryptocurrency to the study (②). After the end of the study, the platform distributes the rewards according to the participants' level of involvement (③).

V. PILOT PLATFORM IMPLEMENTATION

The proposed platform is designed for a participatory trial. The characteristic of a participatory trial in terms of data collection is that, through systematic collection of data based on crowdsourcing, ordinary people can become researchers and conduct experiments. They can then share the results with the public [44], [77].

To achieve this goal, we developed a platform that integrates all the methods presented above, so that each feature is systematically connected to another. The working logic of the integrated platform is implemented as follows (Figure 5). Creators are people who recruit participants to suggest research and evaluate the effectiveness of healthcare applications. They can build studies in the data preparation phase by using the protocol database and protocol discovery methods. When the created study has been activated, users can join the study. The participants in the study are curious about the effectiveness of healthcare applications, and therefore, their attitude toward entering data is relatively serious. The research data and basic information about the study are stored in the blockchain to prevent data falsification in vulnerable systems [74]. The study ends after its predefined duration. Even if the number of participants is not sufficient to collect an adequate amount of data, the study is closed at the pre-set endpoint. The system then calculates the effectiveness score using the study data. The research results are open to both the public and experts in the data sharing phase, and participants receive rewards based on the data they enter.

Technically, we designed the pilot platform according to the interface segregation principle [78], which provides a user with a tool that utilizes only the services that the user expects. The platform consists of three layers: Web interface, application programming interface (API) engine, and database. The functions of each layer are based on Node.js, the Java programming language, and the PostgreSQL database. In addition, essential functions and fundamental database schema related to the EDC system were designed using OpenClinica 3.12.2 [79]. We also applied the EDC function to receive data in the Operational Data Model (ODM) XML format, which is compatible with the Clinical Data Interchange Standard Consortium [80], [81]. Finally, we integrated the blockchain system to prevent data falsification in the crowdsourced data collection platform in participatory trials [74]. See the Supplementary Document for more information on the platform development.

VI. RESULTS

The evaluation results of the proposed method consist of both quantitative measurements, i.e., the test results for each method, and qualitative measurements, i.e., a comparison of the platform functions with those of other platforms. We conducted the quantitative measurements with the participants in the platform. The measurements included simplicity of the protocol creation function in the data preparation stage, DRW validation in the data storage stage, and influence of expected rather than immediate rewards in the data sharing

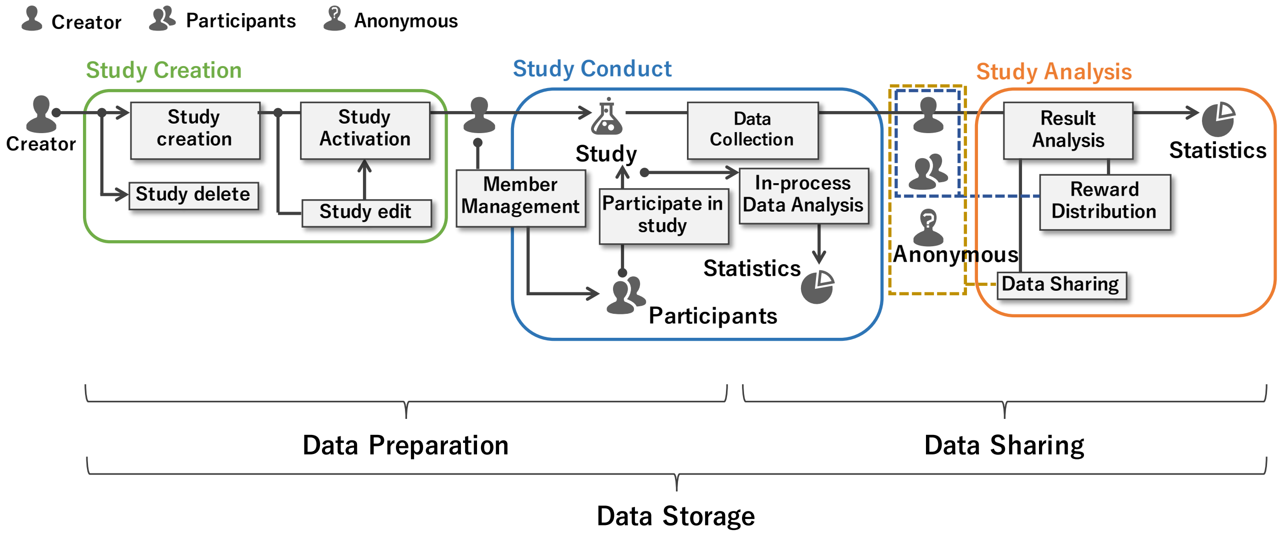


FIGURE 5. **Working logic overview of the platform.** The platform integrates and connects methods for reliable data collection at all stages of the study: creation, conduct, and analysis. In the creation stage, the creator can create, modify, or delete the study. When the creator has activated the study, it is open to other members of the public and they can participate. In the conduct stage, the creator can collect the data from the participants and monitor them. In addition, the creator and participants can monitor statistical information related to the study. In the analysis stage, all the data generated after the study is completed, are disclosed to the study creator, participants, and anonymous users. Thus, the data are available to the experts and the public. The platform services allow users to prepare, collect, and share reliable data in an integrated environment at each stage.

stage. All participants filled and signed an electronic consent form. In the quantitative measurements, the participants' data input was used for the purpose of method validation. We did not collect any personally identifiable information during the tests. We used qualitative measures to demonstrate the benefits of a participatory trial, as compared with a clinical or pragmatic trial.

A. COMPARISON OF THE SIMPLICITY OF DATA PREPARATION

We developed a protocol retrieval method based on the clinical trial protocol database and validated the methods developed in our previous studies [57], [59]. The methods constitute the core functions of data preparation. We validated that the semantic filters of the database can provide more appropriate protocols than a keyword search [59]. We used clinical trial protocols and corresponding disease conditions extracted from ClinicalTrials.gov as the golden standard set [82]. The F-1 score was 0.515; this score is higher than that of keyword search, 0.38. The concept embedding model to find the semantic relatedness of the clinical trial protocols showed a Spearman's rank correlation coefficient of 0.795 with the benchmark set from the Mayo Clinic [57], [83]. The results were higher than those of the previous method, with a Lesk of 0.46 and vector of 0.51. Finally, we conducted a user evaluation test of the protocol retrieval system from the point of view of 10 clinical trial experts [59]. Our system presented a score of 1.6 for difficulty and 6.5 for satisfaction on a 7-point Likert scale. The scores of ClinicalTrials.gov were 6.2 and 2.3, respectively.

However, previous results were limited, because tests were conducted solely to obtain expert views. In addition, we obtained the results for a clinical trial, not a participatory trial. In contrast, the interpretable data preparation method we propose, is a publicly available system for applying the above technique from the perspective of a participatory trial. Therefore, further validation that the proposed method is sufficiently simple to allow a member of the general public to create an interpretable protocol for preparing data collection, is required.

For the validation, we designed an experiment that collected the Usefulness, Satisfaction, and Ease of Use (USE) scores from human participants as a reflection of their experience of the different methods in terms of simplicity [84]. We recruited human participants who represented general users using MTurk [31]. We used an expected mean of 1.1 and an expected SD of 1.14, as in the previous USE study [85]. We set a power of 95% and a one-sided level of significance of 5% to calculate the number of participants. The number of participants obtained by adjusting the sample size for t-distribution was 20 [86]. The participants in the experiment compared the data preparation function, which is related to protocol creation, of the proposed method and the existing expert method. We prepared the data preparation method as a simulation program (PM); the complicated method used Openclinica 3.14, an open-source EDC program (OC) [79]. The two programs recorded a randomly generated ID for each participant to allow their test scores to be identified. The participants responded to the USE questionnaire, which consisted of 30 items scored on a 7-point Likert scale to obtain the comparison score and 6 short items for recording their opinion after each use [84]. We reversed the 7-point Likert scale so that the participants would concentrate more on the questionnaire [69], [87]. We restored the modified scores in the results calculation.

