Reliable data collection in participatory trials to assess digital healthcare apps.

Junseok Park[1,2], Seongkuk Park[3], Kwangmin Kim[1,2], Gwangmin Kim[1,2], Jaegyun Jung[2,4], Sungyong Yoo[5], Gwan-su Yi[1], 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

*** 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

Digital healthcare mobile apps are experiencing exponential growth due to the development of the mobile network and widespread usage of smartphones. Unlike high growth rates, sufficient validated apps are few. Because mobile health apps are low-risk technology, and it tends to reduce regulatory oversight. With this tendency, the apps have a direct trade characteristic between developers and end-users. The existing platforms are not suitable to collect reliable data for evaluating the effectiveness of the apps. Moreover, these platforms only reflect the perspectives of developers and experts, not end-users. For instance, the methods of data collection in clinical trials are not appropriate for participant-oriented assessment of healthcare apps since its complexity and high cost. Thus, we identified a need for a participant-oriented data collection platform for end-users as an interpretable, systematic, and sustainable tool—as a first step to validate the effectiveness of the apps. To collect reliable data in the participatory trial format, we have defined data preparation, storage, and sharing stage. Interpretable data preparation consists of a protocol database system and semantic feature retrieval method to create a protocol without professional knowledge. Collected data reliability weight calculation belongs to the systematic data storage stage. To sustainable data collection, we integrate the weight method and reward distribution function. We validate the methods with 718 of human participants to conduct statistical tests. The validation results showed that the methods have significant differences in the comparative experiment, and we convinced the methods are essential to reliable data collection. Furthermore, we created a web-based system for our pilot platform to collect reliable data in an integrated pipeline. We validate the platform features with existing clinical and pragmatic trial data collection platforms. In conclusion, we show that the method and platform support reliable data collection, forging a path to effectiveness validation of digital healthcare apps.

1 Introduction

Digital healthcare has become one of the most promising fields in the healthcare industry with the widespread popularity of wireless devices such as smartphones. Notably, approximately 200 mobile health apps are listed daily in mobile app stores, and investments in digital health are booming [25, 89]. While many of the apps are general wellness applications that help users to manage their activities, some healthcare apps claim to improve a symptom or disease directly. The treatments proposed by the apps take the form of information or visual stimuli, which are effectively delivered to users as digital therapy. reSET-O, an 84-day prescription digital therapeutic app for the treatment of opioid use disorder, has been approved by the U.S. Food and Drug Administration (FDA) [51]. Akili Interactive has clinically demonstrated that interactive digital treatment through a video game, which is under review by the FDA, may improve symptoms of attention deficit hyperactivity disorder (ADHD) and sensory processing disorder [4, 119]. Omada Health showed that a web-based diabetes prevention program could significantly lower hemoglobin A1c levels. Moreover, the web-based program showed a higher rate of patient engagement than the traditional in-person diabetes prevention program [94]. They have received tremendous attention from researchers and the public as digital medicine [113].

The increased interest in digital health has led researchers to study the effectiveness of current apps. As a practical example, Pokémon GO, which is a game app with augmented reality, has positive effects on social interaction: 43.2% of people spent more time with their family [57]. In contrast, 6 weeks of an online intervention experiment has revealed that Headspace, which is a healthcare app for mindfulness by guided meditation, has relatively small effects on mindfulness [79]. The outcome contradicts the results of previous randomized controlled trials for the app [49]. Because of such contradictions, there is a need for a reliable mechanism to identify validated digital health apps [69].

Currently, there is no platform that evaluates the effectiveness of an app using a study based on a systematic and objective validation method in digital health [69]. In particular, mobile health apps have a very fast development cycle and are low-risk technology, which tends to reduce the burden of regulation. With this circumstance, the apps have a direct trade characteristic between developers and end-users. However, the few existing platforms are not suitable to evaluate the effectiveness of the apps in this case. For example, review platforms and guidelines are emerging to solve the problem, but they only provide limited effectiveness information obtained from the perspectives of developers, experts, and regulatory authorities [22, 61, 77, 100]. Thus, we found a need for a participant-oriented data collection platform for end-users as an interpretable, systematic, and sustainable tool to validate the effectiveness of the apps.

The first step is a reliable data collection in a path to validation. The validation method begins with a collection of reliable data; next, the data analysis is performed for verification [8]. Hence, data collection is the basis of data analysis. The utilization conditions of healthcare apps are different from an ideal research-intensive environment, and the collected data relate to the term "real-world data" [98]. An incomplete collection of real-word data can cause inaccurate data analysis due to inconsistency, poor data quality, and noisy or missing data [121]. Thus, the data collection is an essential stage of the verification workflow. We compare previous expert-oriented data collection platforms in existing research fields to develop a participant-oriented data collection platform.

A general study with high effectiveness and efficacy of the measurement method is a clinical trial. The clinical trial includes a reliable data collection procedure based on a clinical protocol in a controlled environment [88]. The clinical protocol is a translational guide for the data collection, and advanced digital technology has been developed as an

electronic data capture (EDC) platform to prepare, store, share, examine, and analyze the data of electronic case report forms (eCRF) with the protocol. In November 2017, Tuft center for the study of drug development showed that 100% of companies were using EDC as a data source [1]. For example, the EVO game app has evaluated the impact of the treatment of attention improvement, working memory, and inhibition in 80 children, ages 8 to 12, with diagnosed ADHD or not, under the platform circumstance and clinical trial protocol [26]. Professional researchers conducted the clinical trial in three site locations for 29 days. However, the methods of data collection in clinical trials are not appropriate for participant-oriented assessment of healthcare apps (which are exponentially growing) because of its complexity and high cost [28, 106].

A study on the real-world measure of the effectiveness of the intervention in broad groups is a pragmatic trial [33]. The 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, or mailed questionnaires, or specific healthcare apps developed from research kits [10, 15, 18]. Recently, David et al. developed a meditation-inspired, closed-loop software program (MediTrain) for an improvement in the continuous attention and working memory of healthy young adults [122]. To validate the efficacy, they have conducted a clinical trial that showed a definite difference in the key neural signatures of attentional control. They also conducted a pragmatic trial to compare the significant differences in the effectiveness of MediTrain from participants recruited through Amazon Mechanical Turk (MTurk). MTurk is a crowdsourcing marketplace for the recruitment of research participants and a platform for conducting surveys [82]. Although these kinds of data collection platforms are valuable when adequately utilized or developed by experts, they are not appropriate for the continuous collection of effective data from end-users of millions of apps with generic views.

In addition, Google Play and Apple App Store are potential data collection platforms because they keep user ratings and reviews in the public domain [74]. An expert-led assessment platform such as NHS Apps Library publishes safe and secure apps through questions designed by experts from technical and policy backgrounds [3]. Both platforms have challenges of the possibility of bias, subjectivity, and opaque results on the effectiveness of an app [12, 67].

Finally, all the platforms we mentioned have limitations that do not reflect the prevention of drop-out and software characteristics of digital healthcare apps. A study with a daily collection of pain data using digital measures has shown that the self-reporting completion rate is at an average of 71.5% (261/365) days. The latest researches are attempting to develop an incentivized program or in-game rewards to increase the self-reporting rates [48, 54, 104]. Furthermore, unlike drugs, functional foods, and mechanical devices that are difficult to modify after the market launch, healthcare apps can be potentially updated as software [30]. The app features may require iterative evaluation upon the upgrade. In summary, a new platform for digital healthcare apps should have a sustainable data collection function.

Here, we propose a participant-oriented reliable data collection method for participatory trial platforms as a first stage to understand the effectiveness of healthcare apps. The method consists of three steps as understandable data preparation, systematic data storage and sustainable data sharing. We utilize a clinical trial protocol database and semantic relatedness method for the participatory trial protocol to prepare data collection. We develop a data weight reliability formula that collects data systematically. We propose a reward distribution function connected to data reliability weight to sustainable collect data. In the result section, we describe the following experiments to validate the reliable data collection method; comparison to the simplicity of data preparation, data reliability weight validation and reward distribution effect observation.

We test a total of 718 human participants for the experiments. The Institutional Review Board (IRB) of KAIST approved an IRB exemption for the experiments.

Moreover, we have developed a web-based pilot platform accessible to the public with real-world data as a crowdsourcing tool based on citizen science concept. The pilot platform organically integrates all proposed methods. We conduct case studies on the pilot platform to validate efficient recruitment. To demonstrate the advantages of the proposed platform, we compare the platform functionality to existing platforms.

2 Definition of Participatory Trial

A participatory trial is an expanded form of a human-involved trial in which the public is used to test the effectiveness of products within their daily life. The concept follows crowdsourcing and citizen science in the aspect of a data-driven science [45, 58]. The participatory trial includes voluntary participation, unlike selective inclusion of the clinical trial and broad inclusion of the pragmatic trial. The participants from the public operate the trials. The objective of the participatory trial is to inform consumers about the effectiveness of a product, and the method consists of a protocol reflecting daily life care to maximize the reliability of the clinically relevant results. We have compared the continuum between clinical trials, pragmatic trials, and participatory trials in Table 1 [75]. We have defined a participatory trial to ameliorate current trial definitions in modern society.

Limited to this manuscript, we also defined the meaning of a word or phrase as following;

- Platform: Digital platform that is an environmental software system associated with each functional part of application programs.
- Data preparation: Determining the data collection structure and having the same meaning as designing the protocol.
- Data storage: Evaluating collected data to store reliably and to determine the value of the data.
- Data sharing: Integrating collected data to share with others and to provide benefit with the data provider.

3 Related Work

3.1 Clinical Trials and Platforms

Clinical trials have been the gold standard for evaluation or development of medical interventions for over 70 years [11]. Focused on evaluating the efficacy of specific interventions, clinical trials often use controlled settings and strict criteria for participants and practitioners. The EDC system is increasingly recognized as a suitable method that has advantages in real-time data analysis, management, and privacy protection [112]. The system is most useful in trials with complicated contexts, such as international, multi-centered, and cluster-randomized settings [109]. However, because of their nature of strictness and complexity, clinical trials still have certain limitations in the thorough evaluation of effectiveness or external validity [90].

Table 1. Continuum between current trials and participatory trial

Continuum	Clinical trial	Pragmatic trial	Participatory trial
WHAT	Efficacy	Effectiveness	Effectiveness
WHY	Assessment	Decision making	Information delivery
HOW	Rigid protocol	Quasi-explanatory protocol	Interpretable protocol
WHO	Selective inclusion	Broad inclusion	Voluntary inclusion
WHERE	Controlled environment	Daily environment	Daily environment
OPERATION	Expert		Public

3.2 Pragmatic Trials and Platforms

Clinical trials are often 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 the intervention in actual clinical practice settings [33]. As introduced in the "PRECIS" assessment by Thorpe et al., these trials try to maximize the external validity of the research by including the heterogeneous population and setting patient-centered endpoints [105]. Various platforms, including mailed questionnaires, web-based forms, and research kits, are used for these kinds of trials [10, 15, 18]. Consequently, pragmatism is an emerging source of clinical evidence, from pediatric asthma and cardiovascular diseases to monetary incentives for smoking cessation [6, 42, 95]. However, it is subject to several challenges such as patient recruitment, insufficient data collection, and treatment variability [33, 101]. In addition, inconsistency in data platforms is another major limitation of pragmatic trials [107].

4 method

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

4.1 Interpretable Data Preparing

At the data preparation stage, we determined what kinds of data should be collected, how to collect the data, and from where we would get the data. We had to consider the data as an interpretable resource to determine the effectiveness of a healthcare application. In the fields of clinical and pragmatic trials, researchers usually develop a protocol to facilitate the assessment of a changing symptom or disease [19, 105].

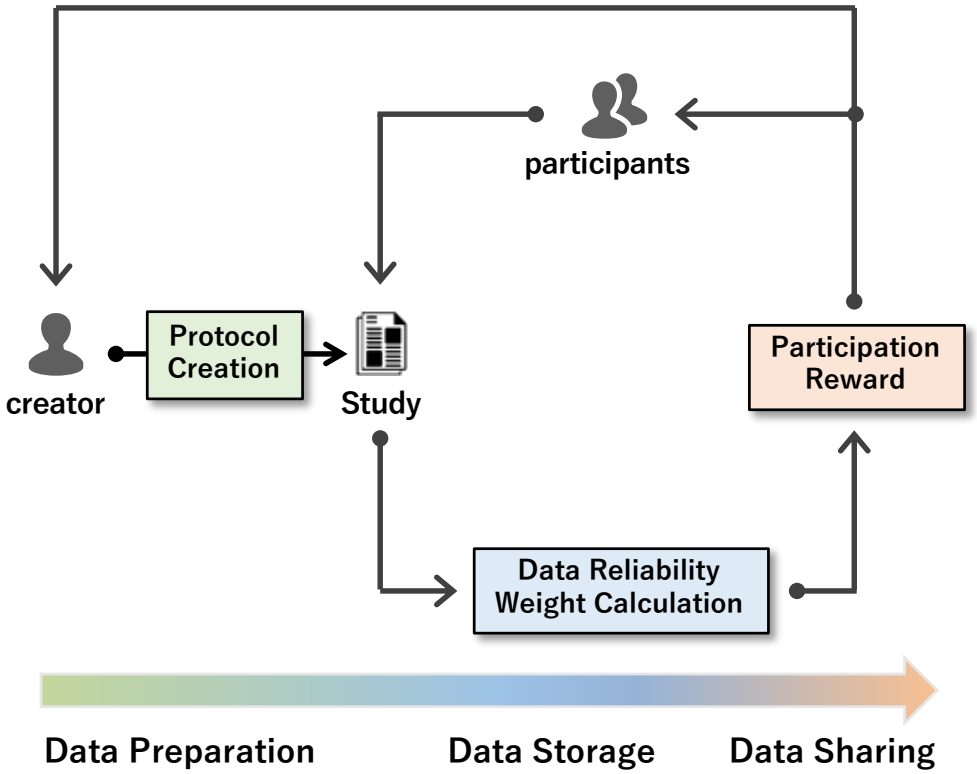


Figure 1. Method overview The method consists of three parts: data preparation, data collection, and data distribution. In data preparation, a creator can make a new study for the evaluation of an app by searching from the protocol creation method. Next, in data storage, the participants conduct a study, and the platform stores all the replies of the participants. At the same time, it collects statistics of the replies and calculates the data reliability weight of each reply. After finishing the study, in data sharing, the platform calculates a reward depending on each person's reliability of both the participants and the creator and distributes the reward to each of them.

Healthcare applications also focus on improving a symptom or treating a disease. A phenotype is an observable physical characteristic of an organism, including a symptom and disease. Therefore, we can effectively measure the change of a phenotype from an existing protocol that is semantically related to the phenotype.

We developed a protocol database system and a concept embedding model to calculate semantic relatedness for a protocol retrieval system for interpretable data preparation. The database contains 5,765,054 phenotypes for context-dependent protocol-element-selection [83]. We created the model to find a vector on a latent space based on distributed representations of a documented method. Furthermore, we calculated the semantic relatedness between input phenotype and protocols with cosine-similarity [60, 84]. Finally, we integrated the methods into the proposed platform for the participatory trial, so that anyone can create a study to easily determine the effectiveness of a healthcare application. Figure 2 shows the overall procedure of the method.

