

Research and Applications

Using dataflow diagrams to support research informed consent data management communications: participant perspectives

Brian J. McInnis, PhD¹, Ramona Pindus, BS², Daniah H. Kareem, BSPH², Julie Cakici, PhD, RN², Daniela G. Vital, MPH, BSPH², Eric Hekler, PhD², Camille Nebeker, EdD, MS^{*,2}

¹School of Information, University of Texas Austin, Austin, TX 78712, United States, ²Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA 92093, United States

*Corresponding author: Camille Nebeker, EdD, MS, Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, 9500 Gilman Drive, MC 0811, La Jolla, CA 92093, United States (nebeker@health.ucsd.edu)

Abstract

Objectives: Digital health research involves collecting vast amounts of personal health data, making data management practices complex and challenging to convey during informed consent.

Materials and Methods: We conducted eight semi-structured focus groups to explore whether dataflow diagrams (DFD) can complement informed consent and improve participants' understanding of data management and associated risks ($N=34$ participants).

Results: Our analysis found that DFDs could supplement text-based information about data management and sharing practices, such as by helping raise new questions that prompt conversation between prospective participants and members of a research team. Participants in the study emphasized the need for clear, simple, and accessible diagrams that are participant centered. Third-party access to data and sharing of sensitive health data were identified as high-risk areas requiring thorough explanation. Participants generally agreed that the design process should be led by the research team, but it should incorporate many diverse perspectives to ensure the diagram was meaningful to potential participants who are likely unfamiliar with data management. Nearly all participants rejected the idea that artificial intelligence could identify risks during the design process, but most were comfortable with it being used as a tool to format and simplify the diagram. In short, DFDs may complement standard text-based informed consent documents, but they are not a replacement.

Discussion: Prospective research participants value diverse ways of learning about study risks and benefits. Our study highlights the value of incorporating information visualizations, such as DFDs, into the informed consent procedures to participate in research.

Conclusion: Future research should explore other ways of visualizing consent information in ways that help people to overcome digital and data literacy barriers to participating in research. However, creating a DFD requires significant time and effort from research teams. To alleviate these costs, research sponsors can support the creation of shared infrastructure, communities of practice, and incentivize researchers to develop better consent procedures.

Key words: digital health; data management; privacy; informed consent; dataflow diagram; research ethics.

Introduction

Digital health technologies enable people to receive personalized health information conveniently via their smart and connected devices. These tools, including apps and wearable devices, have a high level of precision due to the array of data sensors embedded within the devices and the availability of cloud computing platforms that can rapidly translate streams of sensor-generated information and other data into health status updates and, in many cases, recommendations to improve wellness. Researchers have used digital health technologies for over two decades with the goal of advancing precision health.¹ The promise is that by providing people with in-situ and real-time feedback about their health, the technologies can facilitate people toward decisions that promote their well-being (ie, prompts to promote physical activity, medication adherence, anxiety management).^{1,2}

While digital health technologies hold the promise of promoting healthier lifestyles, the systems collect highly

personalized behavioral, biological, and ecological information about research participants using these devices. Furthermore, data collection and processing methods can involve sharing participant personal information with third-party services, which raise important questions such as who has access to participant data, for how long, and for what purposes.^{3–5} These questions and corresponding answers are important to convey during the informed consent process as they help prospective participants to evaluate study risks and benefits.^{4,6,7}

Information visualizations can communicate the nuance associated with data collection, processing, and sharing and should be explored in addition to text-based materials to make the consent process truly informed in instances, such as digital health technology research, when the evaluation of participation risks and benefits is especially complex or obtuse. The process of translating content into a visualization naturally involves making editorial choices about what information to include, exclude, highlight, and how. However,

choosing to highlight some content inherently means that other content is de-emphasized, if not removed from the diagram. For instance, digital health studies can involve the collection of a considerable range of data about study participants, but presenting every type might lead to a confusing visualization that is challenging for prospective participants to understand. While reducing the amount of data presented on a dataflow diagram (DFD) might help to improve readability, doing so may inadvertently impact trust in the study and the research team. Navigating these tensions in design is challenging. This paper presents study results exploring the potential value of incorporating information visualizations about data privacy policies into the informed consent process for digital health research.

The DFDs apply specific strategies for highlighting study risks (ie, points, pathways, regions, levels). The analysis sheds light on how prospective study participants in research respond to each strategy for highlighting risk and to the amount of detail presented (ie, minimal, maximal). In addition, our study invited participants to consider the process of authoring the information visualizations. While it is standard practice for a research team to develop the informed consent materials that prospective research participants use to consider digital health study risks, researchers might also apply a

community-based approach and even use a large language model (LLM), which is a type of generative artificial intelligence, to support parts of the consent design process. After reviewing each strategy for presenting study risk, participants were asked how different authorship conditions may factor into their feelings of trust in the DFDs: ie, created by the research team, created by an ethics board, created by a community panel, created by using a generative artificial intelligence platform.

Methods

This study called the Consent Language Evaluation and Redesign was supported by a bioethics administrative supplement as part of a larger digital health research project funded by the National Cancer Institute. A goal was to identify facilitators that may lead to improved informed consent processes for digital health research. This study was reviewed and approved by the Institutional Review Board at University of California San Diego.

Participant recruitment procedures

Participants were recruited from mid-March through May 2023 with study information shared via an internal listserv at