In the experiment description, we provided only an elementary data preparation guide and no detailed manual. We also divided participants into two groups to prevent recall bias occurring in the order of the program usage [36]. We reversed the order of use for each group. Finally, we identically configured the hardware environment on the cloud-computing resources of Amazon Web services to identify only software differences [88].

After the experiment was completed, we excluded the results of participants who did not complete all the USE items, recorded the same values for all the items, or did not complete the test of method systems. Consequently, we could obtain the results of 42 participants from the combined data set of the two groups. See Supplementary Data. For conducting the descriptive statistics of both methods, we divided the USE items into four dimensions [85]. Accordingly, the 30 USE items were broken down into 8 usefulness (UU) items, 11 ease of use (UE) items, 4 ease of learning (UL) items, and 7 satisfaction (US) items. The statistics indicated that the scores of the proposed method were above 5 for all the USE dimensions (Figure 6 and Table 2). We also found that the scores showed negative skewness for all aspects.

As shown in Table 2, the average USE score of PM was 5.548 (SD=0.897) and of OC was 4.3 (SD=1.5). We conducted a statistical test to confirm a significant difference between the average scores. To check the USE score's normality, we conducted a Shapiro-Wilk test [89]. The test result p-value of our score was 0.082, and the compared score was 0.238, which satisfied normality (p > 0.05). Therefore, we conducted paired sample t-tests; the USE score presented a significantly higher value for PM than OC, with t = 7.243, p <.001. Because the four dimensions in USE did not satisfy normality, we conducted a Wilcoxon signed-rank test [90]. The results of the test showed noticeable differences between the two methods in all dimensions (Table 3).

In addition to the above results, we also found that the participants completed 100% of tasks on PM, whereas they generated only basic metadata of the given tasks on OC. Thus, we concluded that PM is more suitable for interpretable data preparation in participatory trials.

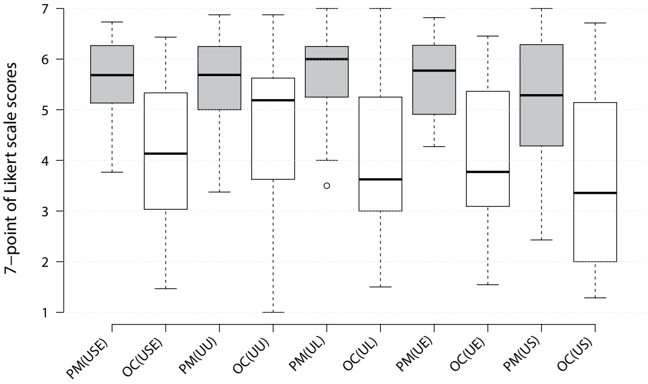


FIGURE 6. Whisker box plots of the dimensions in USE for the two groups. The proposed data preparation method (PM) showed higher records than the existing expert method (OC) in all domains of the tests.

TABLE 2. Descriptive statistics of USE in the proposed data preparation method and the existing expert method. Statistical description of proposed data preparation method (P) and existing expert method (O).

P O	Mean	Std	Min	Max	Skewness	Kurtosis
USE	**5.5** 4.1	0.8 1.4	**3.8** 1.5	**6.7** 6.4	-0.4 -0.1	-0.7 -1.0
UU	**5.6** 4.7	0.9 1.4	**3.4** 1.0	**6.9** 6.9	-0.6 -0.7	-0.3 -0.3
UL	**5.8** 4.0	0.8 1.5	**3.5** 1.5	**7.0** 7.0	-0.9 0.2	0.1 -1.0
UE	**5.6** 4.0	0.7 1.4	**4.3** 1.5	**6.8** 6.5	-0.3 0.0	-1.2 -1.1
US	**5.2** 3.6	1.1 1.7	**2.4** 1.3	**7.0** 6.7	-0.4 0.3	-0.5 -1.3

Note. N=42. Min = minimum; Max = maximum

TABLE 3. Related-samples Wilcoxon signed rank test of the dimensions in USE of the proposed data preparation method and the existing expert method

Dimension	N	s.t.s.	p
Usefulness (UU)	42	-4.327	<.001
Ease of Use (UE)	42	-5.196	<.001
Ease of Learning (UL)	42	-5.125	<.001
Satisfaction (US)	42	-4.848	<.001

Note. s.t.s = standardized test statistic

B. DATA RELIABILITY WEIGHT VALIDATION

The DRW is a weight, in the calculation of which a score that is the measure of the effort that a user expends when entering data is included. We examined the correlation between the Human Intelligent Task Approve Rate (HITAR) of MTurk and the DRW to assess whether this score is an effective measure. The HITAR is a score that evaluates how effectively a worker is performing tasks assigned by a requestor on MTurk. Consequently, we assumed that a higher HITAR would lead to a higher DRW. To verify this, we performed the following tests.

We prepared the first test to confirm that the DRW indexes correlate with the HITAR. The DRW indexes include RT, LSA, IR, and SR [65], [69]. Our objective was to organize the DRW using empirical DRW indexes to allow simple calculation in the eCRF. Because the DRW is a subset of IER, we built a questionnaire to calculate the DRW based on detecting IER [69]. The questionnaire contained 60 items of the international personality item pool—neuroticism, extroversion, and openness (IPIP-NEO)—to calculate IR [91]. We also added eight items of the infrequency IER scale to raise the concentration level of the participants at the beginning of the test and an SR item for the DRW score [70], [92]. We placed a self-input HITAR item box at the start of the questionnaire as the LSA item. Thus, the questionnaire consisted of 70 items.

To calculate the total sample size, we used a 0.13 expected mean of the paired differences based on the results of our operational test on the pilot platform, which used a total of 100 participants working in MTurk. We set a power of 95%, a one-sided level of significance of 5%, and an equal group size for sample size calculation [86]. The calculated sample size for each group was 153, and the total size was 306. We recruited the participants using MTurk; that is, MTurk workers took part in the study as participants. We divided the participants into two groups, those with HITAR levels above 95 and those with HITAR levels below 90 (called 95HITAR and 90HITAR, respectively. Participants in the 95HITAR group needed a record of more than 500 completed tasks to participate. The purpose of this was to select participants with an average of more than one year's experience, given that the average number of tasks completed in 2.5 years is 1,302, as analyzed in a previous study [93]. Participants in the 90HITAR group had no additional restrictions. This is because the initial HITAR given is 100, from which points are deducted when a task is not performed well. Participants read the task description posted on MTurk. All the settings of the test were identical for the two groups, except for the HITAR levels. The endpoint of the test was set to the completion time of the recruitment of one of the groups.

A total of 340 participants were involved in the test (see Supplementary Data 2). Each group consisted of 170 participants, and we considered that the number of participants reflected the calculated sample size. According to the data collected from the participants, we compared the average value of the DRW and DRW indexes for each group. In all the DRWs and DRW indexes, the 95HITAR group scored higher than the 90HITAR group (Figure 7(a)). We conducted a statistical analysis to understand the differences between the two groups. Table 4 displays the statistical description. The DRW index values in the table were converted to CDF values to calculate the DRW. The standard deviation (SD) in all areas, except RT and LSA, showed similar values. The reason was that the RT and LSA values of the 95HITAR group were very high. We calculated the cutoff values using the CDF values obtained through the DRW calculation method. We checked that only a range of values of approximately 0.5 or more passed. We verified the statistical significance between the mean values of the two groups using the independent samples t-test. We performed a Levene's test to consider unequal variance in the t-test [94]. The p-values indicate that the results are statistically meaningful. In the two groups, only the values of SR are similar. We presumed that the mean value difference of SR was small; however, SR would have only a small effect on the calculation of the DRW. Furthermore, we calculated effect size, an indicator of the amount of difference between the two groups [95]. Because we were comparing two groups having the same sample size, we calculated the effect size using Cohen's d; the result was close to a large effect (0.8) according to the interpretations of Cohen's d [96].