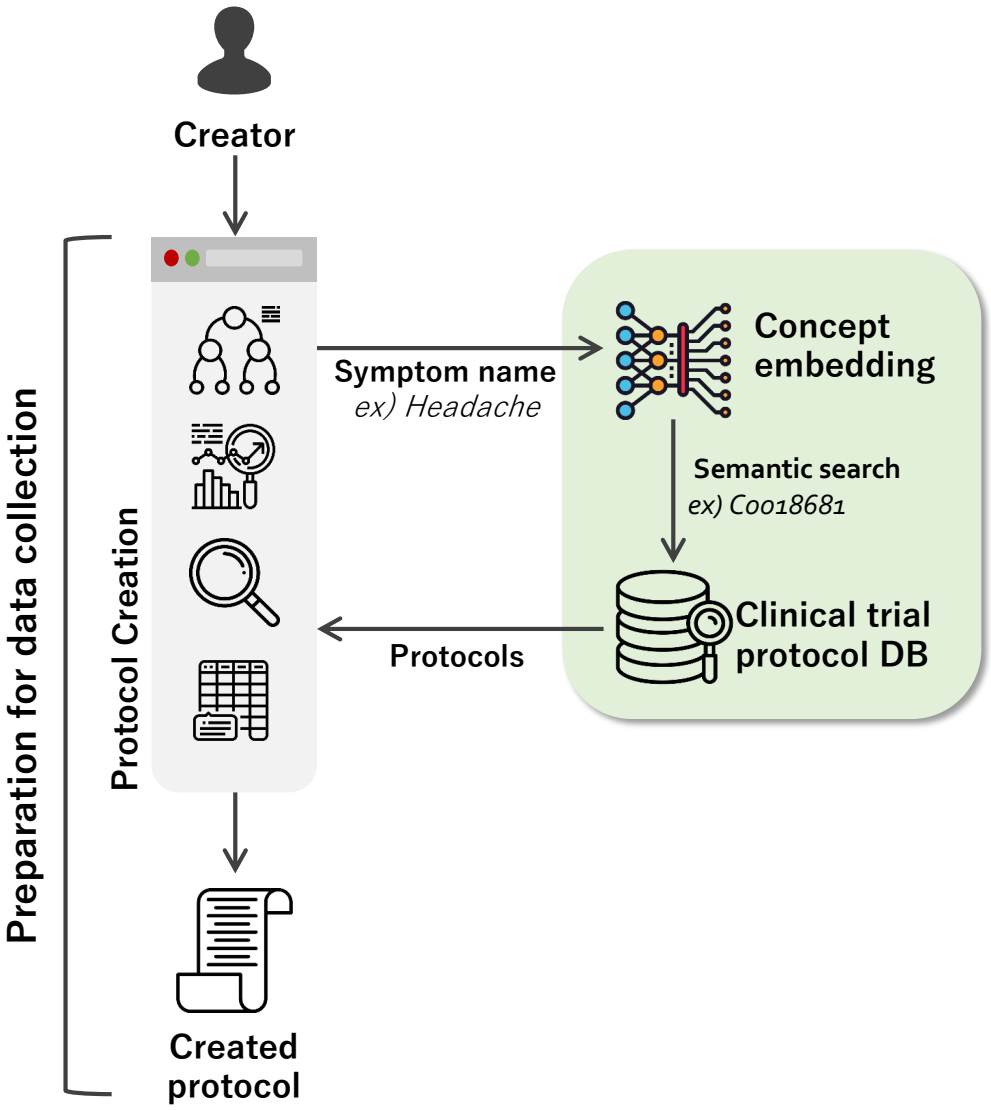


Figure 2. Data preparation to collect The creator retrieves the previous protocol based on the symptom name of the app to verify effectiveness. Then the creator prepares to collect data on the retrieved protocol and create the participatory trial protocol for data storing.

4.2 Systematic Data Storage

The participatory trial platform uses crowdsourcing methods to collect data. Crowdsourcing is obtaining data and using it for scientific research through voluntary participation from 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 [53, 56]. However, data collected from crowdsourcing is often pointed out to be less reliable than results obtained by systematic approaches [29, 64]. Therefore, calculating a measure to acquire credible data is the main challenge of the data reliability.

The primary purpose of this method is reliable data collection and delivery for effectiveness evaluation, which is the next level in the path to validation. We needed a method to guarantee the reliability of input data to achieve this purpose. Therefore, we proposed a formula to calculate data reliability weight (DRW) through combined effort measures of input data.

Careful, relevant, and purposive data input by the participant can improve the quality of data in systematic data storage [71]. Huang et al. developed an insufficient effort responding (IER) detection method to determine the input data for providing accurate responses and correctly interpreting the data [47]. The recommended composition of IER is response time, long-string analysis, and individual reliability. Response time (RT) is a method to determine the effort based on execution time [117]. Long-string analysis (LSA) uses intentional long-string patterns to assess the effort of a participant [24]. Individual reliability (IR) divides a pair of data items into two halves and uses the correlation between two halves to determine the normal response [93]. In addition, Huang et al. also presented a single self-report (SR) IER item, which has empirical usefulness [46]. The SR is a single item for detecting IER, and SR consists of the same 7-point Likert scale [73]. We developed the DRW calculation method based on the IER methods.

We calculated $DRW_{idx}(X)$ according to the type of electronic case report form (eCRF) for obtaining DRW (1). f_{RT}, f_{LSA}, f_{SR} and f_{IR} are the indexes of DRW. The created protocol contains eCRFs, and an eCRF is a tool used to collect data from each participant. We provided four types of eCRF described in Table 2. We calculated DRW from the symptom and data type eCRF. The symptom type measures the presence and the degree of targeted symptoms only once. The data type measures the effectiveness of the targeted apps repeatedly. $X = \{x_1, x_2, ..., x_n\}$ is the set of items in eCRF. The platform receives 11 types of item data in eCRF, as shown in Table 3. Type codes ED, TEL, SET, DATE, PDATE, and FILE were not used to calculate the DRW.

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

We calculated f_{RT} and f_{LSA} for symptom and data type of X. f_{RT} calculates the difference between the start and end time when the participant input the data to X. f_{LSA} uses the length input strings. On the other hand, 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 response by the participant on the iterative procedures. To measure whether a participant has attention to enter data at the end, we placed an item of SR 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 obtained a Pearson correlation coefficient [35]. 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 generate Z and $Z^{'}$ so that the correlation is close to one. z_{m_k} and $z^{'}_{m_k}$ the each item value of X_k.

K is the number of participants, and k is kth participant of K.

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

Table 2. eCRF types and their description

CRF Type ID	Name	Description
1	SYMPTOM	Symptom investigation of a participant
2	ELIGIBILITY	Eligibility checks to participate in a study
3	CONSENT	Voluntary consent to participate in a study
4	DATA	Data collected to find effectiveness of healthcare apps

Table 3. Item data types for eCRF

Item_data_type_id	Code	Name
1	BL	Boolean
2	BN	BooleanNonNull
3	ED	Encapsulated data
4	TEL	Telecommunication address
5	ST	Character string
6	INT	Integer
7	REAL	Floating
8	SET	Group values
9	DATE	Date
10	PDATE	Partial date
11	FILE	File

Each index has its cutoff value and calculates the cutoff values for each prepared CRF in a study. In detail, we calculated the mean (μ) value and the standard deviation (σ) to remove outliers. We removed an outlier if it was 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 DRW indexes (3). The one reason for CDF is to find a cutoff value that has the lowest area under the probability density function among the input values. The other reason is to represent random variables of real values whose distribution is unknown using the collected values [16]. Accordingly, we used the CDF values to find the μ'' and σ'' again to get a normal distribution. On the distribution, we gained cutoff-value using z-score for the p-value

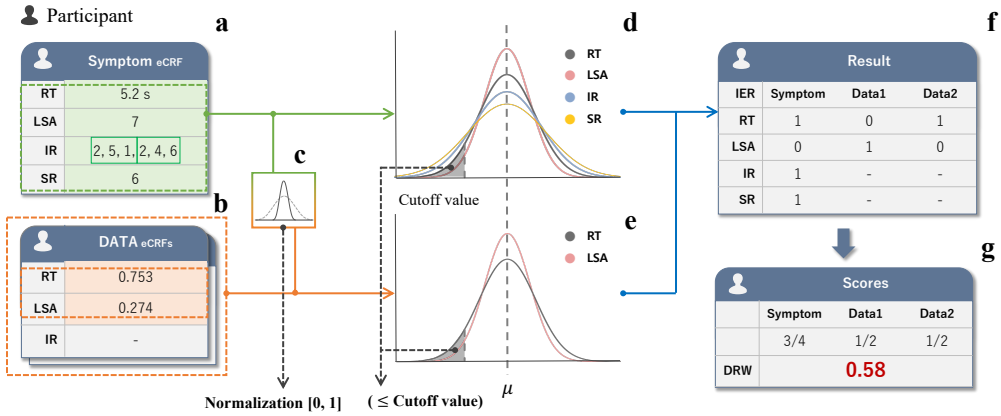


Figure 3. Example of calculating the DRW of one participant in the study (a) The platform calculates RT, LSA, IR, and SR based on the data entered to calculate the DRW of the symptom of eCRF. Every participant has one symptom eCRF. (b) The platform stores the collected data in data eCRF. Data eCRF, unlike symptom eCRF, does not display a specific pattern, so it calculates RT and LSA excluding IR and SR. (c) The platform calculates the cumulative distribution function for the normalization of input values. (d) The platform calculates 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 the same items of 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 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.

(default is 0.05). $h(X')$ return 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 with the binary decision function.

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

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

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

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

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

4.3 Sustainable Data Sharing

A reward can drive active participation, which is an intuitive fact. Recent scientific findings support the fact that a reward has an impact on data acquisition [34, 54]. We

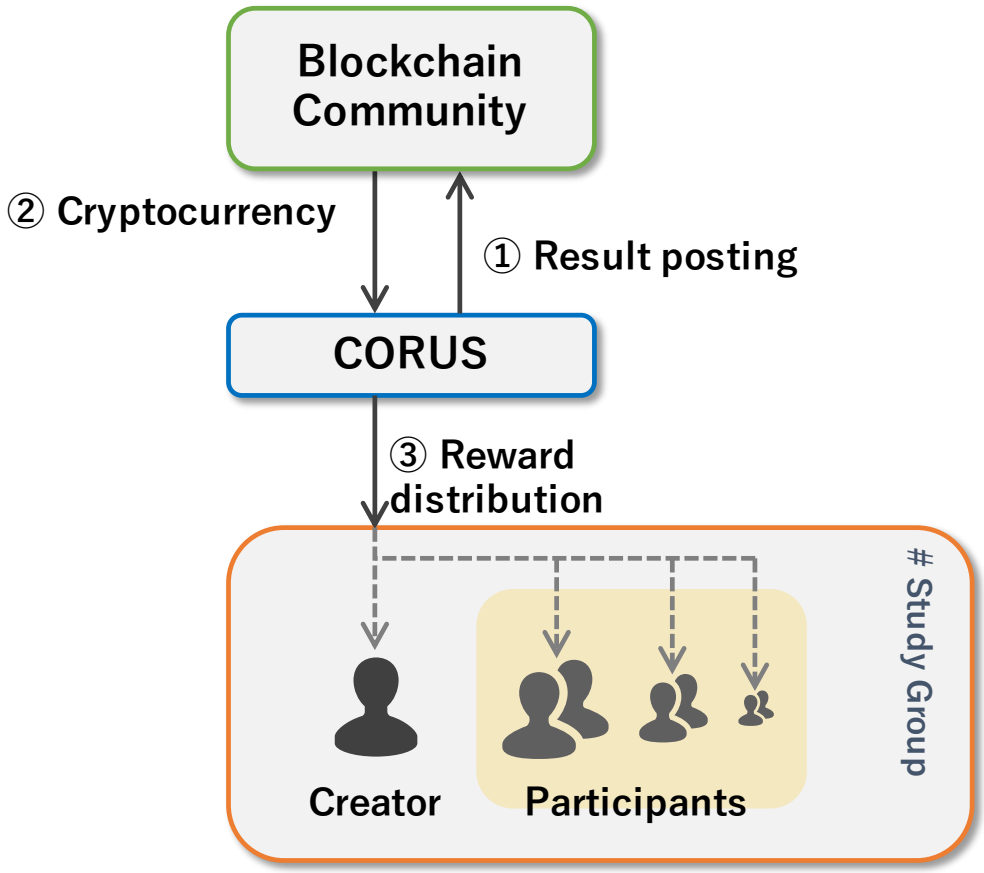


Figure 4. Participation and rewards of the trial On participating in the trial, the results of each participant are posted on the blockchain community (①), and the community provides cryptocurrency to the CORUS (②). The platform distributes rewards differently depending on the participant's level of involvement after the end of the trial (③).

provided financial rewards for sustainable data collection as exchangeable cryptocurrency [55, 85]. To provide the reward, we developed a study result transfer function that delivers statistical information of the completed study to an external cryptocurrency system (Figure 4). The study participants periodically receive cryptocurrency as a reward for their data. However, the reward is an expected compensation after the end of the study, not an immediate profit. Therefore, we tried to induce sustainable participation with an expectation that active participation will earn high rewards [110]. The expectation established a reward distribution method to draw continuous management motivation from the creator and the active involvement of participants.

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

We collect rewards per each completed study. R_{total} is $R_{creator} + R_{participants} + R_{system}$. R are rewards. $R_{creator}$ calculates the reward amount based on the compensation proportion of participants and μ_{DRW} value of the study (7). If the μ_{DRW} is low, this acts as a kind of penalty. We introduced this to encourage the

study creators to create a better plan and stimulate a study

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

Here, $R_{participant}$ is that R_{total} substrates $R_{creator}$ (8). K is the total number of participants. DRW_k is the calculated DRW for kth participant. The platform rewards participants for their efforts to input data. A participant receives more rewards if he or she made a more than average effort. Otherwise, the participant gets a lower reward. R_{system} acquire high compensation when the μ_{DRW} is low. We added R_{system} to recover the cost of wasting platform resources when careless participants generated unreliable data. We will use R_{system} as a platform maintenance cost.

5 Pilot Platform Implementation

The proposed platform is designed for a participatory trial. The characteristic of participatory trial in terms of data collection is that people can become researchers and conduct experiments, systematically collecting data based on crowdsourcing and sharing the results with the public [52, 58].

To achieve this goal, we have developed a platform that integrates all the methods presented above, so that each feature is organically connected to another. The business 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 apps. Creators build studies in the data preparation phase by using the protocol database and protocol discovery methods. Once the created study is activated, users can join the study. Participants of the study will be curious about the effectiveness of healthcare apps, so that they are more serious about entering data. Research data and basic information about the study are stored in the blockchain to prevent data falsification in vulnerable system [85]. The study ends after the predefined duration of the research. Even if the number of participants is not enough to develop a sufficient amount of the data, the study would be closed at the pre-set endpoint. The system then calculates the effectiveness score from the study data. At the same time, the research results are open to both the public and experts in the data sharing phase, and participants receive rewards based on the data entered.

5.1 Blockchain to prevent data falsification

The clinical research field often reports data falsifications. According to a National Institute of Health (NIH) study, 33% of researchers have experienced data falsification [68]. Blockchain is the most promising technology to prevent data distortion for any reason during the data storage. Characteristics of the blockchain in a clinical trial are immutable, traceable, and more trustworthy [118]. Recently, we demonstrated an integrated blockchain system to prevent data falsification for the crowdsourced data collection platform in participatory trials [85]. The system consists of a two-stage blockchain separated into a private chain for short-term data collection in progress and a public chain for long-term data storage from the completed study, respectively. The blockchain does not expose progressing data to participants to prevent the bandwagon effect and provide complete data to everyone to maintain transparency at the end of the study [80, 102]. We used the blockchain system as the underlying data storage system of the participatory trial platform.