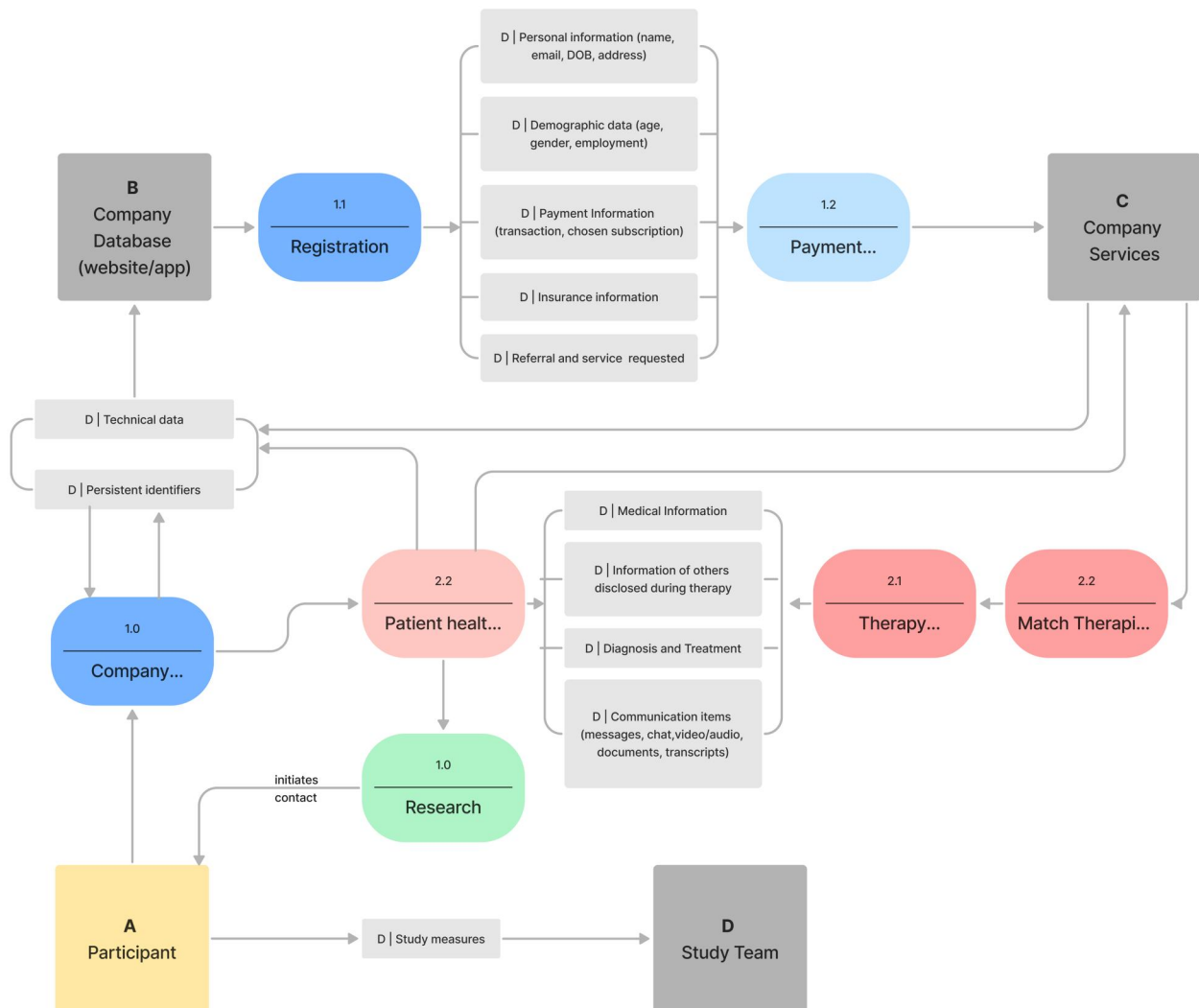


Figure 1. Baseline design for the Commercial DFD is based on the data privacy policy documents published on the company’s website. This Commercial DFD included a legend describing entities and processes depicted in this visualization.

University of California San Diego and to individuals who had previously expressed interest in research participation. Interested individuals were prescreened via an electronic survey to assess their eligibility, which included fluency in English and the ability to access an online collaboration system used to facilitate the focus groups. Those eligible and interested in participating reviewed the informed consent information electronically prior to participating in study activities.

Focus groups were assigned to a case study context

To prototype the focus group protocol, our team conducted a pilot co-design workshop that involved participants interacting virtually with a sample DFD. The pilot identified ways to improve the study instructions and procedures. After incorporating these learnings into the protocol, eight focus groups were held between April and May 2023. Each of the eight focus groups was assigned to review one of two study contexts (ie, commercial vendor privacy policy, Academic informed consent document).

The study contexts involved different data management practices and types of risk. The commercial vendor context reflected a mental health services provider involving talk therapy, which collects personally sensitive data from clients directly and during their therapeutic sessions. The data management policies specify that this data may be used for future academic and internal research (Commercial—Figure 1). Our

research team used the information presented on the company’s website to construct a DFD based on the data management policies, including information about data collection, storage, and data sharing for the purposes of improving clinical practices and secondary research.

The second context was modeled after an academic research study that involved participants completing a survey about their well-being and physical activity twice daily, while also using a commercially available smartwatch to monitor their physical fitness and receive automated notifications from the study team to encourage physical activity (Academic—Figure 2). In this case, the DFD was authored by the researchers involved with the study context and approved for use in their informed consent process by the local IRB.

A semi-structured protocol was used to facilitate the focus groups

To foster discussion about the study context information and DFDs, the focus group facilitator followed a semi-structured protocol that involved introducing the study context through four major parts. At each part, the facilitator provided participants with ample time to ask any clarifying questions and share their perspectives. The focus group protocol involved the following parts:

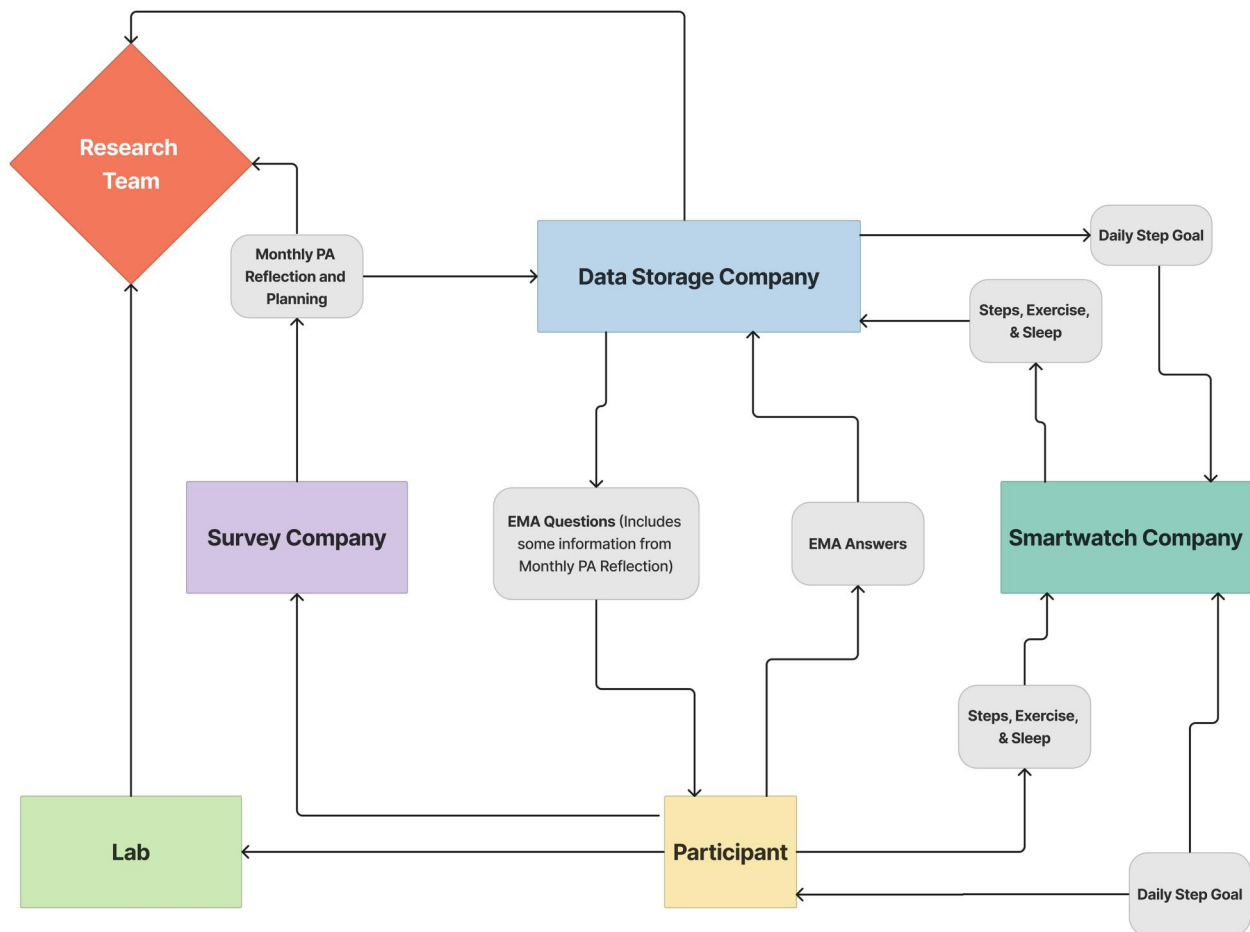
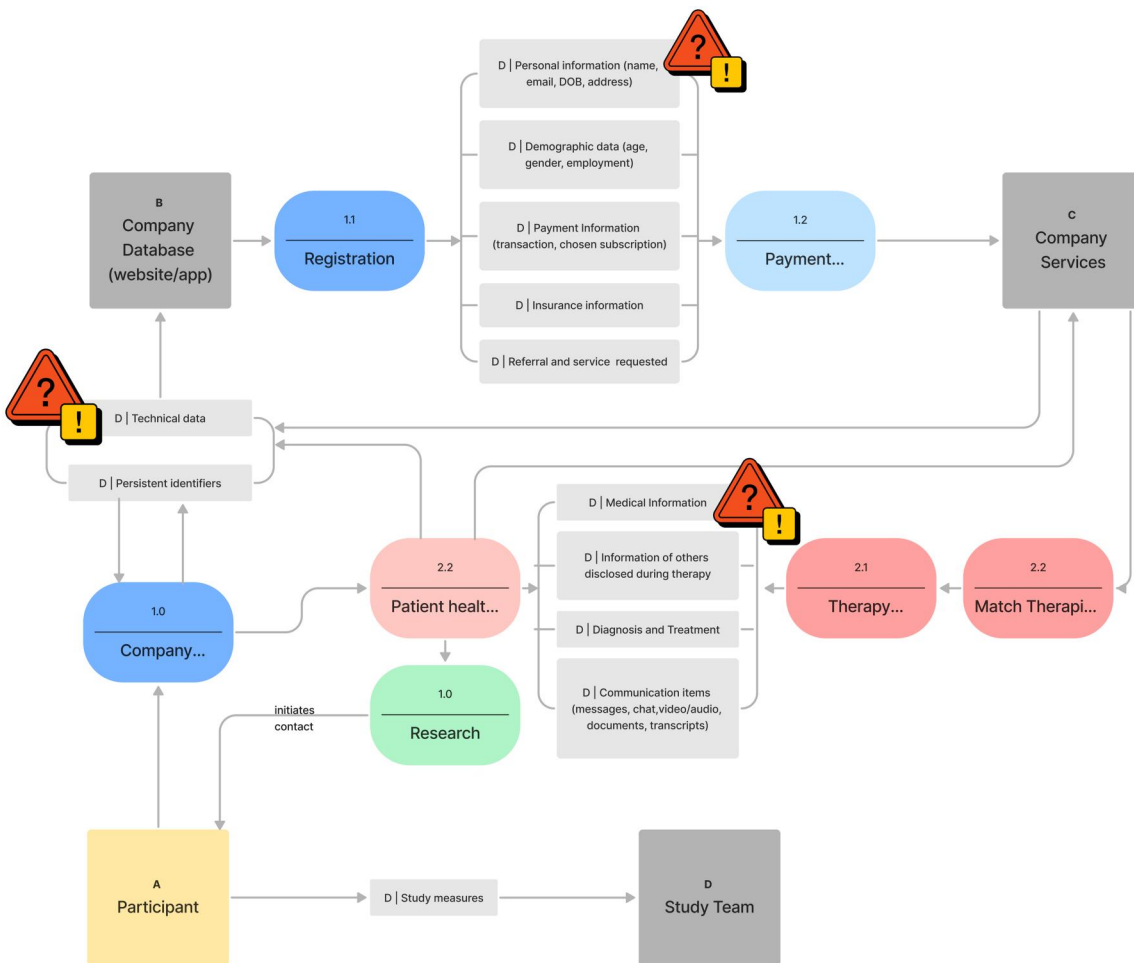


Figure 2. Baseline design for the Academic DFD. The Academic DFD used for the focus groups was based on the IRB-approved informed consent document to be shared with prospective participants as part of their consent process. A legend was not included in the original design of the Academic DFD.

- 1) *Text-based materials*: Participants learn about their assigned study context by reviewing a privacy policy or informed consent form. This approach mirrors standard approaches to reviewing data management practices with text-based materials.
- 2) *Incorporating a DFD*: Participants were introduced to a DFD depicting key aspects of the data management policy reviewed in Part 1. Participants can simultaneously review the DFD, as well as the materials presented in Part 1.
- 3) *Alternate ways of highlighting risk*: Participants were asked to consider modifications to the DFD to present risks associated with the data management policy, both with design strategy and level of detail. The research team applied visual strategies based on fundamental design principles through the use of color, line thickness,

icons, and text to visually convey risk.⁸ The baseline DFDs were modified for each study context to highlight specific points, pathways, and areas of risk and at different levels (ie, minimal, maximal).