Then, we considered the correlation between the DRW and

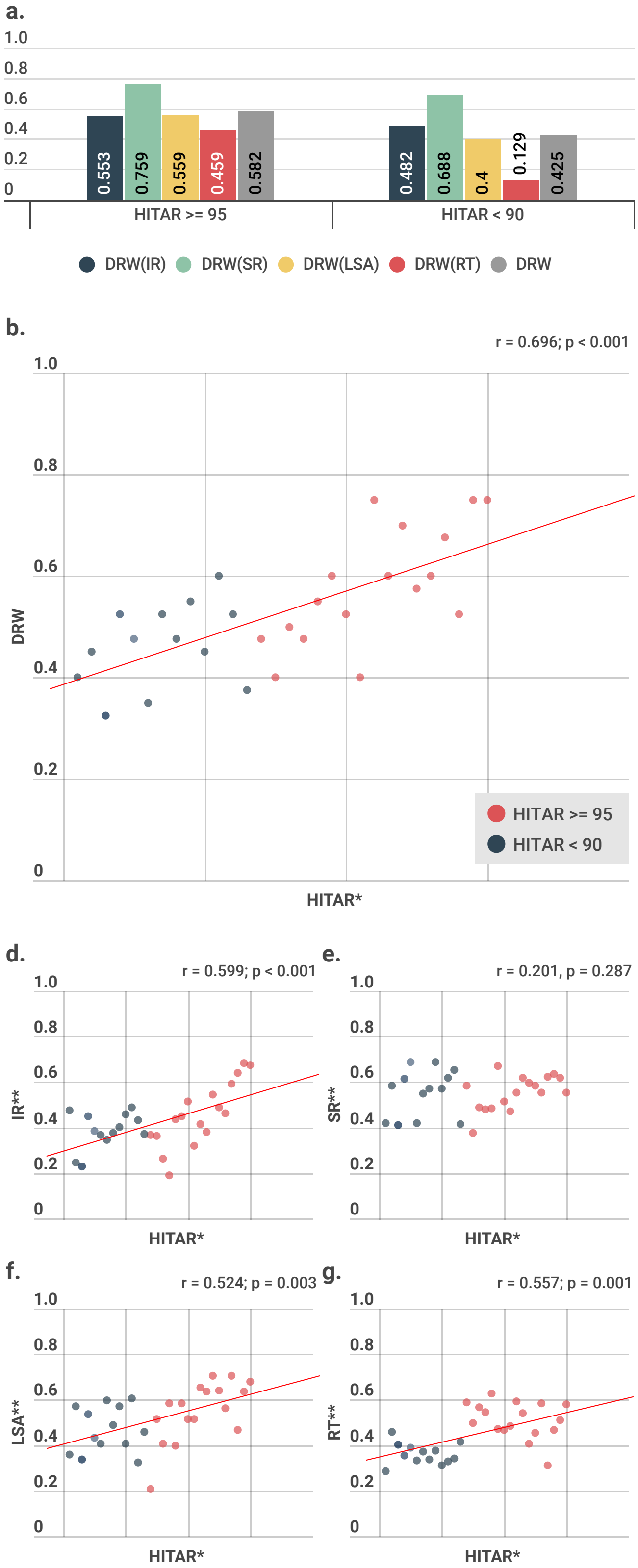


FIGURE 7. Correlation between data reliability weight (DRW) and the Human Intelligent Task Approve Rate (HITAR). (a) Average DRW indexes of worker groups classified by the HITAR; (b) correlation between the HITAR and the DRW; c (d–g) correlation between the HITAR and the DRW indexes. * signifies scaled values and ** normalized values.

TABLE 4. Statistical description and independent samples t-test of individual reliability (IR), long-string analysis (LSA), response time (RT), self-report (SR), and data reliability weight (DRW) of two groups having HITAR values greater than 95 and less than 90.

Index	HITAR condition	Number of participants	Mean	Standard deviation	Cutoff value	t	d.f. (degree of freedom)	p	Cohen's d
DRW	>= 95	170	0.582	0.232	N/A^a	6.009^b	338	<.001	0.652
	<= 90		0.425	0.251					
IR^c	>= 95		0.460	0.272	0.418	2.310	338	0.021	0.251
	<= 90		0.391	0.276					
LSA^c	>= 95		0.553	0.193	0.473	6.690^b	294.743	<.001	0.726
	<= 90		0.375	0.289					
SR^c	>= 95		0.555	0.254	0.577	0.723	338	0.470	0.078
	<= 90		0.535	0.255					
RT^c	>= 95		0.518	0.263	0.452	6.987^b	263.457	<.001	0.758
	<= 90		0.357	0.146					

Note. ^{a} is not applicable. ^{b}: the t-test result was reported on the unequal variances based on the Levene test. ^{c}: the value is normalized by CDF.

the HITAR values. The obtained HITAR values presented high skewness (-2.138), because the HITAR average we obtained had a relatively high value of 83.91. To improve the performance of the data analysis and the interpretation, we established a binning strategy. We thus divided the data group by the same frequency and smoothed the grouped data values to the average value. Specifically, the actual data processing was as follows. We sorted the 340 data items of the participants based on the descending order of their HITAR values. We grouped the data items into groups of 10 as frequency and averaged the HITAR, DRW, IR, SR, LSA, and RT values. Based on these groups, we ranked the HITAR and took the reverse order. Thirty-four participants did not enter the HITAR values correctly, and 6 data items of participants remained ungrouped. Accordingly, we removed 40 data items and divided the remaining 300 data items into 30 groups. To examine the effectiveness of the binning, we used an information value (IV) that expressed the overall predictive power [97]. The measured IV value was greater than 0.3 (0.465), and thus, we determined that the binning was good. Based on the binning, the data obtained showed significant results that indicated a noticeable correlation between the DRW and the HITAR values (Figure 7(b)–(g)). The DRW and HITAR results also present an interesting correlation score (r=0.696, p < 0.001). In summary, we validated that the DRW correlates with the HITAR and thus can be used as a reliability measure.

C. OBSERVATION OF FUTURE REWARD DISTRIBUTION EFFECT

Recent studies have confirmed that systems with immediate virtual rewards positively affect continuous data collection [98], [99]. However, we evaluated the validity of a reward distribution system that gives rewards to participants in the future, rather than immediately. We show the validity as the correlation between rewards and the DRW. Accordingly, we evaluated the following scenario, in which the future reward system of our platform is seen to increase the participation rate.

We created a simulation environment in MTurk to reduce the observation time that exists in the real world. In a real-world study of the reward distribution effects, a significant time would have been required to collect the data and cases. For example, PatientsLikeMe has required approximately five years to collect a sufficient number of cases, which was related to the extent of the site use to show the benefits to communities [30]. Thus, we conducted two tests in the simulation environment. We designed one test in which information about the reward distribution in the future (as in our platform) was provided. In the second test, this information was not provided. The remaining settings were the same for both tests. For the sample size calculation, we conducted a pilot test with 100 participants in MTurk. We obtained an expected mean of the paired differences of 0.11 and an SD of the results of the pilot test of 0.25. Based on the results, we calculated the sample size as 168 in each group to have a power of 95% and a one-sided level of significance of 5% [86]. The total sample size was 336. Then, we used a questionnaire for data reliability validation and weights for consistency comparison, except in the case of the HITAR. However, we needed to consider the casual observation characteristics of MTurk workers, because in this experiment, there was no HITAR constraint [100]. Therefore, we added five questions that induced the participants to enter as many words as possible in the LSA index of the DRW [65].

We recruited 336 workers from MTurk to participate in the test. We allocated 168 workers to the test that contained future reward information (RI condition) and 168 to the test without future reward information (NRI condition). We designed the test to be completed within two hours and recorded the test start and end times for each worker to obtain the RTs. We successfully collected data from 336 workers (Supplementary Data 3). First, we analyzed the effect of RI on the self-reporting rate using a two-way ANOVA test. Table

TABLE 5. Statistical description and independent samples t-test of individual reliability (IR), long-string analysis (LSA), response time (RT), self-report (SR), and data reliability weight (DRW) of groups receiving and not receiving future reward information.