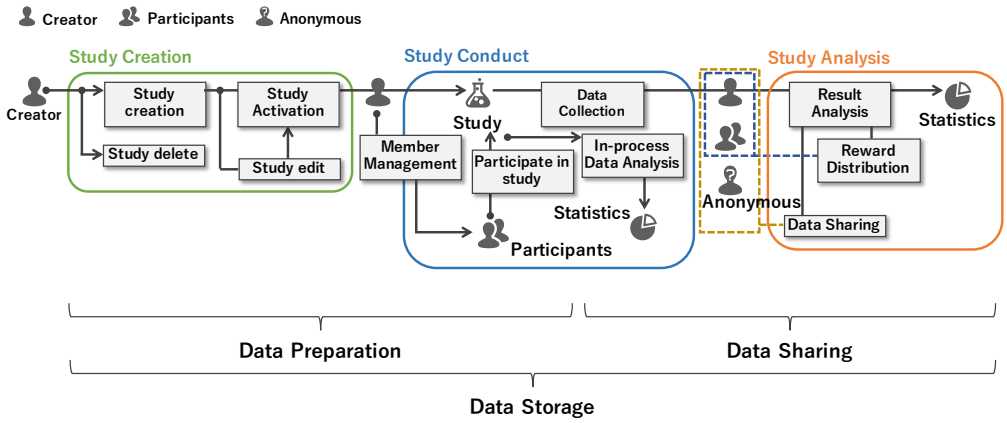


Figure 5. Business logic overview of the platform The platform integrates and connects methods for reliable data collection at all stages of study creation, study conduct, and study analysis. In the study creation stage, the creator can create, modify, and delete the study. When the creator activates the study, it is open to other participants, and they can participate. In the study conduct stage, the creator can collect the data from the participants and can monitor them. In addition, the creator and participants can monitor statistical information related to the study. In the study analysis phase, all data generated after the study is completed will be disclosed to the study creator, participants, and anonymous users. This makes data available to the experts and the public. The platform services allow users to prepare, collect, and distribute reliable data at each stage in the integrated environment.

5.2 Data distribution

Data sharing can assist with ongoing research and positively impact derivative research [2, 32]. In this regard, we expect that data sharing of the participants arouses a sustainable interest for ongoing or future studies. On the proposed platform, data sharing consists of two directions: transparently sharing research findings with the public and delivering them to a group of experts. Their detailed description is as follows. Sharing data with the public is aimed at transparently disseminating research findings and encouraging more participants to continue to access the platform based on trust. After the study is completed, anyone can view the study results through the platform interface. Users can easily share the results with other communities by using the SNS delivery function on the study results page. Furthermore, raw data can also be downloaded from the study. However, there is a problem with anonymity during this process [108] To solve this, it is now possible to use IDs that are randomly given when users participate in research. [91] Each table is stored separately in the database, and only relevant research information is provided to the researcher. For example, if a user with an ID of example@corus.com participates in a study number 2849, they will be assigned a random ID of 2849-1. Randomly assigned numbers are determined within the range of numbers of the target participants. Study creators can only view anonymous participant information. Participants can provide information on gender and age in participatory studies, but this is optional, and even if such information is provided, identification of participants is very difficult in the absence of specific regional information [103]. Consequently, study results are transparent to the public, while personal information is opaque. This structure was developed based on Opensource EDC's Database Schema [41]. The proposed platform aims to collect reliable information. However, the data collected for detailed verification must be readily available to a group of experts. Convenient data utilization can only be achieved in a

standardized format. Similarly, platforms such as PatientsLikeMe allow patients with rare diseases to share their treatment information, and clinical professionals can test the exact effect of shared information when interest in efficacy increases [70]. As such, in some cases, it is necessary to verify large-scale partial trial results using small-scale clinical trials precisely. This is a procedure that can express the effectiveness as efficacy [62]. This required a service that would independently exchange results or store and deliver clinical data. To this end, we have inherited the EDC function to receive data in the Operational Data Model (ODM) XML format that is compatible with the Clinical Data Interchange Standard Consortium [13, 38].

5.3 Technical Implementation

Technically, we designed the platform according to the interface segregation principle [116], which provides an object with an object that uses only the services that the object expects. We categorized this pilot program as a platform because it could be the basis for external applications beyond its original purpose. We designed a business logic to integrate all methods mentioned above: data preparation, data storage, and data sharing. We chose Amazon Cloud as the environment for this platform. Moreover, we designed our server architecture using Amazon services to create a highly available, stable, and secure system environment. The software consists of three layers: web interface, API engine, and database. The functions of each layer are based on Node.js, Java programming language, and PostgreSQL database. In addition, essential functions and fundamental database schema related to the EDC system are designed using OpenClinica 3.12.2 [17]. See the Supplement for more information on platform development.

6 Results

The results of the proposed method are quantitative measurements (the test results of each method) and qualitative measurements (a comparison of the platform functions with other platforms). We conducted quantitative measurements with the participants of the platform in terms of simplicity of protocol creation function in the data preparation stage, DRW validation in the data storage stage, and the influence of expected rewards rather than immediate rewards in the data sharing stage. All participants filled and signed an electronic consent form. In the quantitative measurements, the participants' data input was 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 clinical and pragmatic trials.

6.1 Comparison to the simplicity of data preparation

We developed a protocol retrieval method based on the clinical trial protocol database and validated the method in our previous studies [83, 84]. The methods are the core function of data preparation. In details, we validated semantic filters of the database that can provide more appropriate protocols than a keyword search [84]. We used clinical trial protocols and corresponding disease conditions as extracted golden standard set from clinicaltrials.gov [120]. Our F-1 score was 0.515, and the score was higher than the keyword search score 0.38. Concept embedding model to find semantic relatedness of clinical trial protocols showed 0.795 Spearman's rank correlation coefficient with the benchmark set from Mayo Clinic [44, 83]. The result was higher than the previous method Lesk of 0.46 and Vector of 0.51. Finally, we conducted a user evaluation test of protocol retrieval system from ten clinical trial experts point of

view [84]. Our system presented 1.6 difficulties and 6.5 satisfaction using a 7-point Likert scale. Clinicaltrials.gov was 6.2 and 2.3, respectively.

However, previous results have a limitation that tested solely for expert views. And we conducted the results for a clinical trial, not a participatory trial. In contrast, the interpretable data preparation method we proposed is a publicly available system to apply the above technique from the perspective of a participatory trial. Therefore, we need further validation of whether the proposed method has a simplicity that anyone can create an interpretable protocol to prepare data collection.

Table 4. Descriptive Statistics of USE in the proposed data preparation method and existing expert method.Statistical description of proposed data preparation method (P) and existing expert method (O)

O \ **P**	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

For the validation, we designed an experiment that collects Usefulness, Satisfaction, and Ease of Use (USE) score from human participants to reflect their simplicity experience of the different methods [65]. We recruited human participants represented general users using MTurk [82]. We used 1.1 expected mean and 1.14 expected standard deviation (SD) of previous USE study [36]. 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 was 20 by adjusting the sample size for t-distribution [27]. The participants of the experiment compare the data preparation function, which is related to protocol creation between the proposed method and the existing expert method. We prepared the data preparation method as a simulation program (PM), and the complicated method used Openclinica 3.14, an open-source EDC program (OC) [17]. Two programs recorded randomly generated ID of participants to identify their test completion. The participants respond to the USE questionnaire, which consists of 30 items of 7-point of Likert scale for the comparison score and 6 short items for their opinion after each use [65]. We reversed the 7-point of Likert scale to make participants more concentrate on the questionnaire [14, 46]. We restored the modified scores in the result calculation. In the experiment description, we only provided an elementary data preparation guide and no detailed manual. And we divided participants into two groups to prevent recall bias occurring in the order of the program usage [81]. We reversed the order of use for each group. Finally, We configured the hardware environment identically on the cloud computing resources of Amazon web services to only identify software differences [20].

After completing the experiment, we excluded participants who did not complete all USE items, has all the same item values or skipped test of method systems. Consequently, we could obtain forty-two participants from the combined data set of two groups Supplementary data 1). We conducted the descriptive statistics of both methods

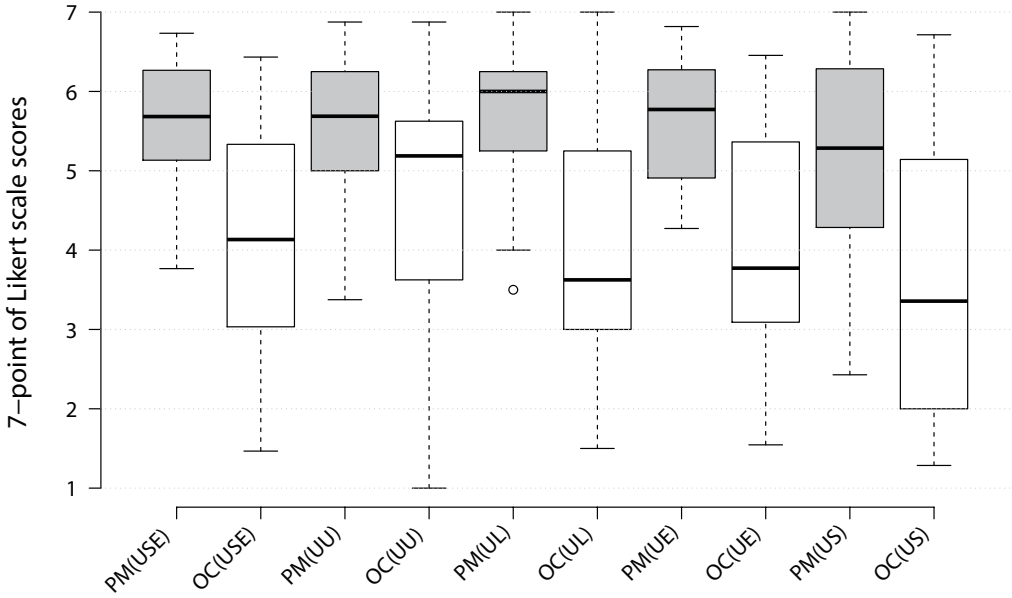


Figure 6. Whisker box plots of the dimensions in USE between two groups The black line demonstrates the median within the boxplot, the length of the box represents the interquartile range, and the whiskers 1.5 times the interquartile range of the data. All data points outside the whiskers are considered outliers. The proposed data preparation method (PM) showed higher records than the existing expert method (OC) in all domains of tests.

as four divided dimensions [36]. Accordingly, the 30 USE items became to belong to 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 proposed method scores were higher than 5 for all USE dimensions (Figure 6 and Table 4). We also found that scores showed negative skewness for all aspects.

As described in Table 4, Average USE score of PM was 5.548 (SD=0.897), and OC was 4.3 (SD=1.5). We conducted a statistical test to confirm a significant difference between the average scores. To conduct a test of normality the USE score, we did the Shapiro-Wilk test [97]. The test result p-value of our score was 0.082, and the compared score was 0.238. They satisfied normality ($p > 0.05$). Therefore, we did the paired sample t-tests and, the USE score presented significantly higher value in PM than OC, $t = 7.243$, $p < .001$. Since the four dimensions in USE did not satisfy the normality, we conducted the Wilcoxon signed-rank test [115]. The outcome of the test showed noticeable differences between the two methods in all dimensions (Table 5).

Besides the results, we also found that the participants completed 100% tasks on PM, while they only generated basic metadata of given task on OC. Thus, we concluded that PM is more suitable for understandable data preparation in participatory trials.

6.2 Data Reliability Weight Validation

DRW is a weight that is calculated by incorporating a score that evaluates the effort spent when a user enters data. We examined the correlation between the Human Intelligent Task Approve Rate (HITAR) of MTurk and DRW to assess whether this score is measured well. The HITAR is a score that evaluates how well a worker is performing tasks assigned by the requestor on MTurk. Consequently, we assumed that a higher

Table 5. Related-samples Wilcoxon signed rank test on the dimensions in USE between proposed data prepration and existing expert method

Dimension	s.e.	s.t.s.	p
Usefulness (UU)	71.531	-4.327	<.001
Ease of Use (UE)	74.387	-5.196	<.001
Ease of Learning (UL)	74.341	-5.125	<.001
Satisfaction (US)	77.144	-4.848	<.001

Note. s.e. = standard error, s.t.s = standardized test statistic

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 to HITAR. DRW indexes contain response time (RT), long string analysis (LSA), individual reliability (IR), and a self-report (SR) [46, 47]. We intended to organize DRW with empirical DRW indexes for simple calculation in eCRF. Because DRW is a subset of IER, we made a questionnaire to calculate DRW based on detecting IER research [46]. We prepared a questionnaire to contain 60 items of the international personality item pool—neuroticism, extraversion, and openness (IPIP-NEO)—to calculate IR [50].We also added 8 items of the infrequency IER scale to increase the concentration of participants at the beginning of the test and SR item for DRW score [73, 86]. We placed a self-input HITAR item box at the start of the questionnaire as LSA item. Thus, the questionnaire included 70 items. To calculate the total sample size, we used 0.13 expected mean of the paired differences based on the result of our pilot DRW test with a total of 100 participants in MTurk. We set a power of 95%, a one-sided level of significance of 5%, and equal group size for sample size calculation [27]. The calculated sample size was 153 for each group, and the total size was 306. We recruited the participants using MTurk. Workers of MTurk participated in the study as participants. We divided all participants into two groups, with HITAR levels above 95 and below 90. Participants over 95 HITAR had to have a record of more than 500 tasks before participating. This was to select a participant with an average of more than one year from the average number of 1,302 tasks of approximately 2.5 years analyzed in the previous study [43]. Participants below 90 HITAR had no other restrictions. This is because the first granted HITAR is 100, and it is a deduction method when the task is not performed well. Participants read the task description posted on MTurk. All settings of the test are identical except for the HITAR levels. The endpoint of the test was set to the recruitment completion time of one of the group.

Consequently, a total of 340 participants participated in the test. Each group has 170 participants, and we considered that the number of participants reflected the calculated sample size. Based on the collected data from the participants, we compared the average value of DRW and DRW indexes per each group. In all DRW indexes and DRW, groups with more than 95 HITAR scored higher than the opposite group (Figure 7 (a)). In more detail, we did a statistical analysis to understand the difference between the two groups. Table 6 displays the statistical description. The DRW index values in the table 6 are converted to CDF value to calculate the DRW. SD in all areas except RT and LSA showed similar values. The reason was that 95 HITAR group had very high values in RT and LSA. We calculated cutoff values using CDF values through DRW calculation method. We checked that only passes a range of values with approximately 0.5 or more. We verified statistical significance between the mean values of the two

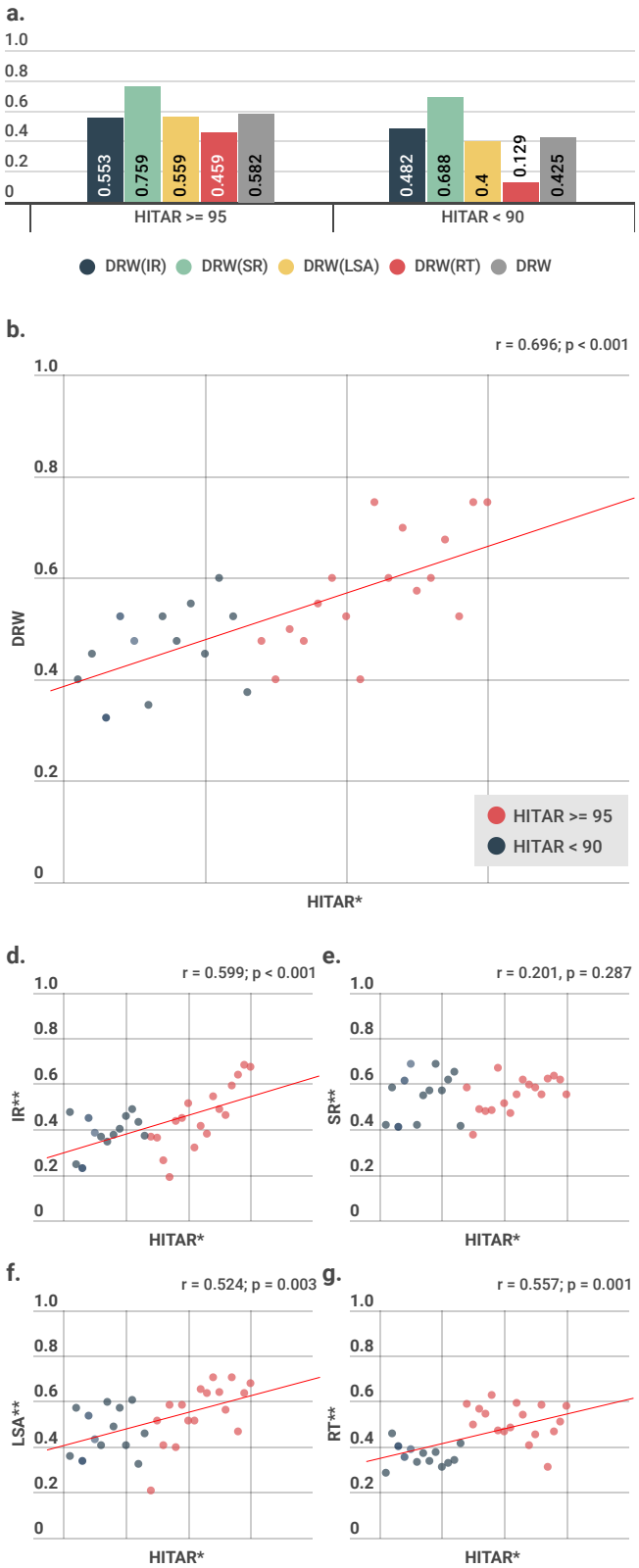


Figure 7. correlation between DRW and HITAR (a) Average DRW indexes of worker groups classified by HITAR; (b) correlation between HITAR and DRW; (d-g) correlation between HITAR and DRW indexes; * is scaled values. ** is normalized values.