- a) *Design strategy*. Risk was presented visually on each diagram as points, pathways, regions, and levels of risk. For instance, the point at which global positioning system (GPS) data are transferred from a smartwatch to a cloud-based service for analysis might be depicted as a *point* of risk. The arrows presenting the process that the digital health system takes to translate GPS data into just-in-time adaptive notifications to the user can be presented as a data *pathway*. If this data pathway is part of a cluster of methods that transfer multiple types of data to several cloud services simultaneously, then the cluster might be



A - Participant: any individual visiting the Company website or interacting with the app is a potential participant
 B - Company database: repository of all data Company collects from each site visit, app activity, communication, and/or use of its services
 C - Company services: includes all use of data collected by Company (provide, develop, and improve services, compliance with legal requirements, communication and marketing, security, processing insurance and payment activities, and conducting research)
 D - Study team: includes Company research partners
 1.0; 1.1; 1.2 - Processes involving mostly unprotected data collection
 2.0; 2.1; 2.2 - Processes involving HIPAA protected data collection
 3.0 - Processes involving data collection covered under the Common Rule

Figure 3. Points of risk. Caution signs are added to the DFD to identify three points in the commercial talk therapy system for prospective clients to consider, ie, personal information gleaned through the registration process, medical information provided by the patient healthcare provider, system logged technical data captured by the company.

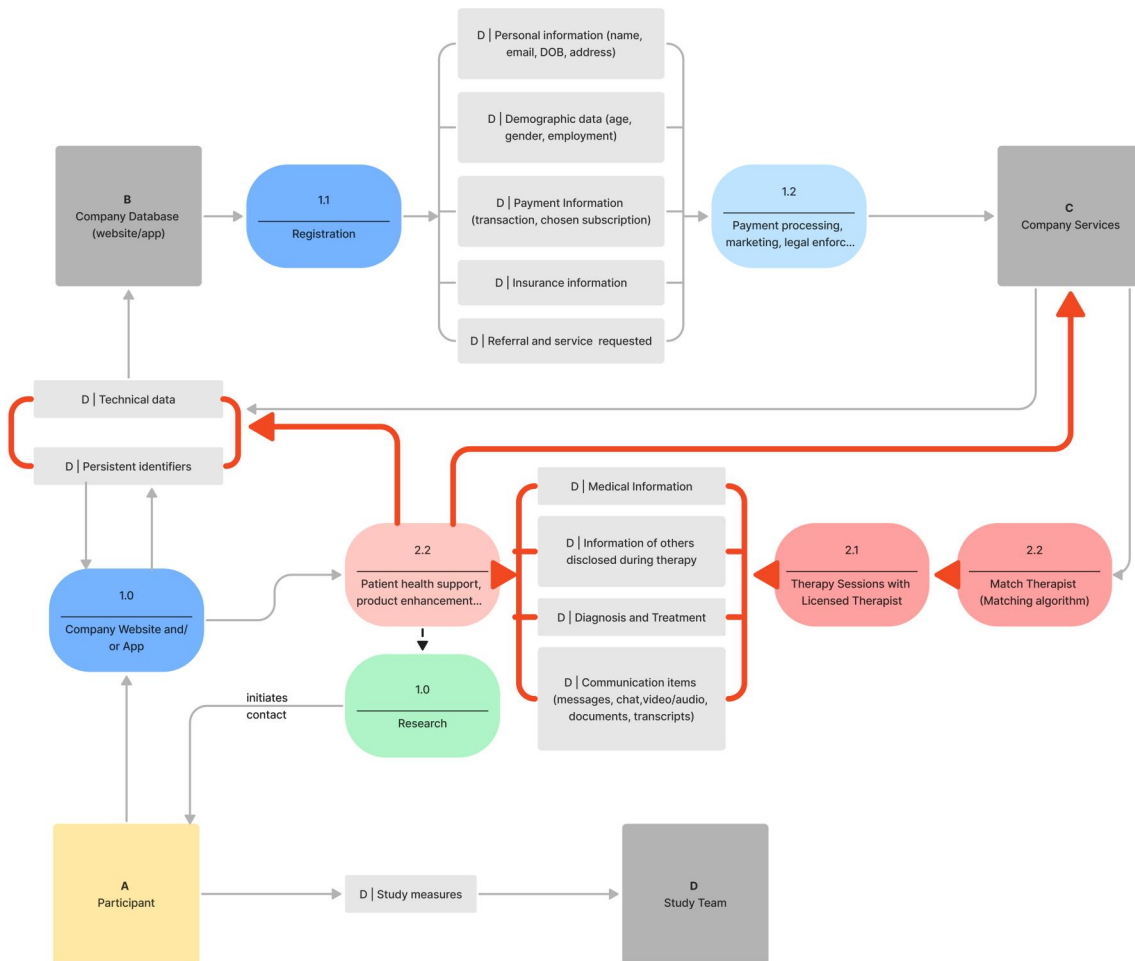
represented as a *region* of risk on a data flow diagram. To highlight differences in the degree of risk among multiple types (eg, points, pathways, regions), the research team applied visual strategies based on fundamental design principles through the use of color, line thickness, icons, and text to visually convey risk.⁸

b) *Level of detail.* During this part, the facilitator introduces each design strategy for highlighting risk, initially with just a few details added to the DFD (called the “minimal” detail view), and then with more details added to the diagram (called the “maximal” detail view). The facilitator was able to alternate the visibility of each view during the focus group, to facilitate discussion among the participants and to highlight key differences.

Figures 3-6 present examples of each strategy for highlighting study-related risks.