Index	Reward condition	Number of participants	Mean	Standard deviation	Cutoff value	t	d.f. (degree of freedom)	p	Cohen's d
DRW	True	168	0.605	0.291	N/A^a	4.127^b	334	<.001	0.450
	False		0.472	0.301					
IR^c	True		0.492	0.284	0.511	-0.697	334	0.486	0.076
	False		0.513	0.291					
LSA^c	True		0.538	0.231	0.47, 0.47, 0.48, 0.42, 0.49	4.671^b	331.783	<.001	0.509
	False		0.415	0.251					
SR^c	True		0.541	0.211	0.574	0.576	334	0.565	0.063
	False		0.528	0.215					
RT^c	True		0.516	0.246	0.485	3.259	334	0.001	0.356
	False		0.427	0.250					

Note. ^{a}: not applicable. ^{b}: the t-test result was reported on the unequal variances based on Levene's test. ^{c}L: the value is normalized by CDF.

6 indicates that the reward condition showed interaction with consecutive LSA items and is statistically significant at an alpha level of 0.05 (p = 0.008).

TABLE 6. Two-way ANOVA test of the interaction of the reward condition and long-string analysis (LSA)

Source	Sum square	d.f.	Mean square	F-value	p
LSA	30.180	4	7.545	93.083	<.001
Reward	4.688	1	4.688	57.836	<.001
Interaction*	1.119	4	0.280	3.451	0.008

Note. *: the interaction of reward condition and LSA.

In other words, we found that consecutive LSA items under the RI condition consistently contained many characters, and we inferred that the RI condition improves the rate of self-reporting as compared with the NRI condition. Further, we found that the average DRW of the RI condition was 0.605 and that of the NRI condition was 0.472 (Figure 8). Table 5 shows the significant difference between the DRW values of the RI and NRI conditions (p=0.03). Figure 8 presents that DRW(SR), DRW(LSA), and DRW(RT) of the RI condition also have higher average values than those of the NRI condition. Table 5 provides detailed statistical information. The DRW index values in this table are values converted to CDF to calculate the DRW. DRW(IR) showed a low average result in the RI condition. However, we interpreted this as an indication that the effect size (0.076) of the index was too low to affect the DRW results, as shown in Table 5. We presumed that this occurred because the selection of the workers was not controlled by their HITAR. Interesting cases were found for both RT and LSA. Both indexes show high average DRW values and a significant difference, as can be seen in Table 5. In particular, we configured the DRW method to measure RT and LSA only in the data type eCRF, and we could confirm that the configuration was appropriate according to the effect of rewards on continuous data collection. Thus, we concluded that the reward distribution not only increased the DRW but also helped improve the sustainability of the data collection platform.

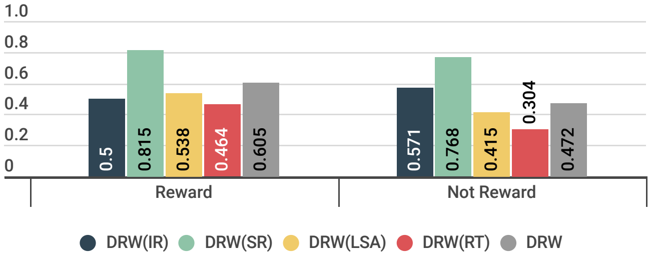


FIGURE 8. Reward distribution effects. Each average score of the data reliability weight (DRW) index for all workers per test.

D. COMPARISON OF OVERALL PLATFORM FUNCTION

We developed "Collaborative research for us" (CORUS, https://corus.kaist.edu) as a pilot data collection platform for participatory trials. The design of CORUS is based on the EDC for existing clinical trials, but it facilitates the design of a study and provides services to allow users to participate voluntarily and continuously in the data collection for the study.

In this section, we analyze the features of CORUS and compare them with those of other systems in detail. For specific comparisons, we selected programs mentioned or used in studies presented in journals related to clinical research from 2012 to 2019, which we described in the introduction section. We summarize the comparison results in Table 7 and provide the details in the following subsections.

1) Comparison of CORUS with clinical trial systems

CORUS includes a protocol creation feature for interpretable data preparation that the public can easily use without the help of experts. In the platform, the creator can utilize the clinical trial protocol database feature according to the semantically related symptom of the effectiveness of healthcare

TABLE 7. Comparison of platform functions and features of participatory, pragmatic, and clinical trials in terms of reliable data collection

Platform Usable Category		**Participatory trial**							
		Pragmatic trial							
		Clinical Trial							
Program Name		Rave EDC [22]	Transparency Life Sciences [23]	TrialChain [101]	PatientsLikeMe [30]	Amazon Mechanical Turk [31]	Google Play Store / Apple App Store [33], [34]	NHS / Rankedhealth [37], [38]	CORUS
Data preparation	Clinical protocol database	○	×	×	×	×	×	×	○
	Protocol creation	△	△	×	×	×	×	×	○
Data storing	Falsification prevention	×	×	○	×	×	×	×	○
	Data reliability weight	×	×	×	×	×	×	×	○
Data sharing	Participation reward	×	×	×	×	○	×	×	○
	Data distribution	○	○	○	○	○	○	○	○
System features	Study management	○	○	○	○	○	×	×	○
	Subject management	○	○	○	○	×	×	×	○
	Clinical monitoring	○	○	○	○	×	×	×	○
	Community	×	×	×	○	×	×	×	○
Target user group		Expert	Expert	Expert	Expert, Public	Expert, Public	Public	Expert	Expert, Public

applications. Moreover, the creator can obtain feedback from the study participants within the community feature, which allows the creator to enhance the protocol in the future.

2) Comparison of CORUS with pragmatic trial systems

The protocol creation feature of CORUS provides interpretable and straightforward data preparation for public users. The system features of CORUS help users lead a study without requiring expert knowledge. Moreover, CORUS has the DRW feature that automatically calculates the reliability scores of collected data to ensure reliable data collection. The scores constitute an objective value that does not include the subjective intention of the study creator. We also developed the scores to evaluate the effectiveness score of applications in further studies of the data storage stage.

3) Comparison of CORUS with other participatory trial systems

We developed CORUS as a participatory trial platform. The data preparation features of CORUS prepare data collection such that immediate data analysis can be performed. Post-processing for unformatted data is not necessary on these features. The features minimize the flawed study design bias [36]. In addition to the DRW feature, the data falsification prevention feature prevents transfer bias by not revealing the collected data to the participants during a study. CORUS, a participant-driven platform, also supports additional analysis for the use of experts. Standardized data sharing is an essential feature for validating the effectiveness of an application. Cryptocurrency is a distinctive participation reward feature of CORUS. We connected CORUS to an external blockchain to encourage participants to share their data continuously. Participants can earn a future reward based on their data input. CORUS calculates the participants' portion of the total cryptocurrency of a study based on the DRW of their input. Thus, the participation reward can induce continuous engagement and reliability.

A common limitation of all the platforms mentioned, except CORUS, is the data falsification problem. The possibility of falsification of the data collected during a study, whether intentional or unintentional, exists in all platforms [102].

TrialChain uses a blockchain to solve the problem [101]. We classify TrialChain as an EDC system based on the explanation of the platform. Thus, we consider that TrialChain is difficult to use in a participatory trial. CORUS also uses blockchain technology. We developed a data falsification prevention feature based on the data immutability characteristics of blockchain technology [74]. The feature solves the problem of data falsification.

VII. CASE STUDY: OPERATIONAL TEST OF THE PILOT PLATFORM

The creators of projects post introductions to their project on the platform and users can select a specific project in which they wish to participate. Thus, they can easily participate in their preferred projects. Participants enrolled in a project can immediately participate in the trial and easily report data. The platform also encourages participants to daily report their data by using features such as leaderboards and alarms.

We conducted an operational test to observe the data collection capabilities of the platform under actual participatory trial conditions. The objective of the trial project used in the test was to confirm that an application that enables blue light filters on smartphones can help improve sleep disorders. Although previous studies showed a relation between blocking blue light from other sources and improvement in certain sleep disorders, to the best of our knowledge, it has not been clinically confirmed that blue light filters on smartphones or other mobile devices have the same effect [103]. The recruitment goal of the trial project was 100 participants.

We designed the participatory trial project and posted it on the pilot platform. In the trial, participants were asked to apply the blue light filter, which blocks blue wavelength light, on their smartphones and report changes in the quality of their sleep. The data that were collected showed no evidence of the effectiveness of the application.