Table 6. Statistical description and independent samples t-test of IR, LSA, RT, SR, DRW between groups of HITAR values greater than 95 and less than 90.

Index	HITAR condition	Number of participant	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]Not applicable [b]The t-test result was reported on the unequal variances based on the Levene's test. [c]The value is normalized by CDF.

groups using the independent samples t-test. We performed Levene's test to consider unequal variance in the t-test [63]. The p-values indicate that the results are statistically meaningful. Only SR had similar values in the two groups. We presumed that the mean value difference of SR was small, but SR would have a small effect on the calculation of DRW. Furthermore, we calculated effect size, an indicator of how much difference between the two groups [72]. Since we are comparing the same sample size between the two groups, we calculated the effect size using Cohen's d, and the result closed to a large effect (.8) according to the interpretations of Cohen's d [21].

Next, we compared the correlation between DRW values and HITAR. The obtained HITAR values presented high skewness (-2.138) because the HITAR average we obtained had an 83.91 relatively high. To improve the performance of the data analysis and interpretation, we have established a binning strategy. As the binning strategy, we divided the data group by the same frequency, and we smoothed the grouped data values to the average value. Specifically, the actual data processing is as follows. We sorted the number of 340 data of participants based on the descending order of HITAR values. We grouped 10 as frequency and averaged the HITAR, DRW, IR, SR, LSA and RT values. Based on the groups, we ranked HITAR and took the reverse order. 34 Participants did not enter HITAR values correctly, and 6 data of participants remained as ungrouped. Accordingly, we removed the 40 data, and we placed 300 data into 30 groups. To check the effectiveness of the binning, we used information value (IV) that express the overall predictive power [96]. The measured IV value (0.465) was greater than 0.3, and we determined that the binning was good. Based on the binning, collected data showed significant results that indicated a correlation between DRW and HITAR (Supplementary data 2). DRW and HITAR showed a noticeable correlation (Figure 7 (b)-(g)). DRW and HITAR also presented an interesting correlation ($r=0.696$, $p < 0.001$) In sum, we validated that HITAR correlates with DRW and DRW can be used as reliability measures.

6.3 Reward Distribution Effect Observation

Recent studies have confirmed that systems with immediate virtual rewards have a positive effect on continuous data collection [76, 87]. However, we evaluated the validity

Table 7. Statistical description and independent samples t-test of IR, LSA, RT, SR, DRW between groups of rewarded and not rewarded.

Index	Reward condition	Number of participant	Mean	Standard deviation	Cutoff value	t	d.f. (degree of freedom)	p	Cohen's d
DRW	True	168	0.630	0.279	N/A[a]	2.982	334	0.003	0.325
	False		0.539	0.279					
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 the Levene's test. [c]The value is normalized by CDF.

of a reward distribution that gives rewards to participants in the future, rather than immediately. We showed the validity as the correlation between rewards and DRW. Accordingly, we evaluated the following scenario, in which the reward system of our platform increases the participation rate.

We created a simulation environment in MTurk to reduce the observation time in the real world. We agreed that a real-world study of the reward distribution effects requires significant time to collect the data and cases. For example, PatientsLikeMe required approximately 5 years to collect enough cases to show the benefits of communities (which were related to the extent of site use) [114]. Thus, we conducted two tests in the simulation environment. We designed one test that informed reward distributions in the future (the same as our platform). The other test did not contain this information. The remaining settings were unchanged. For the sample size calculation, we conducted a pilot test with 100 participants in MTurk. We obtained 0.11 expected mean of the paired differences and 0.25 SD on the result of the pilot test. Based on the result, we calculated the sample size as 168 per each group to have a power of 95% and a one-sided level of significance of 5% [27]. The total sample size was 336. Then, we used a questionnaire as data reliability validation and weights for consistency comparison, except for the HITAR. However, we needed to consider the casual observation characteristics of MTurk workers since this experiment has no constraint on HITAR [7]. Therefore, we added five questions that induce to enter as many words as possible in the LSA index of DRW [47].

We recruited 336 workers from MTurk to participate in the test. We allocated 168 workers to the test that contained reward information (RI) and the other 168 to the test without reward information (NRI). We designed the test to be completed within 2 hours, and we recorded the test start and end times for each worker as RT. Consequently, we successfully collected data from 336 workers (Supplementary data 3). The average DRW of RI was 0.63, and the DRW of NRI was 0.539 (Figure 8). Table 7 shows the significant difference between DRW of RI and NRI (p=0.03). Figure 8 presents that DRW(SR), DRW(LSA) and DRW(RT) of RI also have higher average values than DRW indexes of NRI, and Table 7 provides detailed statistical information. DRW Index values in Table 7 are CDF converted values to calculate the DRW. The DRW(IR) showed a low average result in RI, but we interpreted that the effect size (0.076) of the index is too low to effect to DRW result in Table 7. We presumed that the reason is caused by not

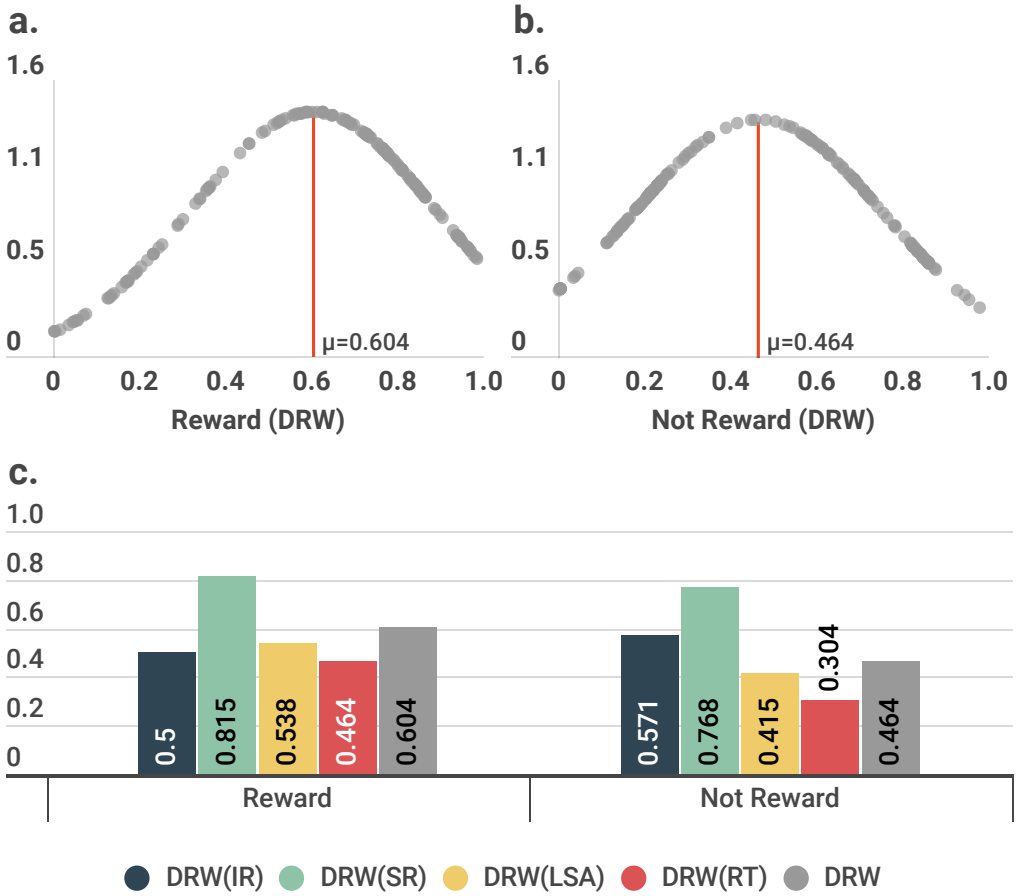


Figure 8. Reward distribution effects (a),(b) Graph of the probability density function of DRW raw scores. Left: reward information informed. Right: reward information not informed. (c) Each average score of DRW index for all workers per test.

controlling workers with HITAR. Interesting cases are RT and LSA. Both indexes showed high average DRW values and significance difference in Table 7 (0.583 (p ¡0.001) and 0.516 (p=0.001) respectively). In particular, we configured the DRW method to measure RT and LSA only in the data type CRF, 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 DRW but also helped to improve the sustainability of the data collection platform.

6.4 Overall Platform Function comparison

We developed "Collaborative research for us" (CORUS, `https://corus.kaist.edu`) as a pilot data collection platform for participatory trials. CORUS is designed based on the EDC of existing clinical trials but makes it easier to design a study and provides services that allow users to participate voluntarily and continuously collecting data in the study. In this section, we compare and analyze the features of CORUS in detail. For specific comparisons, we selected the programs mentioned or used in journals related to clinical research from 2012 to 2019. We summarize the comparison results in Table 8 and provide the details in the following subsections.

6.4.1 Comparison with the clinical trial systems

The main target of the clinical trial platform is clinical specialists. Therefore, the platform has complexity for the public to use. To select the representative clinical trial system, we searched under the "Best Clinical Trial Management Software" category in the G2 Crowd, a software comparison website . Rave, EDC platform made by Medidata, showed the highest level of popularity and satisfaction as a result [78]. Rave follows a typical EDC design approach. Beside, Rave has a more advanced feature to design basic document structure of a protocol and provides optimized protocols using a PICAS database with trial cost information [39]. Despite the functionality, Rave still appropriate for expert use. Transparency Life Sciences(TLS) is a leader in digital drug development services offering virtual and hybrid clinical trials [62]. TLS uses the crowdsourcing method for protocol design. In detail, experts create a survey project for disease and release the project to all the users of the TLS. After the release, the users interested in the project start to participate with their identities such as patient family members, researchers and patients. At the last step of the project, the TLS clinical trial team designs the protocol from the collected data of the participants. However, the expert team, not common users, leads to the creation of a protocol in the TLS. CORUS has a protocol creation feature as interpretable data preparation that makes it easy to use by the public without the help of experts. In the platform, the creator can utilize clinical trial protocol database feature according to the semantically related symptom of the effectiveness of healthcare apps. Moreover, the creator can get feedback from the study participants within the community feature, which allows the creator to enhance the protocol next time.

6.4.2 Comparison with the pragmatic trial systems

The pragmatic trial platform is to use for collecting data from massive and broad study subjects to measure effectiveness, and qualification requirements to the human subjects are not too rigorous like a clinical trial [33]. For instance, PatientsLikeMe is a platform that shares experience data for participants to understand possible treatments treatment of particular disease condition from the experiences of others [114]. We considered PatientsLikeMe as the pragmatic trial platform because the platform encourages the participation of many patients to build communities online for certain diseases. However, PatientsLikeMe does not provide an environment for the public to lead the study preparation, and the platform has no feature to evaluate collected data reliability. Another example is the MTurk. MTurk can belong to the pragmatic trial platform that requester (creator) can recruit a large number of workers (participants) for a study. However, the platforms did not provide standardized method at the data preparation stage. The platforms need clinical experts to prepare data collection procedures and system based on their knowledge, in other words. MTurk provides a feature to approve or reject of an individual worker, but the feature relies on a subjective judgement which has an objectivity limitation. Accordingly, we insist that the platforms also have no suitable feature to measure the reliability of the data collected from the participants. Protocol creation feature of CORUS provides understandable and straightforward data preparation for public users. The system features of CORUS helps the users lead a study without expert knowledge. Moreover, CORUS has the DRW feature that automatically calculates the reliability scores collected data for reliable data collection. The scores are an objective value that does not comprise the subjective intention of the study creator. We also developed the scores to evaluate the effectiveness score of apps in further study of data storage phase.

Table 8. Platform function and feature comparison in a manner of reliable data collection for the participatory trial

Platform Usable Category		**Participatory trial**							
		Pragmatic trial							
		Clinical Trial							
Program Name		Rave EDC	Transparency Life Sciences	Trial Chain	PatientsLikeMe	Amazon Mechanical Turk	Google play store / Apple app store	Guardians	CORUS
Data preparing	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	Expert	Expert	Expert, Public

6.4.3 Comparison with participatory trials

The participatory trial platform is to allow a public user to create and conduct the participant-oriented study for measuring the effectiveness of products within daily life. The core features of the platform should consist of interpretable preparation, systematic storage and sustainable sharing for reliable data collection. No previous platform has these features to the best of our knowledge. Accordingly, we identified comparable platforms that have an alternative, similar or essential features.

Google Play and Apple App Store is the alternative platform [5, 40]. Both platforms have free text app review and scaled rating function as data storage feature. However, the review feature has a natural language processing problem, including text classification [66]. In addition, the functions have a possible risk of transfer biases that could affect new reviews and ratings, because they expose prior collected data to new participants [81]. Other limitations of the platforms are that they do not offer data preparation features to collect systematic data collection protocol, data reliability evaluation features, and system features for participatory trials.

RankedHealth is a similar platform [42, 100]. RankedHealth has a data storage feature that can make a useful and reliable review of the quality of mobile healthcare apps. For collecting a reliable review, an editor and a reviewer role are essential in RankedHealth. The detailed procedure is as follows; (1) Multiple reviewers review an app. (2) An editor analysis of the reviews to identify the disagreement items. (3) The

reviewers discuss the items until they reach a consensus of the review upon the intervention of the editor. The objective of the procedure is to minimize the influence of individual reviewers for data reliability. However, the data storage feature has a limitation. The consensus of the experts is close to evaluate the technical performance of an app. That is not validating the experienced effectiveness of users of the app.

Sustainable data sharing for collecting large amounts of data is a crucial function of the participatory trial platform. The essential feature of the solution is continuous user engagement enticement. The Guardians is a platform which has a participation reward feature for the essential feature [104]. The platform collects data from participants as rewards through games. However, the platform does not have data reliability evaluation feature. In other words, unreliable data collection is possible. The participants can input unreliable data for the sole purposes of earning rewards.