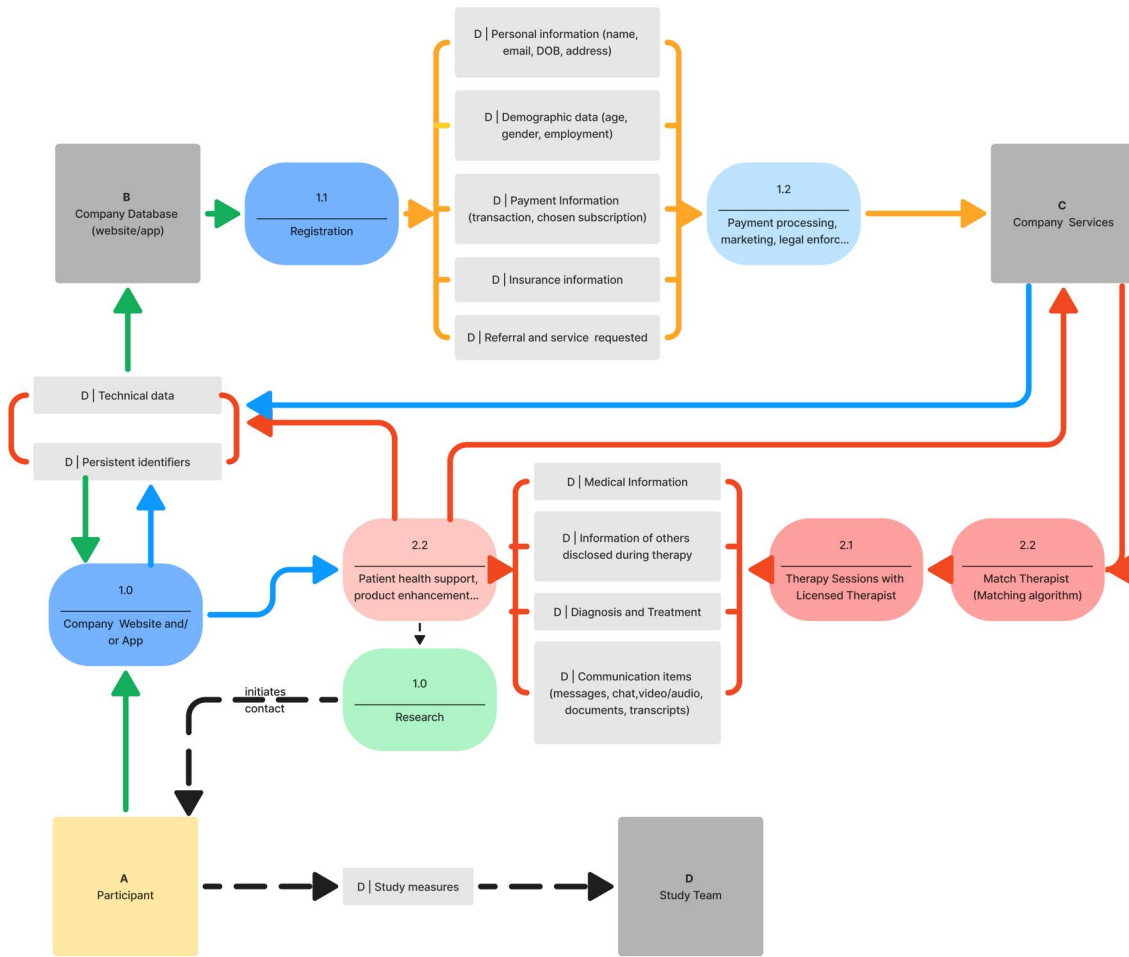
The modified DFDs presented in Part 3 provided participants with an opportunity to reflect on a variety of design considerations, such as ease of navigating the DFD, methods of directing attention toward specific risks, as well as how information about third-party use of data is represented along with other data processes. Participants were asked to provide feedback on each of the modified DFDs.

4) *Authorship considerations:* To explore how the provenance of a DFD plays into feelings of trust about a study context, participants were asked to consider several authorship scenarios. Specifically, participants were invited to share how information about the authorship of the DFDs played into their feelings of trust about



- A - Participant: any individual visiting the Company website or interacting with the app is a potential participant
- B - Company database: repository of all data Company collects from each site visit, app activity, communication, and/or use of its services
- C - Company services: includes all use of data collected by Company (provide, develop, and improve services, compliance with legal requirements, communication and marketing, security, processing insurance and payment activities, and conducting research)
- D - Study team: includes Company research partners
- 1.0; 1.1; 1.2 - Processes involving mostly unprotected data collection
- 2.0; 2.1; 2.2 - Processes involving HIPAA protected data collection
- 3.0 - Processes involving data collection covered under the Common Rule

Figure 4. Pathways of risk—the arrows connecting participant health data to system logged technical data collected by the smartwatch company are emphasized by changing the line color to red and increasing the line width. The arrows show how the participant health data, including therapy session transcripts and healthcare provider information, are integrated with system logged data to inform decision-making related to each client.



A - Participant: any individual visiting the Company website or interacting with the app is a potential participant
 B - Company database: repository of all data Company collects from each site visit, app activity, communication, and/or use of its services
 C - Company services: includes all use of data collected by Company (provide, develop, and improve services, compliance with legal requirements, communication and marketing, security, processing insurance and payment activities, and conducting research)
 D - Study team: includes Company research partners
 1.0; 1.1; 1.2 - Processes involving mostly unprotected data collection
 2.0; 2.1; 2.2 - Processes involving HIPAA protected data collection
 3.0 - Processes involving data collection covered under the Common Rule

Figure 5. Regions of risk—to present the “maximal” view of risk related to participant information collected by the commercial talk therapy platform, multiple data points in the DFD are shaded purple to highlight all parts of the system that collect personal data.