In addition to validating technical issues, we also validated that the system could effectively recruit participants for the trial. The scheduled duration of participant recruitment was one month; however, the actual rate of recruitment was faster. In total, 100 participants were recruited in 21 days. We compared the rate of recruitment in this experiment with those found in previous studies. To represent the rate of recruitment among trials, the recruitment rate is defined as the number of participants recruited per month (or participants per center per month, in the case of multicenter trials). In 151 traditional clinical trials supported by the United Kingdom's Health Technology Assessment program, the recruitment rate was 0.92 [104]. According to the reports of 8 Web- or mobile application-based studies collected from literature databases, an average of 468 participants were recruited during 5 months (recruitment rate = 93.6) [105]. For the platform, the recruitment rate was 142.8. This shows that participant recruitment was more effective when the platform was used than that achieved in traditional clinical trials; it was also competitive with that achieved in other Web- or mobile application-based studies. The response rate also tended to increase over time after the beginning of the project period. The response rate increased to 120% on recruitment days 17 and 18, when many new participants were enrolled, and remained high for a significant time period (Figure 9).

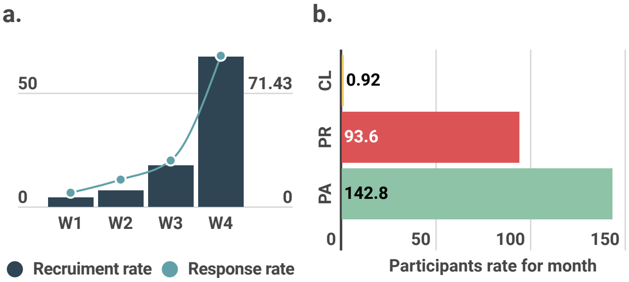


FIGURE 9. Recruitment rate and comparison analysis. (a) Recruitment and response rate during one month in the platform trial project; (b) participant rates among clinical trials (CLs), pragmatic trials (PRs), and participant trials (PAs) based on the evidence presented in the literature

VIII. DISCUSSION

Our method and platform are of technical value in that they constitute a combination of various independent and unrelated methods. We integrated a clinical trial protocol and database, semantic correlation calculations, blockchain, survey data evaluation, intrinsic and extrinsic motivation of human behavior, cloud computing technology, and software engineering methods into a participatory trial platform to assess the effectiveness of healthcare applications. Although many types of participatory trials exist, such as crowdsourcing and crowd-science, these approaches still have critical limitations for unstructured data collection and data sharing. We suggested a new participatory trial concept, where the beneficial characteristics of crowd-science and crowdsourcing are combined. Then, we proposed a solution to solve problems in the interpretable data preparation, systematic data storage, and sustainable data sharing stages. Our method and platform are of great value, because they present a primary technique that can be used to validate the effectiveness of various healthcare applications.

The minimization of bias that occurs intentionally or unintentionally at all stages of data collection is an additional important factor in ensuring data reliability. We identified and considered each source of bias in the data collection stages according to the relevant research studies [36], [106]. At the data preparation stage, we addressed the problems of flawed study design, channeling bias, and selection bias. We attempted to prevent flawed study design by using the existing clinical trial protocols. We minimized channeling bias, which leads to the inclusion of only specific participants, by applying the unconditional eligibility criterion that anyone can participate in a study. The editable eligibility criteria of a creator, such as gender and age, still need to be considered to resolve the channeling bias issue. However, the selection bias, which can distort the data tendency, was not considered. This is a difficult problem to solve in the field of Web-based data collection [107]. As a solution, we propose a social relationship distance estimation method to maintain a certain distance between participants. A social network estimation method of participants to calculate the closeness rank among participants is an example method for excluding participants having a certain distance or less between them [108]. At the data storage and data sharing stages, we included interviewer bias and transfer bias. The interviewer bias refers to the potential influence on the participants derived from the data collector. Our data storage method involves only mechanical data collection functions, and therefore, we assumed that interviewer bias did not occur. However, we do not include a preventive measure against interviewer bias caused by the study description or the eCRF items that a creator manually generates. The solution is to clearly define the digital biomarkers that a creator seeks to evaluate for determining the effectiveness of a digital healthcare application [109]. Further research and the consensus of experts are required to enable standardized participatory trial protocols with digital biomarkers. A transfer bias could be caused by data exposure at the data storage stage and by the possible influence exerted by the stored data on other participants. We avoided this source of bias by preventing exposure of the stored data during a study. Finally, we did not consider the bias generated at the data analysis stage, because it is outside the scope of this study, which considers only data collection methods.

There are additional considerations that may improve our platform. First, the current platform cannot provide a quantitative comparison of the search results of the clinical trial protocol. The platform takes the symptom provided by the creator and returns all the relevant clinical trials. However, many comparable clinical trials are provided as a search result, and additional analysis by the creator is required. In clinical trials, various types of data—not only symptoms but also genes and chemical compounds—appear. This information can be used to retrieve the clinical trial data required by the user. If an advanced platform is developed in the future that can quantitatively assess the relevance of all data involved in the clinical trial, it will reduce the additional analysis burden on the creator. Second, advanced DRW methods should be developed. For example, the basic DRW is used to assess the reliability of the input data based only on the length of the input text, as in LSA. In other words, it cannot consider semantic errors in the input text. Therefore, an advanced method that considers semantic errors in the text is required. In addition, we face a cutoff value calculation problem when the SD is 0 from the collected data of a DRW index. We solved the problem by using a predefined minimum cutoff value on the platform; however, we need to improve this approach so that we can actively determine the cutoff values that are suitable for the collected data. Third, we will devise an optimal strategy to reduce the cost of running the platform. Fourth, an additional study is required to verify the platform. We plan to conduct a small-scale pilot study to examine whether the study results are the same as the verified effects.

All these limitations should be considered for further experiments or improved computational methods. Moreover, we suggest further studies, because data collection methods are the basics of the digital healthcare field, which is currently progressing exponentially. The suggested study objectives are as follows. (1) The development of a standard for a participatory trial protocol method that considers digital markers and confounding factors. (2) The advancement of protocol creation methods for the determination of specific parameters, such as the number of use, effective report counts, study duration, or the number of participants. (3) The use of natural language question models to generate DRW index questions instead of human-managed templates. (4) Optimization of the DRW calculation with the importance value of each index. (5) Implementation of an effectiveness score calculation method based on the collected data with DRW scores. These further improvements and future studies would allow our platform to be used as a valuable tool to assess the effectiveness of various applications in a cost-effective manner.

IX. CONCLUSION

As the first stage of the verification workflow to assess the effectiveness of digital healthcare applications, we proposed a reliable data collection method and platform. We presented a participatory trial concept for a new type of data collection trial that uses the voluntary participation of end users of digital healthcare applications. Then, we described the essential methods of the reliable data collection platform based on the participatory trial concept. Our interpretable data preparation methods consist of a protocol database and a retrieval system to allow a person to create a protocol without expert knowledge. We validated the simplicity of the methods by comparing the USE scores of the proposed system and the existing system. We developed a DRW calculation method for systematic data storage. The DRW score was shown to be a reliable measure through its correlation with the

HITAR. To achieve sustainable data collection and sharing, we developed a reward distribution method. We indirectly observed the effect of rewards using the DRW. The effect of rewards presented as an increased DRW; i.e., it increased the participants' effort and thus achieved sustainable data collection. Finally, we implemented a pilot platform, CORUS, which integrated all the methods and essential features for a participatory trial platform. We compared its features with those of existing platforms in the field of clinical, pragmatic, and participatory trials. We assert that CORUS has all the necessary features to collect reliable data from the users of digital healthcare applications on the path to the validation of their effectiveness.

CODE AVAILABILITY

We will provide all the source code at https://github.com/bisllaboratory/corus upon publication. The platform is open-source to promote development in the public domain. The repository describes software requirements and distributes a README.MD file.

ACKNOWLEDGMENT

We thank Jooyeon Lee and Inhwan Hwang of Amazon Web Services Korea LLC (AWS Korea) Worldwide Public Sector for the support they provided through cloud computing platform services and demonstrations.