We developed CORUS as a participatory trial platform. Data preparation features of CORUS prepare data collection that can perform immediate data analysis. Postprocessing for unformatted data is not necessary on the features. The features minimize the flawed study design bias [81]. Besides DRW, data falsification prevention feature prevents transfer bias by not exposing the collected data to the participants during a study. CORUS, a participant-oriented platform, also supports additional analysis for experts. Standardized data distribution is the feature. The feature is essential for validating the effectiveness of an app. Cryptocurrency is a distinctive participation reward feature of CORUS. We connected CORUS to an external blockchain to encourage participants sharing their data continuously. Participants can earn a reward based on their data. CORUS calculate the participant portion of total cryptocurrency of a study based on DRW of the participant. Thus, the participation reward can induce continuous engagement and reliability.

A common limitation of all mentioned platform, except CORUS, is the data falsification problem. The falsification possibility of the collected data, whether intentionally or unintentionally during a study, exists in the platforms [37]. Trialchain used blockchain to solve the problem [118]. We classified Trialchain as EDC based on the explanation of the platform. Thus, we considered that Trialchain has difficulty to use for a participatory trial. CORUS also uses Blockchain technology. We developed a data falsification prevention feature based on the data immutability characteristics of blockchain technology [85]. The feature solves the problem of data falsification.

7 Case Study: operation test of the pilot platfom

We conducted an operation test to observe the platform's data collecting capabilities in actual participatory trial conditions. The objective of the test was to confirm whether an app that enables blue light filters on smartphones would help to improve sleep disorders. Although there have been previous studies that showed the relation between blocking blue light and improvement of certain sleep disorders, to our best knowledge, it has not been clinically confirmed whether the blue light filters on smartphones or other mobile devices have the same effect [31]. The test has been designed to recruit 100 participants.

We designed the participatory trial project and posted it on the test platform. In the trial, participants were asked to apply the blue light filters (which blocked blue-wavelength lights on their smartphones) and report changes in the quality of their sleep. When users selected a specific project on the platform, introductions to the projects by the creators were shown. This made it convenient for users to participate in their preferred projects. Participants enrolled in the project could immediately participate in the trial and conveniently report data. The platform also encouraged participants to report their data daily by using features such as leaderboards and alarms.

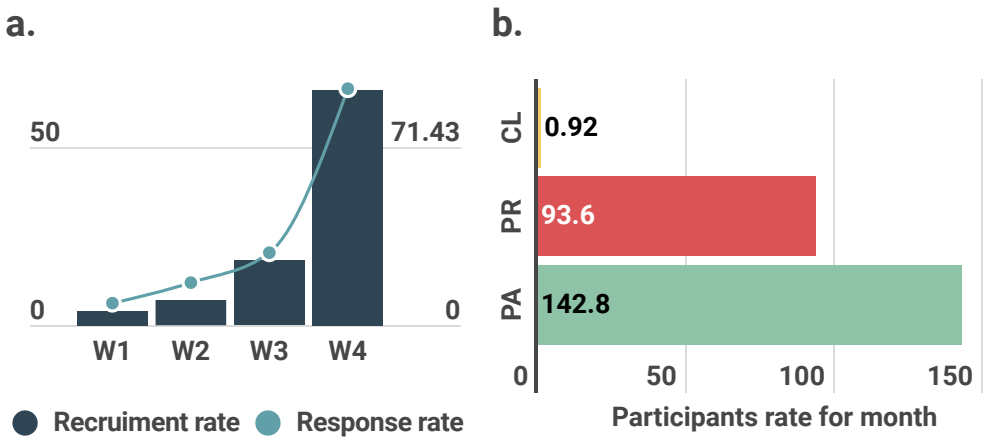


Figure 9. Recruitment rate and comparison analysis (a) recruitment and response rate for a month from a platform test study; (b) participant rates among clinical trial (CL), pragmatic trial (PR), and participant trial (PA) based on the literature evidence

In addition to technical validation, we also validated whether the system could effectively recruit participants of the trial. The duration of participant recruitment was originally scheduled for a month, but the actual rate of recruitment was faster than that. In total, 100 participants were recruited in 21 days. We compared the rate of recruitment in this experiment with those known from previous studies. To denote the rate of recruitment among the trials, the recruitment rate was 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 [111]. According to 8 web-based or mobile app-based studies collected from literature databases, an average of 468 participants was recruited for over 5 months (recruitment rate = 93.6) [59]. For the platform, the recruitment rate was 142.8. This shows that participant recruitment using the platform was significantly more effective than participant recruitment in traditional clinical trials; it was also competitive with other web-based or mobile app-based studies. In this stage, which was the beginning of the study period, the response rate also tended to increase over time. The response rate increased to 120% on recruitment days 17–18, when many new participants were enrolled, and the response rate remained high for a significant period of time (Figure 9).

8 Discussion

Our method and platform have technical value in combining various independent and unrelated methods. We integrated a clinical trial protocol and database, semantic correlation calculation, 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 apps. Although there are many types of participatory trials, such as crowdsourcing and crowd-science, these approaches still have critical limitations in unstructured data collection and data sharing. We suggested a new concept, which is a participatory trial, to combine beneficial characteristics of crowd-science and crowdsourcing. Then, we proposed a solution to solve problems on interpretable data preparation, systematic data storage, and sustainable data sharing stages. Our method and platform have great value in presenting the primary technique that can validate the effectiveness of various

healthcare apps.

Minimizing bias occurred intentionally or unintentionally at all stages of data collection is another important factor in ensuring data reliability. We identified and considered each bias of the data collection stages according to bias in research [81, 99]. In the stage of data preparation, we included flawed study design, channelling bias and selection bias.We tried to prevent flawed study design by using the existing clinical trial protocols. We minimized channelling bias, which passes only specific participants, by placing unconditional eligible criteria that everyone can participate in a study. The editable eligible criteria of a creator such as a gender and age is still a matter to consider for the channelling bias. However, the selection bias that can distort the data tendency is not considered. This is a difficult problem to solve in the field of web-based data collection [9]. As a solution, we propose the social relationship distance estimation method of participants to keep them beyond a certain distance. A social network estimation method of participants to calculate a closeness rank among participants is an example method for cutting off a certain distance or less [92]. In the stage of data storage and data sharing, we included interviewer bias and transfer bias. The interviewer is that affects participants by the data collector. Our data storage method does only involve mechanical data collection functions, so we predicted that cannot occur interviewer bias. However, we do not have a preventive measure against interviewer bias caused by the study description or the CRF items that a creator manually generates.The solution is to clearly define the digital biomarkers that a creator seek to evaluate the effectiveness of a digital healthcare app [23]. Further research and consensus from experts are needed to enable it for standardized participatory trial protocols with digital biomarkers. The transfer bias in our method is that data is exposed during the data storage stage and other participants may be affected by the stored data. We prevented the bias by hiding the stored data from being exposed during a study. Finally, we did not consider the bias generated in the stage of data analysis because it is out of the scope of this article only dealing with data collection methods.

In the DRW test, we did not get enough data from the "less than 95 HITAR" group of workers due to sequential access procedures of the external platform, unlike the test in MTurk. Actually, we prepared a test to see what the results were with DRW in a real experimental environment of our platform. We created a temporary ID generator for participants to ensure anonymity and issued a temporary ID per request. Participants participated in a test study to investigate the efficacy of the healthcare app using a random ID. This experiment was designed to be as close as possible to the actual research process. Participants entered consent, eligibility, symptom, and data type eCRFs. Participants agreed to an electronic consent, confirmed their eligibility for the study, and measured their current symptoms. They entered their data into an eCRF that was configured to measure the effectiveness of the app. Data is entered twice for two weeks on a weekly basis. We measured each DRW based on the data entered for the symptom and data type eCRF. We recruited the participants from MTurk, and we decided on the group threshold to find a correlation between DRW and HITAR in our platform. As a result of the test, we had 58 participants from 103 workers (Figure 10). We found that only 28 participants completed all data inputs to calculate DRW. We calculated the DRW based on the average value of RT, LSA, IR, and SR. The average DRW of the "greater than 95" group was 0.94, and the opposite group was 0.91 (Supplementary data 4). This result would support that HITAR and DRW have a relation in the actual platform. However, the collected data size was too small to conduct a statistical test. Furthermore, we could not collect data from all distributions of HITAR participants. The reason was that most of the workers of MTurk had more than 95 HITAR to our best knowledge. This made it challenging to calculate the correlation between HITAR and DRW based on all the distributions. We also found that

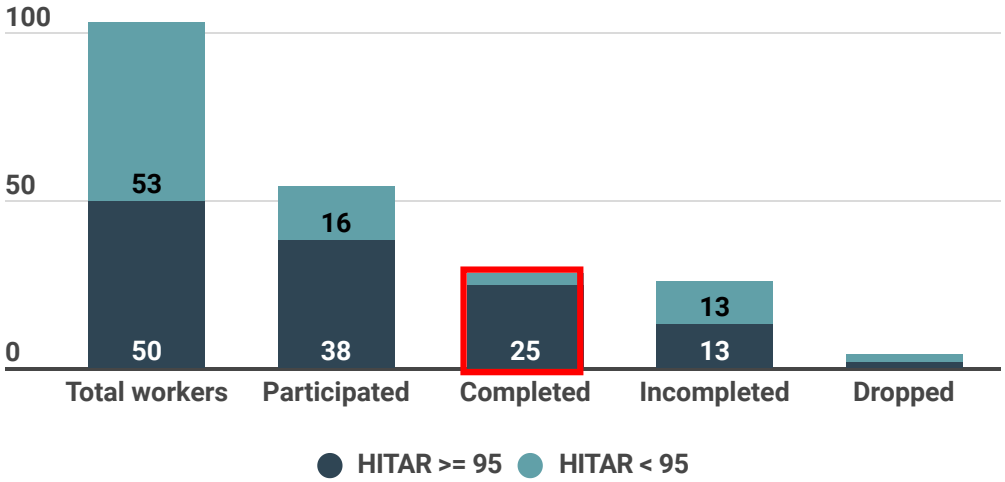


Figure 10. Participation statistic HITAR statistics worker groups classified by participation level.

not specifying the endpoint of the recruitment of the DRW test brought slightly different results. We estimated that the reason is that late participants of the survey tended to fill out the questionnaire to improve their HITAR scores rather than those who participated earlier.

There are additional considerations that may improve the given platform. First, the current platform cannot provide a quantitative comparison of the search results of the clinical trial protocol. The platform takes a symptom from the creator and provides all relevant clinical trials. However, many clinical trials result in a search; this requires additional analysis by the creator. In clinical trials, various types of elements—not only symptoms but also genes and chemical compounds—exist. This information can be used to retrieve clinical trial data required by the user. If an advanced platform is developed in the future that can quantitatively assess the relevance of all the elements involved in the clinical trial, it will reduce the inconvenience of the creator having to perform additional analysis. Second, advanced DRW methods should be developed. For example, the basic DRW can be used to assess the reliability of the input data based only on the length of the input text, as in long-string analysis (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. 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 check whether the study results are the same as the verified effects.

All these limitations can be considered for further experiments or improved computational methods. Moreover, we suggest further studies since the data collection methods are the basics of the digital healthcare field that is a work in progress. The suggesting studies are following; (1) Developing a standard of participatory trial protocol method considered digital markers and confounding factors, (2) Advancing protocol creation methods for specific parameter determination such as the number of uses, effectiveness report counts, study duration or the number of participants, (3) Natural language question model to generate DRW index questions rather than using human managing templates, (4) Optimizing data reliability weight calculation with the importance value of each index, (5) Effectiveness score calculation method based on the collected data with DRW scores. With these further improvements and the further studies, our platform can be used as a valuable tool in the assessment of the

effectiveness of various apps in a cost-effective manner.

9 conclusion

On the first stage of verification workflow, we proposed the reliable data collection method and platform to assess the effectiveness of digital healthcare apps. We presented a participatory trial concept for a new type of data collection trial that uses voluntary participation of end-users of digital healthcare apps. Then, we described the essential methods of the reliable data collection platform based on the participatory trial concept. Interpretable data preparation methods were protocol database and retrieval system to create a protocol without expert knowledge. We validated the simplicity of the methods from compared USE scores between the proposed system and the existing system. We developed DRW calculation method for systematic data storage. DRW score showed to be a reliable measure through correlation with HITAR. To sustainable data collection and sharing, we developed a reward distribution method. We indirectly observed the effect of reward using DRW. The effect of reward presented increasing DRW, that is to make more effort of participants to sustainable data collection. Finally, we implemented a pilot platform, CORUS. CORUS integrated all the methods and essential features for a participatory trial platform. We compared the features with existing platforms on the field of clinical trial, pragmatic trial and participatory trial. We assert that CORUS has all the necessary features to collect reliable data from the users of digital healthcare apps on the path of validation of their effectiveness.

Code Availability

We provide all the source code at `https://github.com/junseokpark/corus` upon the 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

Thanks to Jooyeon Lee and Inhwan Hwang of Amazon Web Services Korea LLC (AWS Korea) Worldwide Public Sector for their support with cloud computing platform services and demonstrations.

References

1. Industry research shows 97% of companies to increase use of real-world patient data for more accurate decision-making.
2. Data sharing and the future of science. *Nature Communications*, 9(1), 2018.
3. How the assessment works, 2019.
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*, 12(4):e0172616, 2017.
5. Apple. App store, 2008.
6. A. J. Apter. Understanding adherence requires pragmatic trials: lessons from pediatric asthma. *JAMA pediatrics*, 169(4):310–311, 2015.
7. 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*, 22(5):441–449, 2019.
8. V. R. Basili. Data collection, validation and analysis. *Software Metrics: An Analysis and Evaluation*, pages 143–160, 1981.
9. J. Bethlehem. Selection bias in web surveys. *International Statistical Review*, 78(2):161–188, 2010.
10. 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*, 3:160011, 2016.
11. 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*, 374(22):2175–2181, 2016. PMID: 27248626.
12. M. N. K. Boulos, A. C. Brewer, C. Karimkhani, D. B. Buller, and R. P. Dellavalle. Mobile medical and health apps: state of the art, concerns, regulatory control and certification. *Online journal of public health informatics*, 5(3):229, 2014.

13. 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*, 13(6):e0199242, 2018.

14. 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*, pages 133–139, 2016.

15. 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*, 371(9624):1595–1602, 2008.

16. G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

17. 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*, volume 5, page S2. Springer, 2015.

18. 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*, 10(1):450, 2015.

19. 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*, 158(3):200–207, 2013.

20. A. E. C. Cloud. Amazon web services. *Retrieved November*, 9(2011):2011, 2011.

21. J. Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

22. F. T. Commission et al. Mobile health apps interactive tool, 2016.

23. A. Coravos, S. Khozin, and K. D. Mandl. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ digital medicine*, 2(1):1–5, 2019.

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

25. N. DANBURY, Conn. & RESEARCH TRIANGLE PARK. Iqvia institute for human data science study: Impact of digital health grows as innovation, evidence, and adoption of mobile health apps accelerate, 2017.

26. N. O. Davis, J. Bower, and S. H. Kollins. Proof-of-concept study of an at-home, engaging, digital intervention for pediatric adhd. *PloS one*, 13(1):e0189749, 2018.

27. N. Dhand and M. Khatkar. Statulator: An online statistical calculator. sample size calculator for estimating a single mean, 2014.

28. 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*, 47:20–33, 2016.

29. C. Eickhoff and A. P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

30. E. Elenko, A. Speier, and D. Zohar. A regulatory framework emerges for digital medicine. *Nature biotechnology*, 33(7):697, 2015.