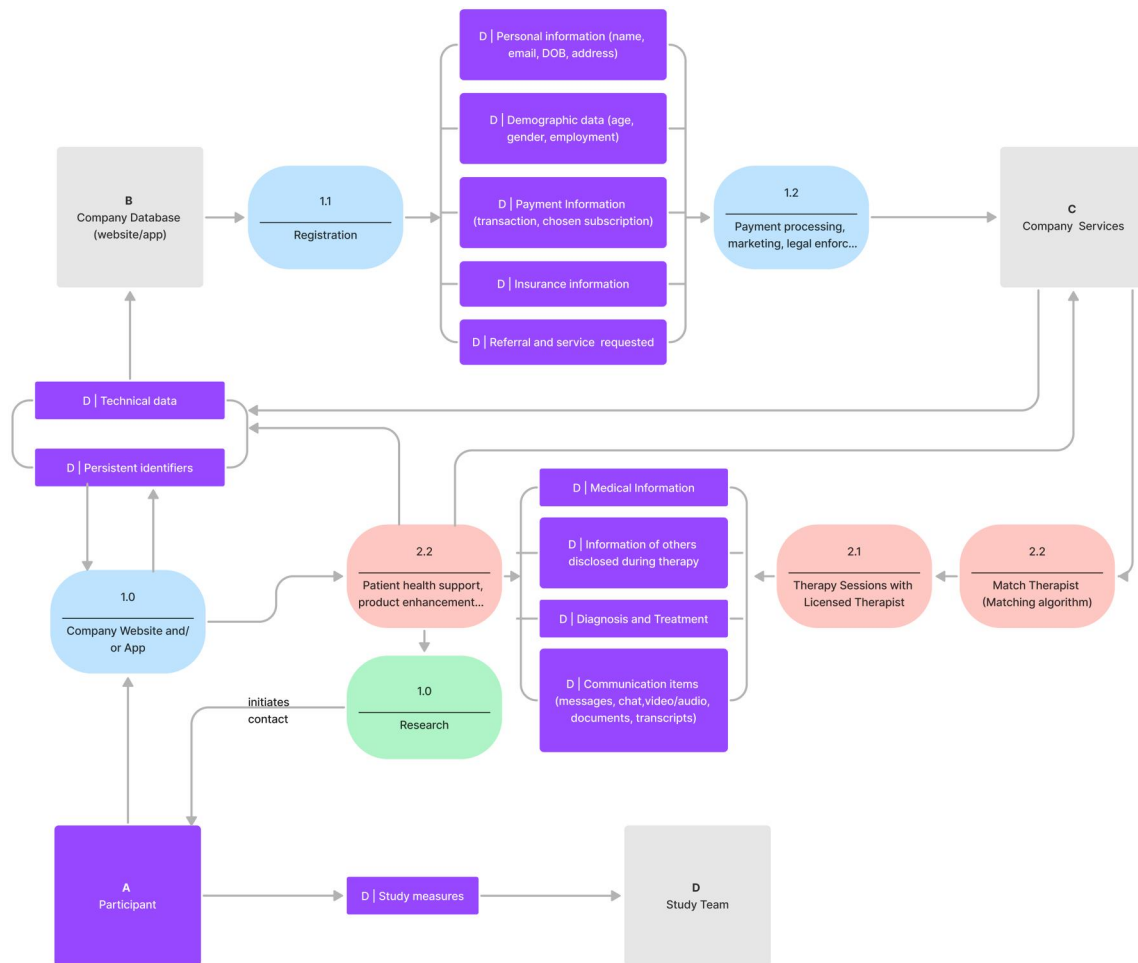
their assigned study context and how (if at all) the DFDs played into their willingness to participate in the study contexts.

- a) *Researchers conducting the study:* “Imagine that members of the research team leading the study constructed a DFD based on their own understanding of the technology and based on the way they intend to use the technology in the study.”
- b) *An institutional review board:* “An Institutional Review Board (IRB) is an administrative panel whose job is to review proposed research studies with the purpose of protecting the rights and wellbeing of research study participants. Imagine that members of an IRB spent time reviewing the study materials and prepared the DFD to reflect their best judgment of the risks.”
- c) *A diverse panel of peers:* “Imagine that the DFD was created by a diverse panel of peers. This may mean

that members of the community were involved in creating the DFD like the way Wikipedia articles are created and reviewed. The panel would have used polling or voting tools to rate the level of risk associated with each aspect of the DFD. The polling might use a series of sliding scales to indicate areas of high risk (red) and low risk (green).”

- d) *A generative artificial intelligence:* “Imagine that an artificial intelligence platform, like ChatGPT, was used to automatically construct the DFD. The AI might base the diagram on a variety of information, such as the data management and privacy policies, code repositories, as well as online reviews and posts about the technologies used in the study.”

During the workshops, the participants received no additional training about how to engage with the DFDs beyond the clarifying questions they asked the facilitator.



- A - Participant: any individual visiting the Company website or interacting with the app is a potential participant
- B - Company database: repository of all data Company collects from each site visit, app activity, communication, and/or use of its services
- C - Company services: includes all use of data collected by Company (provide, develop, and improve services, compliance with legal requirements, communication and marketing, security, processing insurance and payment activities, and conducting research)
- D - Study team: includes Company research partners
- 1.0; 1.1; 1.2 - Processes involving mostly unprotected data collection
- 2.0; 2.1; 2.2 - Processes involving HIPAA protected data collection
- 3.0 - Processes involving data collection covered under the Common Rule

Figure 6. Maximal pathways of risk—multiple colors are used to highlight each of the data pathways through the flow diagram. These include pathways to highlight the user registration process, processes to merge health records with system logged technical data, as well as processes for sharing company data for research purposes with academic collaborators, which is depicted with a dashed line to indicate that the process is not automatic. Unlike the minimal view of pathways presented in Figure 4, this maximal view highlights the multiple stories associated with the data collection, management, and sharing processes.

Data analysis

Focus group recordings were transcribed with an automated system. The transcription service generated statements for each participant, which are the primary unit of analysis for the study. Participant IDs were created for all participant statements. Participant IDs were also created to uniquely identify each participant to all statements associated with their participation in the focus group and their survey responses. Identifiable information within the text was removed from the transcripts, this included when participants referred to other participants by name. Three research team members (DK, RP, and JC) read through the transcripts and removed all personally identifiable information, while

simultaneously coding each statement with a Participant ID associated with each part of the focus group protocol [eg, (1) Text-based materials, (2) Alternate ways of highlighting risk] and the type of participant feedback (ie, Question, Opinion, Idea).

The inductive process of generating themes from the data involved iteratively reviewing the Participant ID associated with each part of the focus group protocol to identify the nature of the questions, opinions, and ideas expressed by the participants. The thematic analysis was led by JC and deliberated amongst the research team. A codebook was developed describing each of the emergent themes. Group consensus was used to resolve any differences of opinion about the themes.

Results

Forty-five people completed the prescreening survey and 34 (76% of 45 prescreened) were enrolled. Four participants were engaged in a pilot workshop and the remaining 30 participants were divided across eight focus groups reported herein. Each focus group involved between 2 and 5 participants. The average length of time for each focus group was 2 h. Of the 34 enrolled participants, 79% ($n=27$) of participants spoke English as their first language (Table 1). Self-identifying females represented 50% of the participants, both White and Black or African American races accounted for 41% ($n=14$ each) of the participants respectively, and 27% ($n=9$) self-identified as Hispanic or Latino (Table 1).

Considerations when adding a DFD to a consent process

Including DFDs in an informed consent process

Our analysis found that the DFD could add value as a supplement to traditional informed consent communications. Some participants said that they appreciated the high-level view of the study depicted by the DFD: “your diagram helps, because it visualizes it [and provides] a clear overview, so that [people] can grasp the main ideas pretty quickly [...] instead of going through all the text and small print” (P7). For a few participants, the DFD surfaced key details that helped them to feel more informed about a study.