REFERENCES

[1] N. DANBURY, Conn. & RESEARCH TRIANGLE PARK. (2017) IQVIA Institute for Human Data Science Study: Impact of Digital Health Grows as Innovation, Evidence, and Adoption of Mobile Health Apps Accelerate. [Online]. Available: https://www.iqvia.com/newsroom/2017/11/impact-of-digital-health-grows-as-innovation-evidence-and-adoption-of-mobile-health-apps-accelerate/

[2] M. Z. Rachana Jain. (2017) 2017 Year End Funding Report: The end of the beginning of digital health. [Online]. Available: https://rockhealth.com/reports/2017-year-end-funding-report-the-end-of-the-beginning-of-digital-health/

[3] S. T. Karen Sharma. (2017) Pear Therapeutics Receives Expedited Access Pathway Designation from FDA for reSET-O™ Prescription Digital Therapeutic to Treat Opioid Use Disorder. [Online]. Available: https://www.businesswire.com/news/home/20171018006174/en/Pear-Therapeutics-Receives-ExpeditedAccess-Pathway-Designation

[4] J. A. Anguera, A. N. Brandes-Aitken, A. D. Antovich, C. E. Rolle, S. S. Desai, and E. J. Marco, "A pilot study to determine the feasibility of enhancing cognitive abilities in children with sensory processing dysfunction," PloS one, vol. 12, no. 4, p. e0172616, 2017.

[5] B. E. Yerys, J. R. Bertollo, L. Kenworthy, G. Dawson, E. J. Marco, R. T. Schultz, and L. Sikich, "Brief Report: Pilot Study of a Novel Interactive Digital Treatment to Improve Cognitive Control in Children with Autism Spectrum Disorder and Co-occurring ADHD Symptoms," Journal of autism and developmental disorders, vol. 49, no. 4, pp. 1727–1737, 2019.

[6] S. C. Sepah, L. Jiang, and A. L. Peters, "Long-term outcomes of a Web-based diabetes prevention program: 2-year results of a single-arm longitudinal study," Journal of Medical Internet research, vol. 17, no. 4, p. e92, 2015.

[7] E. Waltz, "Pear approval signals FDA readiness for digital treatments," Nature Biotechnology, vol. 36, no. 6, pp. 481–482, 2018. [Online]. Available: https://doi.org/10.1038/nbt0618-481

[8] L. Kogan, P. Hellyer, C. Duncan, and R. Schoenfeld-Tacher, "A pilot investigation of the physical and psychological benefits of playing Pokémon GO for dog owners," Computers in Human Behavior, vol. 76, pp. 431–437, 2017.

[9] C. Noone and M. J. Hogan, "A randomised active-controlled trial to examine the effects of an online mindfulness intervention on executive control, critical thinking and key thinking dispositions in a university student sample," BMC psychology, vol. 6, no. 1, p. 13, 2018.

[10] W. P. Jayawardene, D. K. Lohrmann, R. G. Erbe, and M. R. Torabi, "Effects of preventive online mindfulness interventions on stress and mindfulness: A meta-analysis of randomized controlled trials," Preventive medicine reports, vol. 5, pp. 150–159, 2017.

[11] S. C. Mathews, M. J. McShea, C. L. Hanley, A. Ravitz, A. B. Labrique, and A. B. Cohen, "Digital health: a path to validation," NPJ digital medicine, vol. 2, no. 1, p. 38, 2019.

[12] S. R. Stoyanov, L. Hides, D. J. Kavanagh, O. Zelenko, D. Tjondronegoro, and M. Mani, "Mobile app rating scale: a new tool for assessing the quality of health mobile apps," JMIR mHealth and uHealth, vol. 3, no. 1, p. e27, 2015.

[13] NHS. (2018) NHS Apps Library. [Online]. Available: https://apps.beta.nhs.uk

[14] F. T. Commission et al., "Mobile Health Apps Interactive Tool," 2016. [Online]. Available: https://www.ftc.gov/tips-advice/business-center/guidance/mobile-health-apps-interactive-tool

[15] T. T. Lee and A. S. Kesselheim, "US Food and Drug Administration precertification pilot program for digital health software: weighing the benefits and risks," Annals of internal medicine, 2018.

[16] R. Health. (2018) Curated Health Apps and Devices With a Focus on Clinical Relevance, Safety, and Efficacy. [Online]. Available: http://www.rankedhealth.com

[17] V. R. Basili, "Data collection, validation and analysis," Software Metrics: An Analysis and Evaluation, pp. 143–160, 1981.

[18] R. E. Sherman, S. A. Anderson, G. J. Dal Pan, G. W. Gray, T. Gross, N. L. Hunter, L. LaVange, D. Marinac-Dabic, P. W. Marks, M. A. Robb et al., "Real-world evidence—what is it and what can it tell us," N Engl J Med, vol. 375, no. 23, pp. 2293–2297, 2016.

[19] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," Applied artificial intelligence, vol. 17, no. 5-6, pp. 375–381, 2003.

[20] L. G. Portney, M. P. Watkins et al., Foundations of clinical research: applications to practice. Pearson/Prentice Hall Upper Saddle River, NJ, 2009, vol. 892.

[21] NHS. (2020) Business Software and Services Reviews. [Online]. Available: https://www.g2.com/categories/clinical-trial-management?order=popular#product-list

[22] K. A. Getz and R. A. Campo, "New benchmarks characterizing growth in protocol design complexity," Therapeutic innovation & regulatory science, vol. 52, no. 1, pp. 22–28, 2018.

[23] A. Leiter, T. Sablinski, M. Diefenbach, M. Foster, A. Greenberg, J. Holland, W. K. Oh, and M. D. Galsky, "Use of crowdsourcing for cancer clinical trial development," JNCI: Journal of the National Cancer Institute, vol. 106, no. 10, 2014.

[24] R. Tonkens, "An overview of the drug development process," Physician executive, vol. 31, no. 3, p. 48, 2005.

[25] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: new estimates of R&D costs," Journal of health economics, vol. 47, pp. 20–33, 2016.

[26] I. Ford and J. Norrie, "Pragmatic trials," New England journal of medicine, vol. 375, no. 5, pp. 454–463, 2016.

[27] C. Celis-Morales, K. M. Livingstone, C. F. Marsaux, H. Forster, C. B. O'Donovan, C. Woolhead, A. L. Macready, R. Fallaize, S. Navas-Carretero, R. San-Cristobal et al., "Design and baseline characteristics of the Food4Me study: a web-based randomised controlled trial of personalised nutrition in seven European countries," Genes & nutrition, vol. 10, no. 1, p. 450, 2015.

[28] R. Campbell, F. Starkey, J. Holliday, S. Audrey, M. Bloor, N. Parry-Langdon, R. Hughes, and L. Moore, "An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial," The Lancet, vol. 371, no. 9624, pp. 1595–1602, 2008.

[29] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey et al., "The mPower study, Parkinson disease mobile data collected using ResearchKit," Scientific data, vol. 3, p. 160011, 2016.

[30] P. Wicks, M. Massagli, J. Frost, C. Brownstein, S. Okun, T. Vaughan, R. Bradley, and J. Heywood, "Sharing health data for better outcomes on PatientsLikeMe," Journal of medical Internet research, vol. 12, no. 2, p. e19, 2010.

[31] G. Paolacci and J. Chandler, "Inside the Turk: Understanding Mechanical Turk as a participant pool," Current Directions in Psychological Science, vol. 23, no. 3, pp. 184–188, 2014.

[32] M. F. Mendiola, M. Kalnicki, and S. Lindenauer, "Valuable features in mobile health apps for patients and consumers: content analysis of apps and user ratings," JMIR mHealth and uHealth, vol. 3, no. 2, p. e40, 2015.

[33] Google. (2012) Google Play. [Online]. Available: https://play.google.com/store

[34] Apple. (2008) App Store. [Online]. Available: https://www.apple.com/ios/app-store/

[35] W. Maalej, Z. Kurtanović, H. Nabil, and C. Stanik, "On the automatic classification of app reviews," Requirements Engineering, vol. 21, no. 3, pp. 311–331, 2016.