31. 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*, 33(8):1037–1044, 2016.

32. A. S. Figueiredo. Data sharing: convert challenges into opportunities. *Frontiers in public health*, 5:327, 2017.

33. I. Ford and J. Norrie. Pragmatic trials. *New England journal of medicine*, 375(5):454–463, 2016.

34. 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.

35. F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

36. 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*, volume 62, pages 1414–1418. SAGE Publications Sage CA: Los Angeles, CA, 2018.

37. S. L. George and M. Buyse. Data fraud in clinical trials. *Clinical investigation*, 5(2):161, 2015.

38. S. Gessner, M. Storck, S. Hegselmann, M. Dugas, and I. Soto-Rey. Automated transformation of cdisc odm to openclinica. In *GMDS*, pages 95–99, 2017.

39. K. A. Getz and R. A. Campo. New benchmarks characterizing growth in protocol design complexity. *Therapeutic innovation & regulatory science*, 52(1):22–28, 2018.

40. Google. Google play, 2012.

41. D. Haak, C.-E. Page, S. Reinartz, T. Krüger, and T. M. Deserno. Dicom for clinical research: Pacs-integrated electronic data capture in multi-center trials. *Journal of digital imaging*, 28(5):558–566, 2015.

42. 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*, 378(24):2302–2310, 2018.

43. 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*, page 449. ACM, 2018.

44. L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

45. J. Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

46. 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*, 30(2):299–311, 2015.

47. 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*, 27(1):99–114, 2012.

48. 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*, 91(3):277–285, 2001.

49. 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*, 5:150–159, 2017.

50. J. A. Johnson. Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality*, 39(1):103–129, 2005.

51. S. T. Karen Sharma. Pear therapeutics receives expedited access pathway designation from fda for reset-o™ prescription digital therapeutic to treat opioid use disorder, 2017.

52. J. S. Katz and B. R. Martin. What is research collaboration? *Research policy*, 26(1):1–18, 1997.

53. N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, volume 11, pages 1–11. Detroit, Michigan, USA, 2011.

54. 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*, 18(11):e292, 2016.

55. M. Kim and J. Chung. Sustainable growth and token economy design: The case of steemit. *Sustainability*, 11(1):167, 2019.

56. 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*, pages 453–456. ACM, 2008.

57. 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*, 76:431–437, 2017.

58. L. E. Kruger and M. A. Shannon. Getting to know ourselves and our places through participation in civic social assessment. *Society & Natural Resources*, 13(5):461–478, 2000.

59. 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*, 17(7):e183, 2015.

60. Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

61. 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.

62. 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*, 106(10), 2014.

63. 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*, pages 278–292, 1960.

64. 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*, 30(6):1251–1266, 2018.

65. A. M. Lund. Measuring usability with the use questionnaire. *Usability interface*, 8(2):3–6, 2001.

66. W. Maalej, Z. Kurtanović, H. Nabil, and C. Stanik. On the automatic classification of app reviews. *Requirements Engineering*, 21(3):311–331, 2016.

67. W. Martin, M. Harman, Y. Jia, F. Sarro, and Y. Zhang. The app sampling problem for app store mining. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, pages 123–133. IEEE Press, 2015.

68. B. C. Martinson, M. S. Anderson, and R. De Vries. Scientists behaving badly. *Nature*, 435(7043):737, 2005.

69. 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*, 2(1):38, 2019.

70. S. McCaffrey, R. A. Black, M. Nagao, M. Sepassi, G. Sharma, S. Thornton, Y. H. Kim, and J. Braverman. Measurement of quality of life in patients with mycosis fungoides/sézary syndrome cutaneous t-cell lymphoma: Development of an electronic instrument. *Journal of medical Internet research*, 21(1):e11302, 2019.

71. 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*, 136(3):450, 2010.

72. K. O. McGraw and S. Wong. A common language effect size statistic. *Psychological bulletin*, 111(2):361, 1992.

73. A. W. Meade and S. B. Craig. Identifying careless responses in survey data. *Psychological methods*, 17(3):437, 2012.

74. 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*, 3(2):e40, 2015.

75. R. G. Mira Zuidgeest, Iris Goetz. Study design: Pragmatic trial.

76. 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*, 5(5):e70, 2017.

77. NHS. Nhs apps library, 2018.

78. NHS. Business software and services reviews, 2020.

79. 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*, 6(1):13, 2018.

80. T. Nugent, D. Upton, and M. Cimpoesu. Improving data transparency in clinical trials using blockchain smart contracts. *F1000Research*, 5, 2016.

81. C. J. Pannucci and E. G. Wilkins. Identifying and avoiding bias in research. *Plastic and reconstructive surgery*, 126(2):619, 2010.

82. G. Paolacci and J. Chandler. Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188, 2014.

83. J. Park, K. Kim, W. Hwang, and D. Lee. Concept embedding to measure semantic relatedness for biomedical information ontologies. *Journal of biomedical informatics*, 94:103182, 2019.

84. 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*, page 814996, 2019.

85. 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)*, pages 181–184. IEEE, 2018.

86. D. L. Paulhus, P. D. Harms, M. N. Bruce, and D. C. Lysy. The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of personality and social psychology*, 84(4):890, 2003.

87. 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*, page 103216, 2019.

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

89. M. Z. Rachana Jain. 2017 year end funding report: The end of the beginning of digital health, 2017.

90. P. M. Rothwell. Factors that can affect the external validity of randomised controlled trials. *PLoS clinical trials*, 1(1):e9, 2006.

91. B. Saunders, J. Kitzinger, and C. Kitzinger. Participant anonymity in the internet age: from theory to practice. *Qualitative research in psychology*, 12(2):125–137, 2015.

92. 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*, 9(1):3, 2019.

93. M.-G. Seo and L. F. Barrett. Being emotional during decision making—good or bad? an empirical investigation. *Academy of Management Journal*, 50(4):923–940, 2007.

94. 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*, 17(4):e92, 2015.

95. 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.

96. C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

97. S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

98. 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*, 375(23):2293–2297, 2016.

99. A.-M. Simundic. Bias in research. *Biochemia medica: Biochemia medica*, 23(1):12–15, 2013.

100. 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)*, 5(1):11, 2016.

101. H. C. Sox and R. J. Lewis. Pragmatic trials: practical answers to "real world" questions. *Jama*, 316(11):1205–1206, 2016.

102. S. S. Sundar, A. Oeldorf-Hirsch, and Q. Xu. The bandwagon effect of collaborative filtering technology. In *CHI'08 extended abstracts on Human factors in computing systems*, pages 3453–3458. ACM, 2008.

103. L. Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671:1–34, 2000.

104. 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*, 21(1):e11683, 2019.

105. 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*, 62(5):464–475, 2009.

106. R. Tonkens. An overview of the drug development process. *Physician executive*, 31(3):48, 2005.

107. 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*, 100(6):713–729, 2016.

108. K. Tucker, J. Branson, M. Dilleen, S. Hollis, P. Loughlin, M. J. Nixon, and Z. Williams. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC medical research methodology*, 16(1):77, 2016.

109. 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*, 390(10104):1737–1746, 2017.

110. V. H. Vroom. *Some personality determinants of the effects of participation*. Routledge, 2019.

111. 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*, 7(3):e015276, 2017.

112. 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*, 6(9):e25348, 2011.

113. E. Waltz. Pear approval signals fda readiness for digital treatments. *Nature Biotechnology*, 36(6):481–482, 2018.

114. 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*, 12(2):e19, 2010.

115. F. Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

116. 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*, 53(4):43–46, 2014.

117. S. L. Wise and X. Kong. Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2):163–183, 2005.

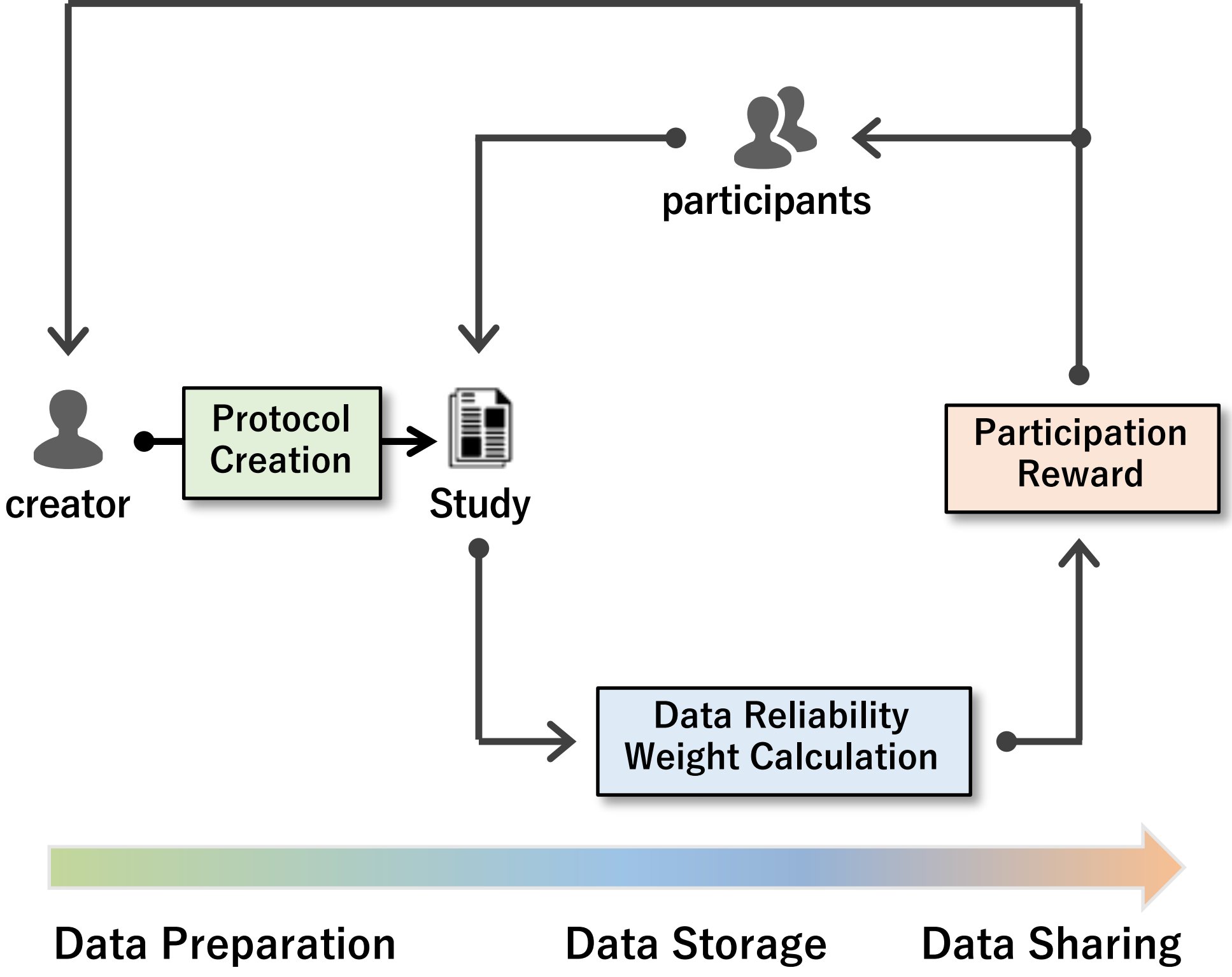
118. D. R. Wong, S. Bhattacharya, and A. J. Butte. Prototype of running clinical trials in an untrustworthy environment using blockchain. *Nature communications*, 10(1):917, 2019.

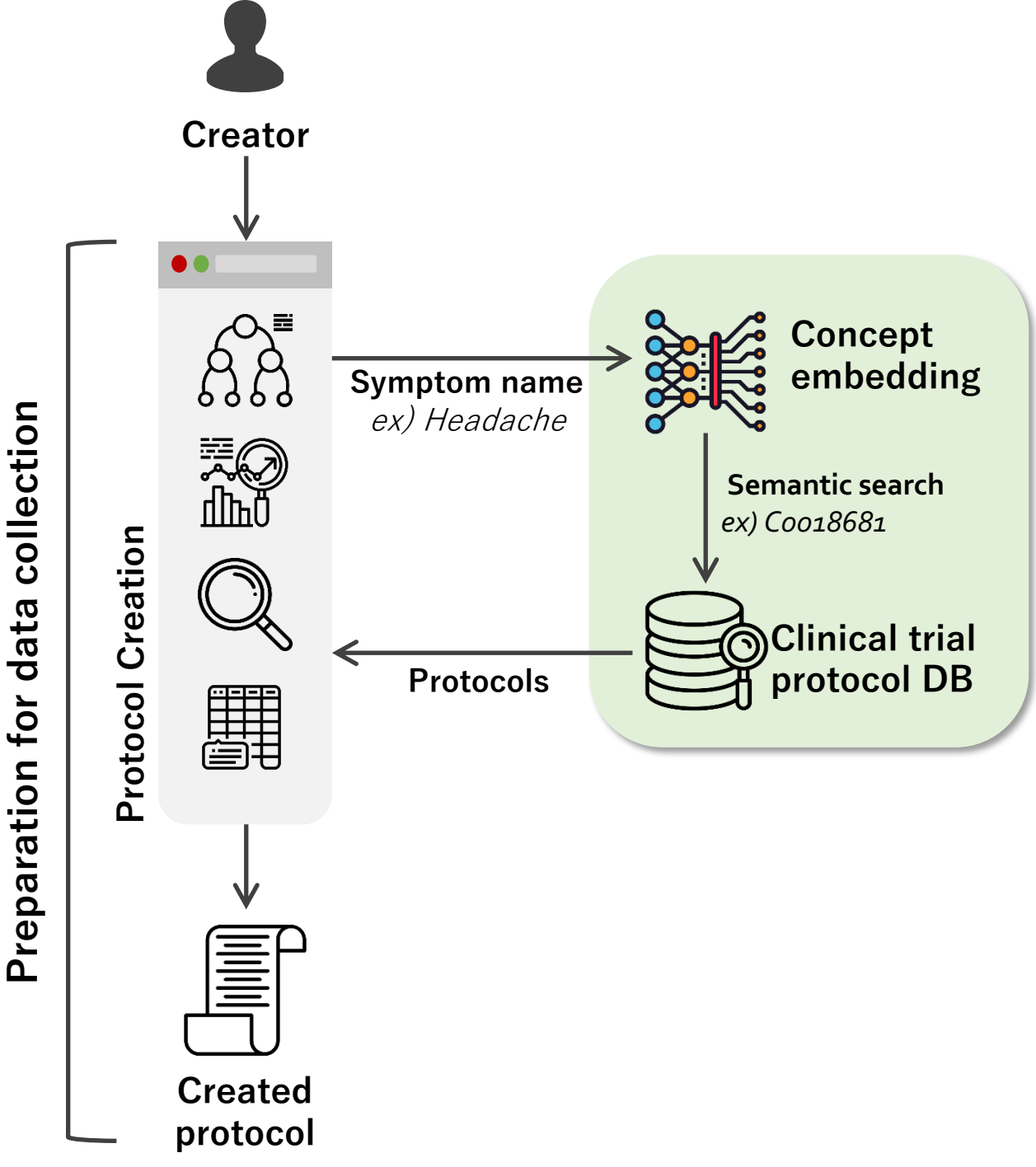
119. B. E. Yerys, J. R. Bertollo, L. Kenworthy, G. Dawson, E. J. Marco, R. T. Schultz, and L. Sikich. Brief report: Pilot study of a novel interactive digital treatment to improve cognitive control in children with autism spectrum disorder and co-occurring adhd symptoms. *Journal of autism and developmental disorders*, 49(4):1727–1737, 2019.