“It does clear things up a little bit more about how the information is interchanged, who we're giving the information to, and how the information is back[ed up], and how the daily goals, for example, are created [...] it's a very, very, very, really small part of I feel what I need to understand in order for me to make an informed decision to participate in the research” (P33).

Participants who felt positive about the DFD regarded it as a resource, but not a replacement for a standard consent form. “I think it complements the consent form, but it definitely does not replace” (P32). The value of a DFD depends on whether it can surface information that is otherwise difficult to infer from a standard consent form and that prospective participants find useful in their evaluation of a study. Otherwise, a DFD may be a distraction.

Several participants found the DFD confusing. “Without prior knowledge about the consent form [presented in Part 1 of the focus group protocol], I'd be totally confused looking at this. It's too busy and then too complicated as well” (P29) and “I think this flowchart is confusing [...] I think there's so much information here” (P24). If people feel confused about study details, they may be less likely to participate.

For a few participants, feelings of confusion played into their feelings of trust in the information. “If, you know, there were things that I thought about that weren't listed on this diagram, then I would be having second thoughts about it before proceeding” (P10) and “if it's easy to read, I would find it more trustworthy, [but] if everything is muddled and sealed and hard to read, and basically ugly, I would find it not trustworthy” (P11). These remarks from participants highlight the risks associated with presenting a DFD that is

Table 1. Demographics of recruited participants.^a

Demographic	Total	Percent of total
Total	34	100%
Age		
21-29 years	12	35.3%
30-39 years	9	26.5%
40-49 years	7	20.6%
50-59 years	4	11.8%
60 years or older	2	5.9%
Education		
Some college but no degree	5	14.7%
Associate's degree	8	23.5%
Bachelor's degree	9	26.5%
Graduate degree	11	32.4%
Prefer not to answer	1	2.9%
Race		
Black or African American	14	41.2%
White	14	41.2%
Asian	3	8.8%
American Indian or Alaskan Native White	1	2.9%
White, Other (specify)	1	2.9%
Prefer not to answer	1	2.9%
Hispanic/Latino		
No	24	70.6%
Yes	9	26.5%
Prefer not to answer	1	2.9%

^a Demographic—demographic characteristics of the participants; Total—number of participants; Percent of Total—percent of total, where total number of participants = 100%; Age—age ranges of participants; Education—level of education attained by participants; Race—self-identified race of participants; Hispanic/Latino—self-identified Hispanic/Latino participants.

not useful, whether due to the information presented or the presentation.

A useful DFD can serve as a guide, pointing people toward relevant information. “[If] that diagram would raise flags or details, or [if] there's something that I would be concerned about, I would definitely go to the specific paragraph and get more information to make a really informed decision” (P7). The DFD for a study could also be shared on request, “maybe it just exists as an asset, like if a participant asks a question regarding data flow and you have this [diagram] to explain it in an easy way” (P33). In these ways, a research team might choose when and how to introduce a DFD during an informed consent conversation.

Esthetics of the diagram: clear, simple, and logical

Across all focus groups, most participants stated that the DFD should be clear and simple to follow. For instance, several participants said that curved lines and looping arrows made the DFD more difficult to understand. “I do think like, when stuff goes back, there is a lot of looping that happens visually, so try to map out one specific [path]” (P30) and “It bugs me. Like, I like straight lines. [Otherwise,] it takes my eye away from what I'm supposed to be looking at, because I'm stuck on the curve” (P25). Such simple aesthetic decisions can promote clarity.

In general, participants expressed an appreciation for simple DFDs that present only enough information as necessary to cover the main points in the consent materials.

“It just has to be clearer or not too fancy. Like have it more just like bullet points or one color or two colors max because I know participants can get kind of overwhelmed with all the verbiage and the words and the lines and what do I have to do there’s caution signs everywhere, it just gonna confuse the participant” (P23). Several participants suggested that a legend would help people to interpret the colors, icons, and other markings on the DFD.

While a DFD can depict the processes associated with a study, our analysis found that which aspects of a DFD to consider first, can be less clear for participants. “Where does it start? This is a study, there’s got to be a point before the study and a point at the end of the study” (P26). Rather than centering the DFD on the data processes, participants recommended focusing the design on conveying the experience of people involved in the study. “[C]entering the participants in the middle would be an appropriate thing to do” (P32).

Many participants recommended that the research team should assume that people do not have any prior experience with the data-sharing processes, let alone the data literacy skills necessary to interrogate a DFD. “Like, if you don’t have any knowledge about data analysis and data flow, are you just going to be like, I don’t even like this, just a bunch of arrows [...] I just don’t think that everybody would understand it” (P33). A related concern is the general accessibility of DFDs, for instance designers might apply a variety of techniques to support people with visual impairments: “I also think about folks who may experience colorblindness [...] so I’ve used like shading or different like squiggles or polka dots, if colors aren’t visually able to be registered?” (P30). Such design decisions can promote understanding, but also help to demonstrate the study team’s care for prospective participants.

Our analysis found that simplicity in design is important and some interactive features in a digital presentation of consent materials can be used to hide, reveal, and highlight key details in a DFD. For instance, a participant shared the following hypothetical user interaction with a DFD:

“When I look at this diagram, obviously it’s a lot more simplified, but as you go through the diagram are you able to like *click* on it, and then it takes you straight into that the end [where] it goes into depth about the information you’re looking at [in the diagram]” (P10).

This user interaction recommendation speaks to a broader question about how best to use both standard text-based consent materials along with novel information visualizations and interactive digital experiences. While these designs have the potential to provide people with multiple entry-points into the material, our analysis also raised potential points of confusion.

Considerations when highlighting study data-related risks

As part of the study protocol, participants were introduced to various ways of highlighting specific risks in the DFDs. Regarding the design of highlighting risk, our analysis found participants wanted to understand the riskiness levels. “There are risks involved, [but] what is risky is necessarily subjective” (P32). For instance, a participant discussed ways of representing different types of risk in the study—from their perspective—through a DFD.

“How do you differentiate the risk of bruising a finger from getting diabetes versus dying of a heart attack on the treadmill? It’s a very different level of risks that cannot be easily put into a color code or pictogram or symbol or something” (P22).

However, the design of risk in a DFD can be confusing, “I would be really confused, and I’d probably have so many questions, and then I wouldn’t have that much confidence in, you know, proceeding” (P10).