[36] C. J. Pannucci and E. G. Wilkins, "Identifying and avoiding bias in research," Plastic and reconstructive surgery, vol. 126, no. 2, p. 619, 2010.

[37] S. D. Halpern, M. O. Harhay, K. Saulsgiver, C. Brophy, A. B. Troxel, and K. G. Volpp, "A pragmatic trial of e-cigarettes, incentives, and drugs for smoking cessation," New England Journal of Medicine, vol. 378, no. 24, pp. 2302–2310, 2018.

[38] K. Singh, K. Drouin, L. P. Newmark, R. Rozenblum, J. Lee, A. Landman, E. Pabo, E. V. Klinger, and D. W. Bates, "Developing a framework for evaluating the patient engagement, quality, and safety of mobile health applications," Issue Brief (Commonw Fund), vol. 5, no. 1, p. 11, 2016.

[39] (2019) How the assessment works. [Online]. Available: https://digital.nhs.uk/services/nhs-apps-library/guidance-for-health-app-developers-commissioners-and-assessors/how-we-assess-health-apps-and-digital-tools#how-the-assessment-works

[40] R. N. Jamison, S. A. Raymond, J. G. Levine, E. A. Slawsby, S. S. Nedeljkovic, and N. P. Katz, "Electronic diaries for monitoring chronic pain: 1-year validation study," Pain, vol. 91, no. 3, pp. 277–285, 2001.

[41] J. Y. Kim, N. E. Wineinger, M. Taitel, J. M. Radin, O. Akinbosoye, J. Jiang, N. Nikzad, G. Orr, E. Topol, and S. Steinhubl, "Self-monitoring utilization patterns among individuals in an incentivized program for healthy behaviors," Journal of medical Internet research, vol. 18, no. 11, p. e292, 2016.

[42] S. Taylor, C. Ferguson, F. Peng, M. Schoeneich, and R. W. Picard, "Use of In-Game Rewards to Motivate Daily Self-Report Compliance: Randomized Controlled Trial," Journal of medical Internet research, vol. 21, no. 1, p. e11683, 2019.

[43] E. Elenko, A. Speier, and D. Zohar, "A regulatory framework emerges for digital medicine," Nature biotechnology, vol. 33, no. 7, p. 697, 2015.

[44] L. E. Kruger and M. A. Shannon, "Getting to know ourselves and our places through participation in civic social assessment," Society & Natural Resources, vol. 13, no. 5, pp. 461–478, 2000.

[45] J. Howe, "The rise of crowdsourcing," Wired magazine, vol. 14, no. 6, pp. 1–4, 2006.

[46] Study design: Pragmatic trial, author=Mira Zuidgeest,Iris Goetz,Rick Grobbee. [Online]. Available: https://rwe-navigator.eu/use-real-world-evidence/generate-real-world-evidence/study-design-pragmatic-trials/

[47] L. E. Bothwell, J. A. Greene, S. H. Podolsky, and D. S. Jones, "Assessing the Gold Standard — Lessons from the History of RCTs," New England Journal of Medicine, vol. 374, no. 22, pp. 2175–2181, 2016, pMID: 27248626. [Online]. Available: https://doi.org/10.1056/NEJMms1604593

[48] B. Walther, S. Hossin, J. Townend, N. Abernethy, D. Parker, and D. Jeffries, "Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data," PloS one, vol. 6, no. 9, p. e25348, 2011.

[49] D. Vinereanu, R. D. Lopes, M. C. Bahit, D. Xavier, J. Jiang, H. R. Al-Khalidi, W. He, Y. Xian, A. O. Ciobanu, D. Y. Kamath et al., "A multifaceted intervention to improve treatment with oral anticoagulants in atrial fibrillation (IMPACT-AF): an international, cluster-randomised trial," The Lancet, vol. 390, no. 10104, pp. 1737–1746, 2017.

[50] P. M. Rothwell, "Factors that can affect the external validity of randomised controlled trials," PLoS clinical trials, vol. 1, no. 1, p. e9, 2006.

[51] K. E. Thorpe, M. Zwarenstein, A. D. Oxman, S. Treweek, C. D. Furberg, D. G. Altman, S. Tunis, E. Bergel, I. Harvey, D. J. Magid et al., "A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers," Journal of clinical epidemiology, vol. 62, no. 5, pp. 464–475, 2009.

[52] N. Sepehrvand, W. Alemayehu, D. Das, A. K. Gupta, P. Gouda, A. Ghimire, A. X. Du, S. Hatami, H. E. Babadagli, S. Verma et al., "Trends in the explanatory or pragmatic nature of cardiovascular clinical trials over 2 decades," JAMA cardiology, 2019.

[53] A. J. Apter, "Understanding adherence requires pragmatic trials: lessons from pediatric asthma," JAMA pediatrics, vol. 169, no. 4, pp. 310–311, 2015.

[54] H. C. Sox and R. J. Lewis, "Pragmatic trials: practical answers to "real world" questions," Jama, vol. 316, no. 11, pp. 1205–1206, 2016.

[55] M. Trusheim, A. A. Shrier, Z. Antonijevic, R. A. Beckman, R. K. Campbell, C. Chen, K. Flaherty, J. Loewy, D. Lacombe, S. Madhavan et al., "PIPELINEs: creating comparable clinical knowledge efficiently by linking trial platforms," Clinical Pharmacology & Therapeutics, vol. 100, no. 6, pp. 713–729, 2016.

[56] A.-W. Chan, J. M. Tetzlaff, D. G. Altman, A. Laupacis, P. C. Gøtzsche, K. Krleža-Jerić, A. Hróbjartsson, H. Mann, K. Dickersin, J. A. Berlin et al., "SPIRIT 2013 statement: defining standard protocol items for clinical trials," Annals of internal medicine, vol. 158, no. 3, pp. 200–207, 2013.

[57] J. Park, K. Kim, W. Hwang, and D. Lee, "Concept embedding to measure semantic relatedness for biomedical information ontologies," Journal of biomedical informatics, vol. 94, p. 103182, 2019.

[58] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in International conference on machine learning, 2014, pp. 1188–1196.

[59] J. Park, K. Kim, S. Park, W. Hwang, S. Yoo, G.-s. Yi, and D. Lee, "An interactive retrieval system for clinical trial studies with context-dependent protocol elements," bioRxiv, p. 814996, 2019.

[60] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2008, pp. 453–456.

[61] N. Kaufmann, T. Schulze, and D. Veit, "More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk." in AMCIS, vol. 11, no. 2011. Detroit, Michigan, USA, 2011, pp. 1–11.

[62] M. Li, J. Weng, A. Yang, W. Lu, Y. Zhang, L. Hou, J.-N. Liu, Y. Xiang, and R. H. Deng, "CrowdBC: A blockchain-based decentralized framework for crowdsourcing," IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 6, pp. 1251–1266, 2018.

[63] C. Eickhoff and A. P. de Vries, "Increasing cheat robustness of crowdsourcing tasks," Information retrieval, vol. 16, no. 2, pp. 121–137, 2013.

[64] R. E. McGrath, M. Mitchell, B. H. Kim, and L. Hough, "Evidence for response bias as a source of error variance in applied assessment." Psychological bulletin, vol. 136, no. 3, p. 450, 2010.

[65] J. L. Huang, P. G. Curran, J. Keeney, E. M. Poposki, and R. P. DeShon, "Detecting and deterring insufficient effort responding to surveys," Journal of Business and Psychology, vol. 27, no. 1, pp. 99–114, 2012.

[66] S. L. Wise and X. Kong, "Response time effort: A new measure of examinee motivation in computer-based tests," Applied Measurement in Education, vol. 18, no. 2, pp. 163–183, 2005.

[67] P. T. Costa Jr and R. R. McCrae, The Revised NEO Personality Inventory (NEO-PI-R). Sage Publications, Inc, 2008.

[68] M.-G. Seo and L. F. Barrett, "Being emotional during decision making—good or bad? An empirical investigation," Academy of Management Journal, vol. 50, no. 4, pp. 923–940, 2007.

[69] J. L. Huang, N. A. Bowling, M. Liu, and Y. Li, "Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions," Journal of Business and Psychology, vol. 30, no. 2, pp. 299–311, 2015.