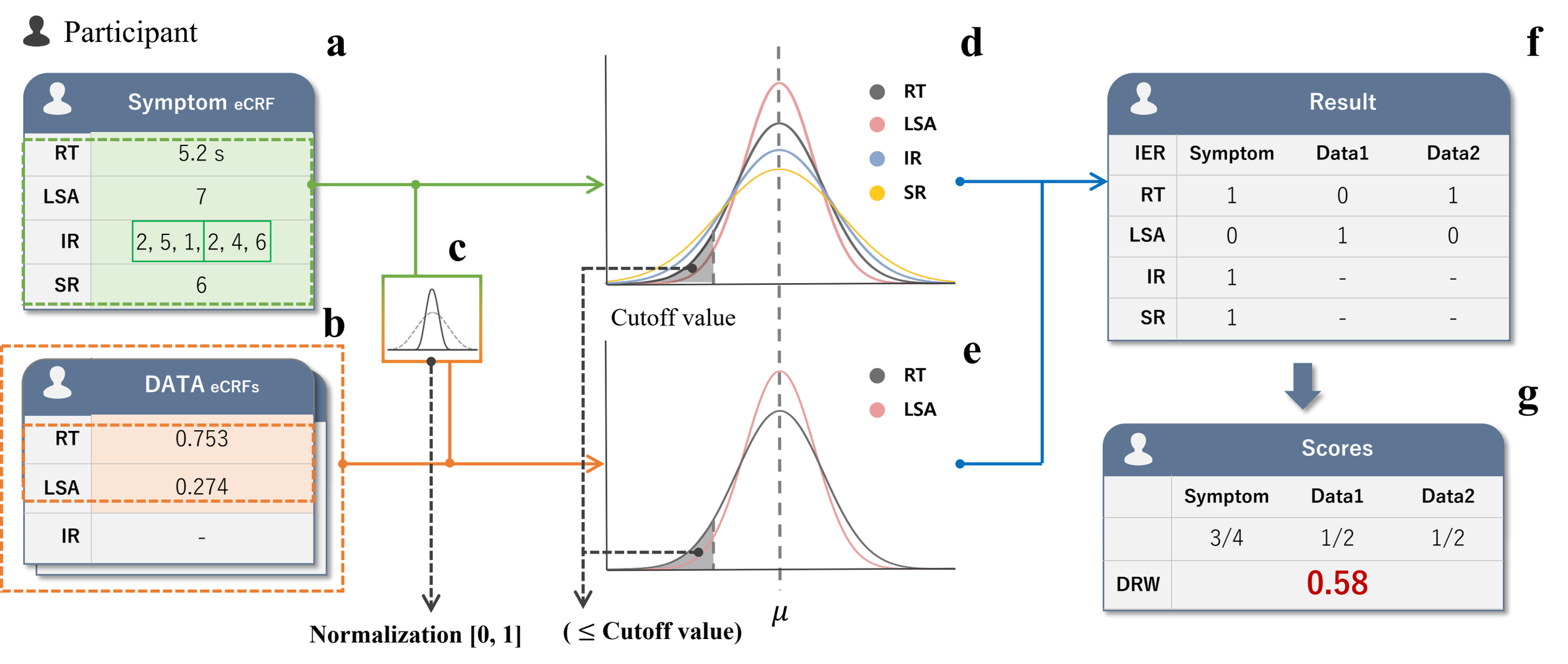
120. D. A. Zarin and A. Keselman. Registering a clinical trial in clinicaltrials. gov. *Chest*, 131(3):909–912, 2007.

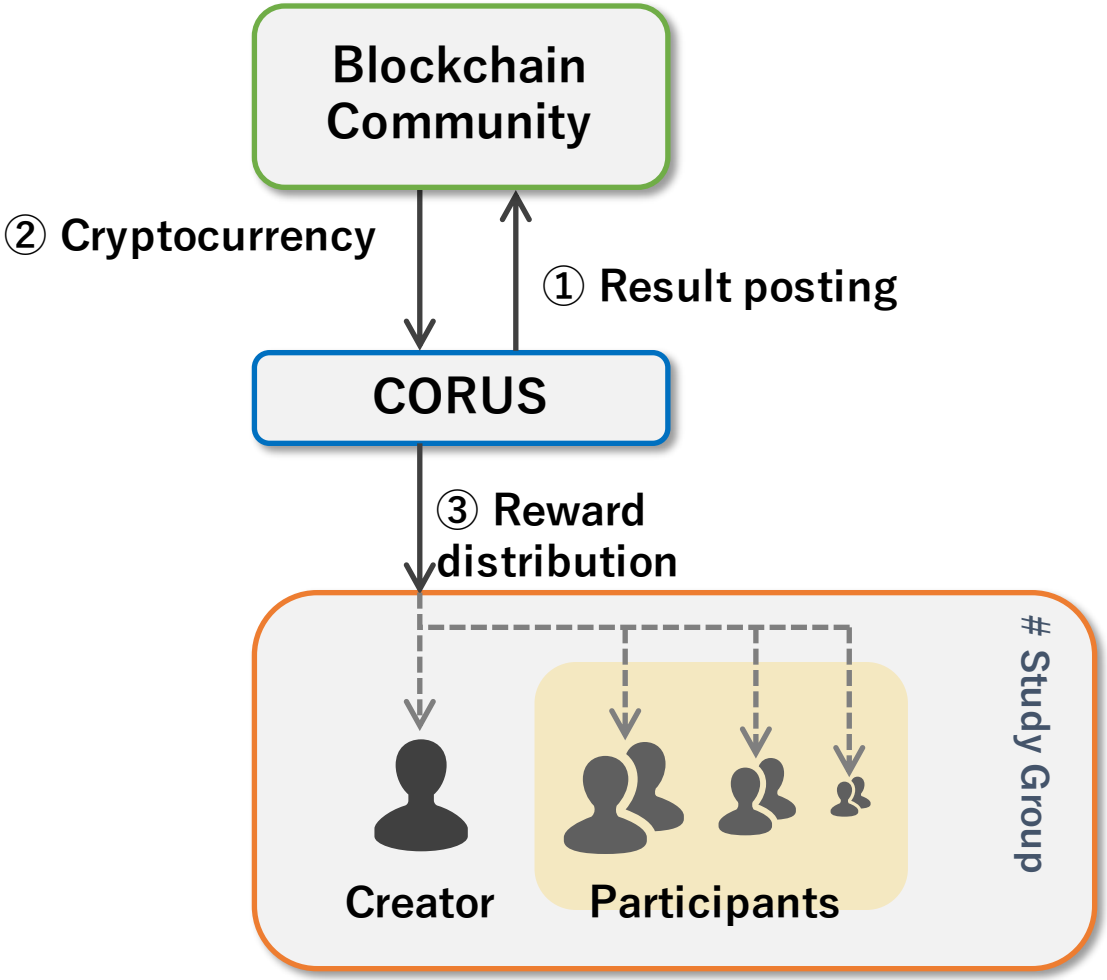
121. S. Zhang, C. Zhang, and Q. Yang. Data preparation for data mining. *Applied artificial intelligence*, 17(5-6):375–381, 2003.

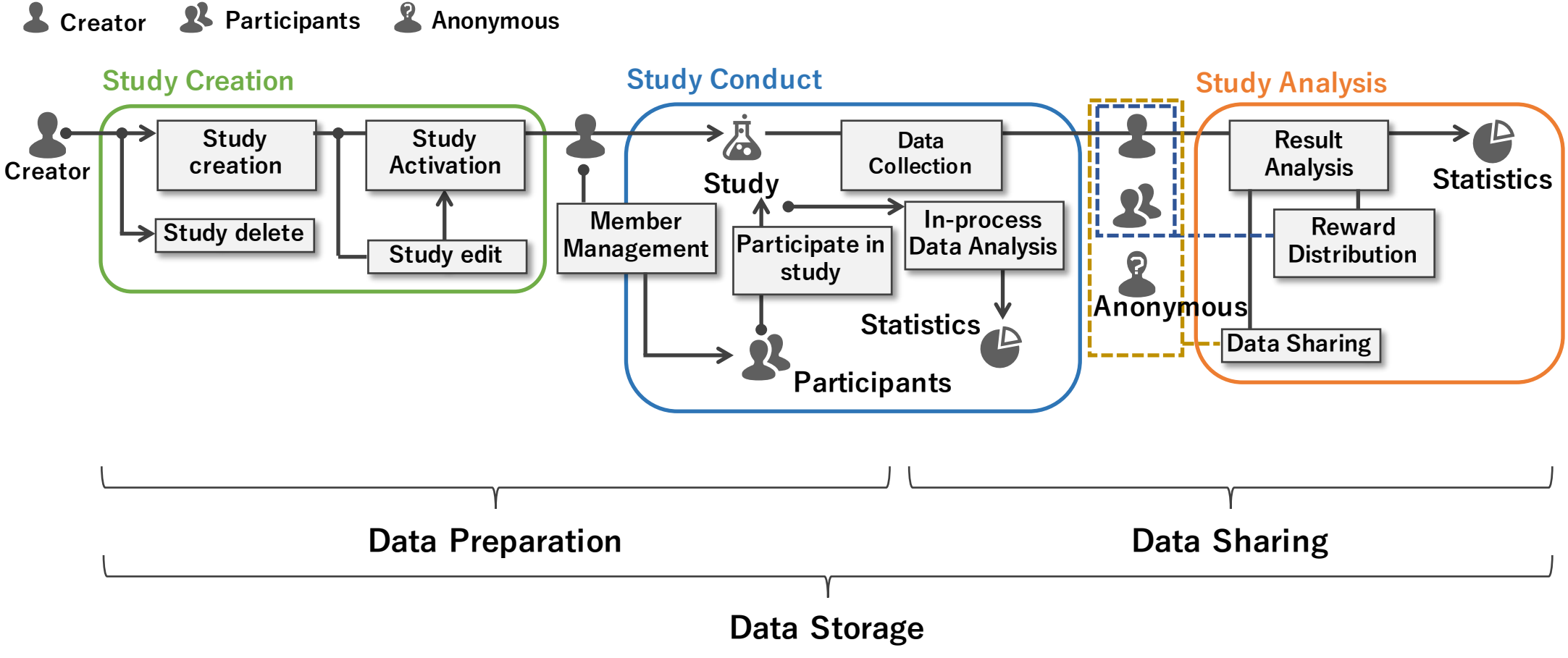
122. D. A. Ziegler, A. J. Simon, C. L. Gallen, S. Skinner, J. R. Janowich, J. J. Volponi, C. E. Rolle, J. Mishra, J. Kornfield, J. A. Anguera, et al. Closed-loop digital meditation improves sustained attention in young adults. *Nature human behaviour*, page 1, 2019.

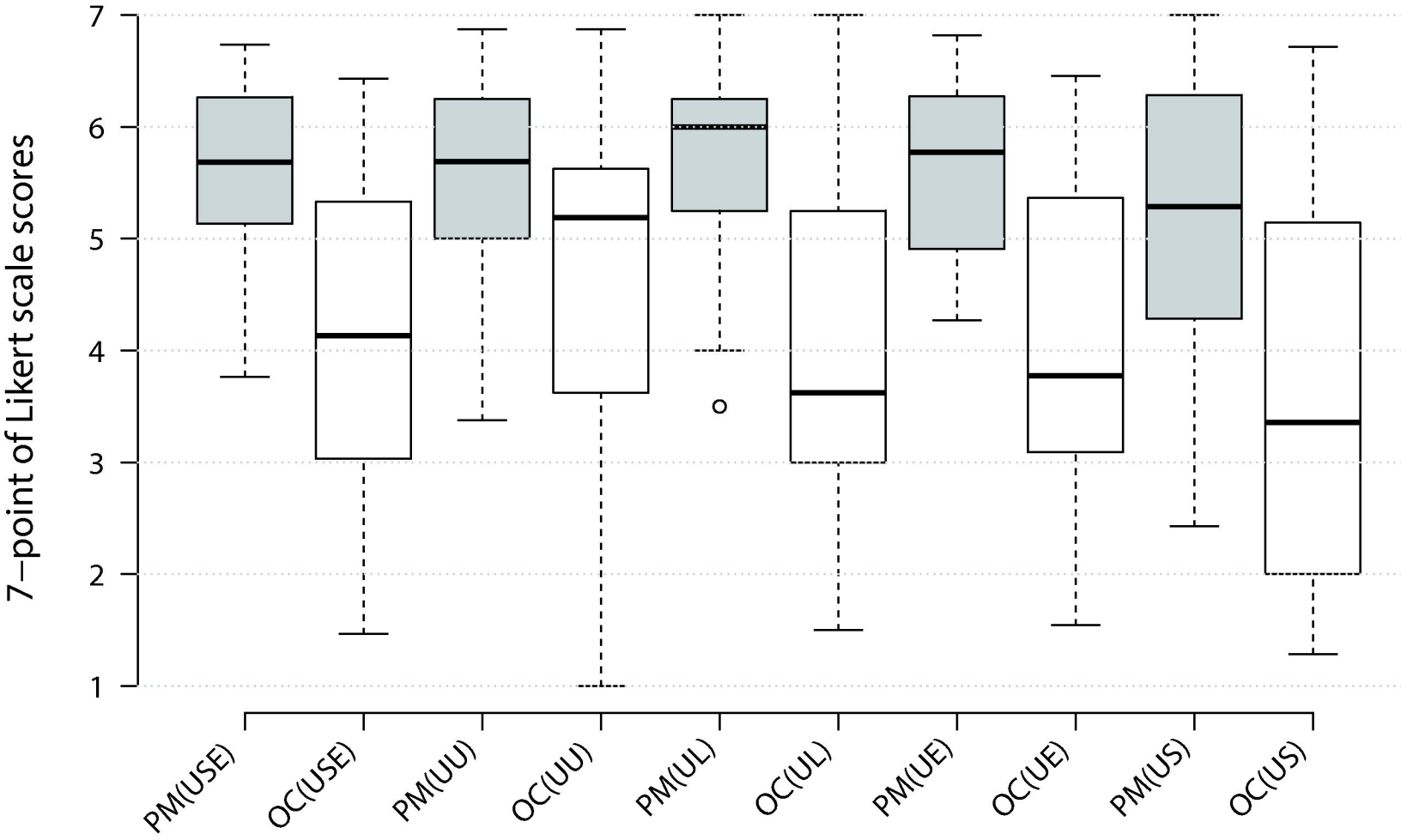

participants
creator
Protocol Creation
Study
Participation Reward
Data Reliability Weight Calculation
Data Preparation
Data Storage
Data Sharing

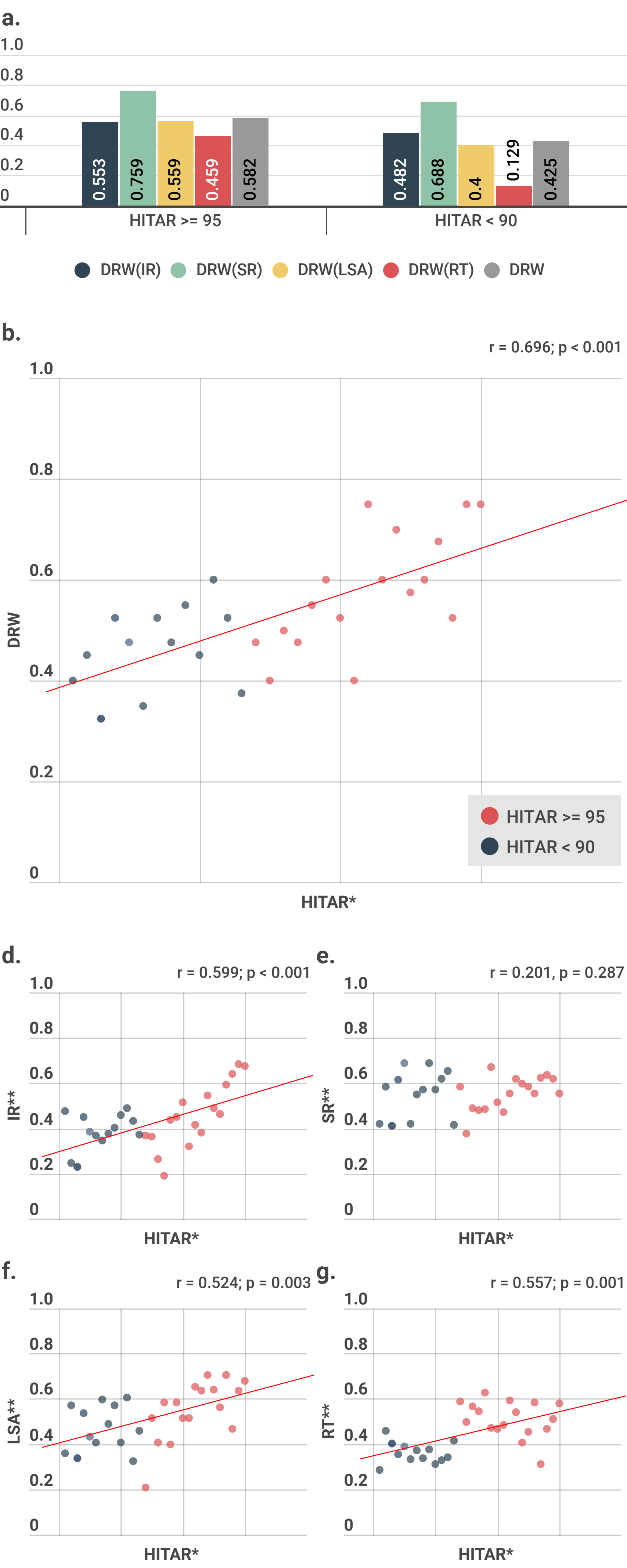

Creator
Preparation for data collection
Protocol Creation
Symptom name
ex) Headache
Concept embedding
Semantic search
ex) C0018681
Clinical trial protocol DB
Protocols
Created protocol

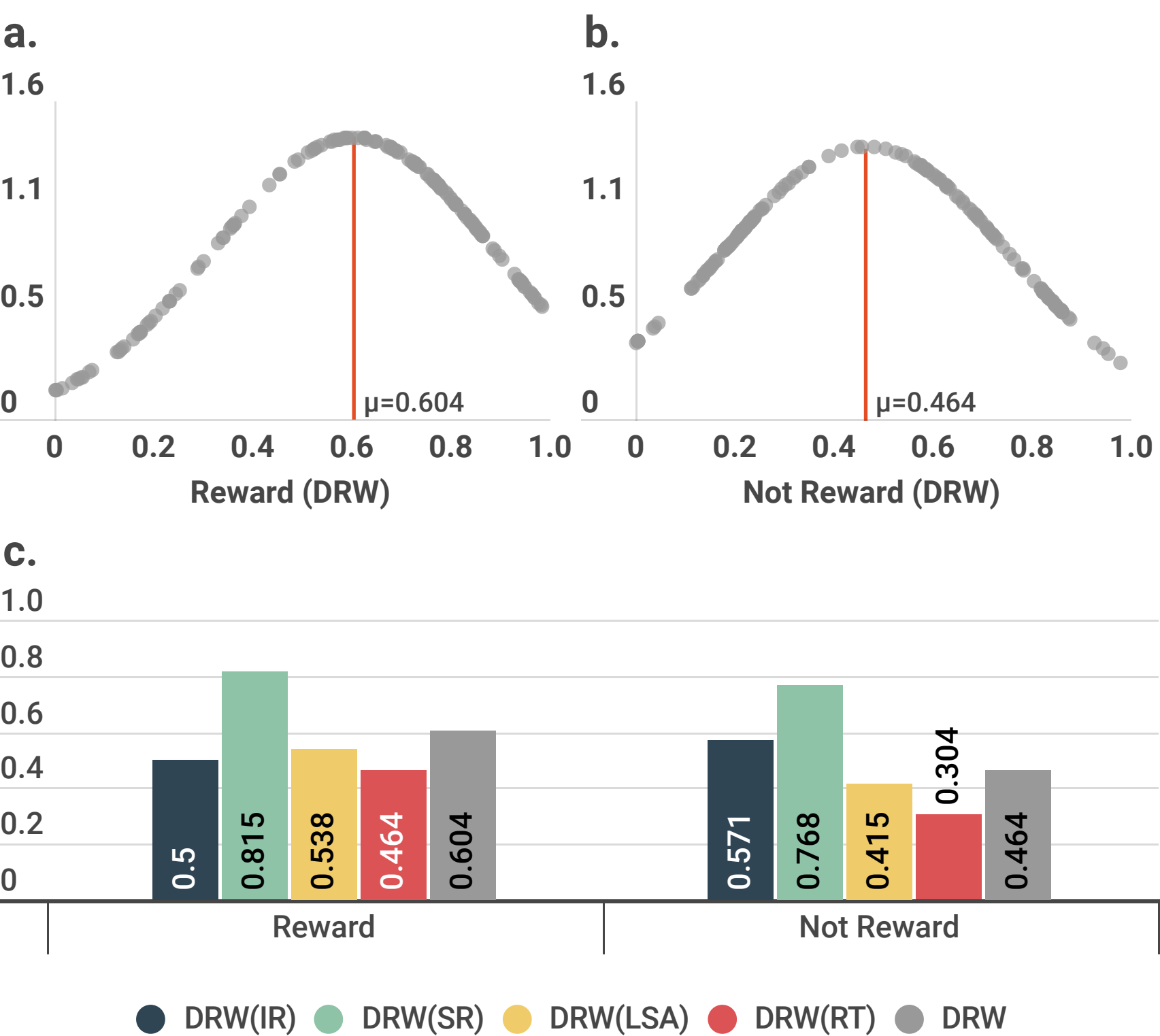

Participant
a
Symptom eCRF
RT 5.2 s
LSA 7
IR 2, 5, 1, 2, 4, 6
SR 6
b
DATA eCRFs
RT 0.753
LSA 0.274
IR -
c
Normalization [0, 1]
d
RT
LSA
IR
SR
Cutoff value
e
RT
LSA
(≤ Cutoff value)
μ
f
Result
IER Symptom Data1 Data2
RT 1 0 1
LSA 0 1 0
IR 1 - -
SR 1 - -
g
Scores
Symptom Data1 Data2
3/4 1/2 1/2
DRW 0.58


Blockchain Community
② Cryptocurrency
① Result posting
CORUS
③ Reward distribution
Study Group
Creator
Participants

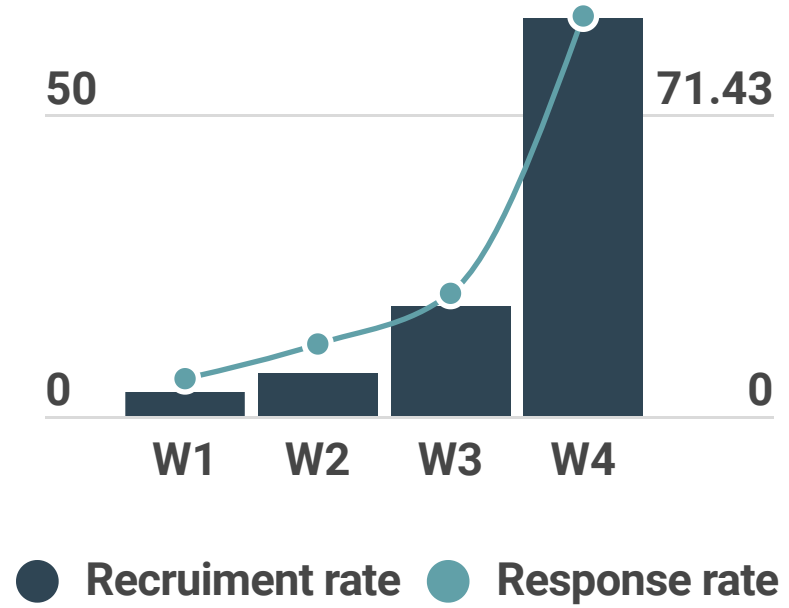

Creator
Participants
Anonymous
Study Creation
Study Conduct
Study Analysis
Creator
Study creation
Study Activation
Study delete
Study edit
Member Management
Study
Data Collection
Participate in study
In-process Data Analysis
Statistics
Participants
Anonymous
Result Analysis
Statistics
Reward Distribution
Data Sharing
Data Preparation
Data Sharing
Data Storage


7−point of Likert scale scores
7
6
5
4
3
2
1
PM(USE)
OC(USE)
PM(UU)
OC(UU)
PM(UL)
OC(UL)
PM(UE)
OC(UE)
PM(US)
OC(US)

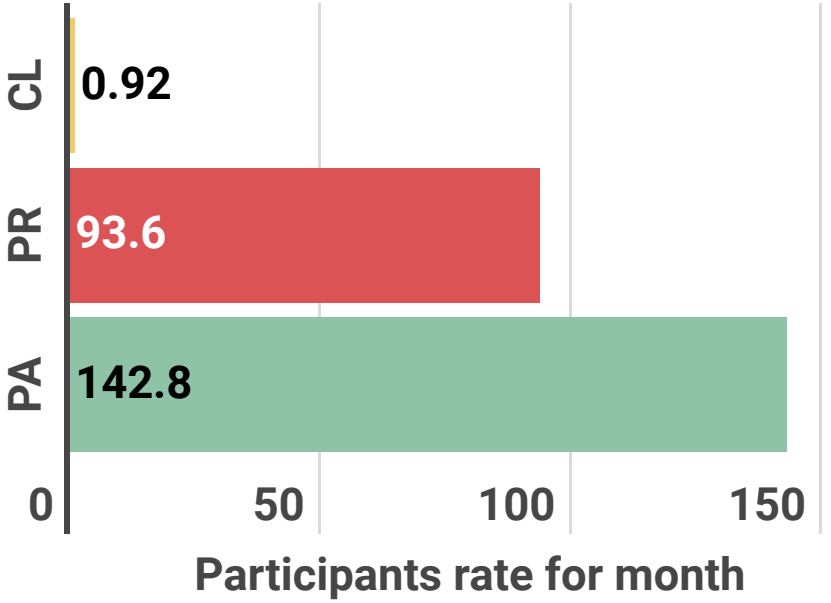

a.
1.0
0.8
0.6
0.4
0.2
0
0.553
0.759
0.559
0.459
0.582
0.482
0.688
0.4
0.129
0.425
HITAR >= 95
HITAR < 90
DRW(IR)
DRW(SR)
DRW(LSA)
DRW(RT)
DRW
b.
r = 0.696; p < 0.001
DRW
HITAR*
HITAR >= 95
HITAR < 90
d.
r = 0.599; p < 0.001
IR**
HITAR*
e.
r = 0.201, p = 0.287
SR**
HITAR*
f.
r = 0.524; p = 0.003
LSA**
HITAR*
g.
r = 0.557; p = 0.001
RT**
HITAR*

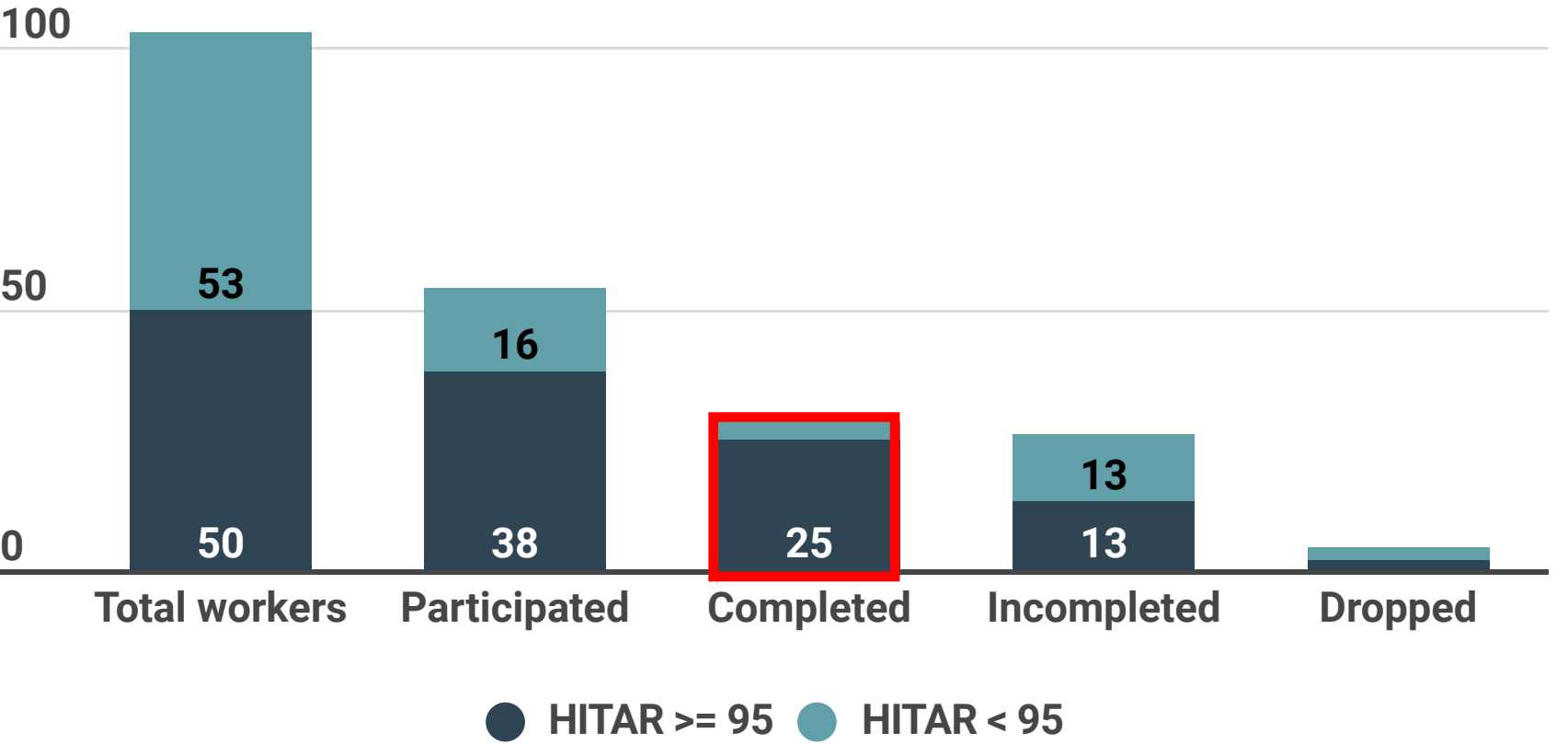

a.
1.6
1.1
0.5
0
μ=0.604
0
0.2
0.4
0.6
0.8
1.0
Reward (DRW)
b.
1.6
1.1
0.5
0
μ=0.464
0
0.2
0.4
0.6
0.8
1.0
Not Reward (DRW)
c.
1.0
0.8
0.6
0.4
0.2
0
0.5
0.815
0.538
0.464
0.604
0.571
0.768
0.415
0.304
0.464
Reward
Not Reward
DRW(IR)
DRW(SR)
DRW(LSA)
DRW(RT)
DRW

a.


50
0
71.43
0
W1
W2
W3
W4
Recruiment rate
Response rate

b.


CL
PR
PA
0.92
93.6
142.8
0
50
100
150
Participants rate for month


100
50
0
53
50
16
38
25
13
13
Total workers
Participated
Completed
Incompleted
Dropped
HITAR >= 95
HITAR < 95