[70] A. W. Meade and S. B. Craig, "Identifying careless responses in survey data." Psychological methods, vol. 17, no. 3, p. 437, 2012.

[71] F. Galton, "Regression towards mediocrity in hereditary stature." The Journal of the Anthropological Institute of Great Britain and Ireland, vol. 15, pp. 246–263, 1886.

[72] G. Casella and R. L. Berger, Statistical inference. Duxbury Pacific Grove, CA, 2002, vol. 2.

[73] D. E. Gaalema, R. J. Elliott, P. D. Savage, J. L. Rengo, A. Y. Cutler, I. Pericot-Valverde, J. S. Priest, D. S. Shepard, S. T. Higgins, and P. A. Ades, "Financial incentives to increase cardiac rehabilitation participation among low-socioeconomic status patients: a randomized clinical trial," JACC: Heart Failure, 2019.

[74] J. Park, S. Park, K. Kim, and D. Lee, "CORUS: Blockchain-Based Trustworthy Evaluation System for Efficacy of Healthcare Remedies," in 2018 IEEE International Conference on Cloud Computing Technology and Science (CloudCom). IEEE, 2018, pp. 181–184.

[75] M. Kim and J. Chung, "Sustainable Growth and Token Economy Design: The Case of Steemit," Sustainability, vol. 11, no. 1, p. 167, 2019.

[76] V. H. Vroom, Some personality determinants of the effects of participation. Routledge, 2019.

[77] J. S. Katz and B. R. Martin, "What is research collaboration?" Research policy, vol. 26, no. 1, pp. 1–18, 1997.

[78] R. J. Winter, "Agile Software Development: Principles, Patterns, and Practices: Robert C. Martin with contributions by James W. Newkirk and Robert S. Koss," Performance Improvement, vol. 53, no. 4, pp. 43–46, 2014.

[79] M. Cavelaars, J. Rousseau, C. Parlayan, S. de Ridder, A. Verburg, R. Ross, G. R. Visser, A. Rotte, R. Azevedo, J.-W. Boiten et al., "OpenClinica," in Journal of clinical bioinformatics, vol. 5, no. S1. Springer, 2015, p. S2.

[80] S. Gessner, M. Storck, S. Hegselmann, M. Dugas, and I. Soto-Rey, "Automated Transformation of CDISC ODM to OpenClinica." in GMDS, 2017, pp. 95–99.

[81] T. J. Brix, P. Bruland, S. Sarfraz, J. Ernsting, P. Neuhaus, M. Storck, J. Doods, S. Ständer, and M. Dugas, "ODM Data Analysis—A tool for the automatic validation, monitoring and generation of generic descriptive statistics of patient data," PloS one, vol. 13, no. 6, p. e0199242, 2018.

[82] "Registering a clinical trial in ClinicalTrials. gov, author=Zarin, Deborah A and Keselman, Alla," Chest, vol. 131, no. 3, pp. 909–912, 2007.

[83] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu, "Accomplishments and challenges in literature data mining for biology," Bioinformatics, vol. 18, no. 12, pp. 1553–1561, 2002.

[84] A. M. Lund, "Measuring usability with the use questionnaire," Usability interface, vol. 8, no. 2, pp. 3–6, 2001.

[85] M. Gao, P. Kortum, and F. Oswald, "Psychometric evaluation of the USE (Usefulness, Satisfaction, and Ease of use) questionnaire for reliability and validity," in Proceedings of the human factors and ergonomics society annual meeting, vol. 62, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2018, pp. 1414–1418.

[86] N. Dhand and M. Khatkar, "Statulator: An online statistical calculator. Sample size calculator for estimating a single mean," 2014. [Online]. Available: http://statulator.com/SampleSize/ss1M.html

[87] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?" Methodological issues and strategies in clinical research, pp. 133–139, 2016.

[88] A. E. C. Cloud, "Amazon web services," Retrieved November, vol. 9, no. 2011, p. 2011, 2011.

[89] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," Biometrika, vol. 52, no. 3/4, pp. 591–611, 1965.

[90] F. Wilcoxon, "Individual comparisons by ranking methods," in Breakthroughs in statistics. Springer, 1992, pp. 196–202.

[91] J. A. Johnson, "Ascertaining the validity of individual protocols from web-based personality inventories," Journal of research in personality, vol. 39, no. 1, pp. 103–129, 2005.

[92] D. L. Paulhus, P. D. Harms, M. N. Bruce, and D. C. Lysy, "The overclaiming technique: Measuring self-enhancement independent of ability." Journal of personality and social psychology, vol. 84, no. 4, p. 890, 2003.

[93] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. P. Bigham, "A data-driven analysis of workers' earnings on amazon mechanical turk," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 2018, p. 449.

[94] H. Levene, "Robust tests for equality of variances. Ingram Olkin, Harold Hotelling, et alia. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling," Stanford University, pp. 278–292, 1960.

[95] K. O. McGraw and S. Wong, "A common language effect size statistic." Psychological bulletin, vol. 111, no. 2, p. 361, 1992.

[96] J. Cohen, Statistical power analysis for the behavioral sciences. Routledge, 2013.

[97] C. E. Shannon, "A mathematical theory of communication," Bell system technical journal, vol. 27, no. 3, pp. 379–423, 1948.

[98] M. Mitchell, L. White, P. Oh, D. Alter, T. Leahey, M. Kwan, and G. Faulkner, "Uptake of an incentive-based mHealth app: process evaluation of the Carrot Rewards app," JMIR mHealth and uHealth, vol. 5, no. 5, p. e70, 2017.

[99] K. Plangger, C. Campbell, K. Robson, and M. Montecchi, "Little rewards, big changes: Using exercise analytics to motivate sustainable changes in physical activity," Information & Management, p. 103216, 2019.

[100] M. S. Aruguete, H. Huynh, B. L. Browne, B. Jurs, E. Flint, and L. E. McCutcheon, "How serious is the 'carelessness' problem on Mechanical Turk?" International Journal of Social Research Methodology, vol. 22, no. 5, pp. 441–449, 2019.

[101] D. R. Wong, S. Bhattacharya, and A. J. Butte, "Prototype of running clinical trials in an untrustworthy environment using blockchain," Nature communications, vol. 10, no. 1, p. 917, 2019.

[102] S. L. George and M. Buyse, "Data fraud in clinical trials," Clinical investigation, vol. 5, no. 2, p. 161, 2015.

[103] Y. Esaki, T. Kitajima, Y. Ito, S. Koike, Y. Nakao, A. Tsuchiya, M. Hirose, and N. Iwata, "Wearing blue light-blocking glasses in the evening advances circadian rhythms in the patients with delayed sleep phase disorder: An open-label trial," Chronobiology international, vol. 33, no. 8, pp. 1037–1044, 2016.

[104] S. J. Walters, I. B. dos Anjos Henriques-Cadby, O. Bortolami, L. Flight, D. Hind, R. M. Jacques, C. Knox, B. Nadin, J. Rothwell, M. Surtees et al., "Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom Health Technology Assessment Programme," BMJ open, vol. 7, no. 3, p. e015276, 2017.

[105] T. S. Lane, J. Armin, and J. S. Gordon, "Online recruitment methods for web-based and mobile health studies: a review of the literature," Journal of medical Internet research, vol. 17, no. 7, p. e183, 2015.

[106] A.-M. Simundic, "Bias in research," Biochemia medica: Biochemia medica, vol. 23, no. 1, pp. 12–15, 2013.

[107] J. Bethlehem, "Selection bias in web surveys," International Statistical Review, vol. 78, no. 2, pp. 161–188, 2010.

[108] A. Saxena, R. Gera, and S. Iyengar, "A heuristic approach to estimate nodes' closeness rank using the properties of real world networks," Social Network Analysis and Mining, vol. 9, no. 1, p. 3, 2019.

[109] A. Coravos, S. Khozin, and K. D. Mandl, "Developing and adopting safe and effective digital biomarkers to improve patient outcomes," NPJ digital medicine, vol. 2, no. 1, pp. 1–5, 2019.