In addition to these general design recommendations, participants shared recommendations for communicating the risks associated with specific types of informed consent content. Specifically, participants raised an interest in the possible health risks associated with a study, “If I hadn’t read the wording of the consent, I would kind of wonder about like *lab measures* and what that entails, just because it’s very general there” (P21). However, participants who did not view the “lab measures” as a particular risk wondered whether it was necessary to highlight at all. “I’m a bit unclear on why you guys would be so focused on highlighting these risks. I mean, is that something that just needs to be done for a reason? Do we feel like these are higher risks than in a normal study?” (P21). Our analysis highlights the challenges a research team might face in trying to accommodate the various levels of interest and tolerance for study risk among prospective participants.

However, participants felt that highlighting risks helped to call attention to potential concerns that were less apparent in the standard text-based consent materials. Several participants commented on the risks associated with third-party access to study data. “This diagram makes me realize that my data will be stored in four different places, which I did not understand that at all from the consent form” (P34). Some of the third-party services were familiar to the participants. “The [Smart Watch Company] is probably a bigger risk, because it has this information, but I trust [Smart Watch Company]. I don’t know the [Data Storage Company], and don’t even know what [Survey Company] is so this kind of makes me think, what do they have” (P24). Participants are less willing to share data when they are less familiar with the third-party services, “I haven’t heard of [Mental Health Provider] before today, so I would be less inclined to provide any information to them” (P8). In these ways, a DFD can raise new questions for prospective participants as well as the research team to consider.

Participants highlighted how the third-party services might use digital health study data is less clear based on the DFD. “They’re collecting all this information, [but] what are they going to do with it? Am I going to get more spam?” (P16). A few participants called attention to the lack of information in the standard consent form about concerns related to third-party services. “To note the informed consent didn’t once mention [Survey Company] It’s letting people know like this is what [Survey Company] is, it is a survey, it’s a database where your surveys will be conducted out of [...] that’s where that assessment is going to live” (P30). These gaps in the informed consent materials are highlighted by the DFDs.

Participants shared that there are several “sensitive data types” that they would want highlighted throughout the informed consent process, if a study incorporated these types.

“I would be concerned about medical information, especially if there's just certain things that I don't want to, you know, be disclosed from my primary care doctors or, you know, to interfere like between the therapist and, you know, my own personal medical care that I get outside of that therapy” (P10).

Participants also want to understand the potential risk associated with a data breach. “In that case basically all your data, including payment information, insurance information, probably also, confidential or non-confidential therapy sessions, personal data, everything [is lost]. Given that there seems to be no separation between all data [...] I think [this] is a big red flag” (P7). DFDs can help highlight risk and opportunities for improved data protection.

Considerations for creating and validating flow diagrams that depict data policies

As a final stage of our study, participants were asked to consider how DFDs for a study might be authored, whether by a research team, institutional review board, panel of peers, or by a LLM like ChatGPT.

Asking researchers to generate DFDs for their own study is not necessarily ideal. Several participants felt that researchers know the most about their own studies, so are well suited to creating DFDs for their work. “Seems like they [the research team] would know what's going on the most, they would have all the details of the study” (P21). However, a few participants felt that the research team might be biased in their representation of the risks in their own study, “they have the most interest in the project. Not that they would intentionally do anything, but just because they're the most embedded in it, they should probably not [design the DFD]” (P11).

Some participants felt strongly that researchers are not the right people to communicate this information about their research:

“[Instead] of relying on the research team members, where I don't feel as confident. I know that they know so much about what they're presenting, and they fully understand it, but for me I would feel more secure if I was to talk to someone that didn't have like a conflict of interest, or anything like that” (P10).

Another participant felt that some research teams might even conceal critical information in a DFD. As a participant shared, “you [researchers] have an investment in up playing the benefits and downplaying the risks” (P32). These comments speak to the distrust that some people have with the practice of scientific research.

By contrast, participants felt that the priorities for the IRB are different than those of a research team, as “their priorities are [to] make sure it's all legal, make sure they follow all the rules that have been established, so that it's easy for [people] to read, make sure it lists all the possible negative outcomes [...] so that the university is covered” (P12). However, the IRB is not typically involved in creating consent materials but, are involved in reviewing documents submitted by the researcher. Reviewing a DFD would be an added responsibility but, one that aligns with current practice.

An option that several participants preferred is to recruit diverse panels of peers to generate DFDs for studies, but

people would need to have information about the panel demographics, interests, and relevant backgrounds.

“I kind of like the diverse panels because [...] you're not bias of people actually doing the study itself” (P12) and “when you come to a diverse panel of peers you're doing this for the people” (P18).

Other participants were less optimistic about this option. “The panel of peers is like a crapshoot; you have no idea what you're gonna get there, so I just couldn't trust it that much” (P26).

Of all the options, participants were the least enthusiastic about using an artificial intelligence, like ChatGPT, to author the DFDs. Artificial intelligence might result in many variations of a DFD, which could cause confusion. However, having multiple options generated could be useful for researchers to evaluate. “ChatGPT could give you an immediate solution, but it will be one of the solutions in 1000s” (P18).

Our analysis found that there are trade-offs associated with each option, whether asking people to develop a DFD or an artificial intelligence to author the work. However, several participants shared that these options could be coordinated as part of a process for creating, enhancing, and reviewing DFDs for a study. Some participants felt that a ChatGPT could be used to generate multiple options for a diverse panel or IRB to consider, “I think it could be a very helpful supplemental tool, but I don't think that it by itself could be completely trustworthy” (P34). In other words, humans would need to be integrated into the process of evaluating any DFDs generated by an artificial intelligence.

Discussion

Our study confirmed that information visualizations can add value to the informed consent process in digital health research if used to augment the standard text-based approach to convey study related information; however, information visualizations are not a replacement for text-based study information. Design considerations for communicating study risks as well as the process of evaluating study risks to prospective research participants include using concise, limited, and clear warnings to emphasize points, pathways, regions, and various levels of risk (ie, minimal, maximal). Information visualizations ideally can help people to find their way into the details of a study, which in practice might involve a “doubly-linked” user interaction to guide people back-and-forth between the text-based materials and components of the visualization.⁹ Our study also highlights how the processes involved with authoring an information visualization also matters to prospective study participants. People need information about how the research team and any other groups or resources were involved in the process of creating and evaluating any information visualizations included in a digital health consent process, whether in academic research or a commercial context. These details help people to evaluate the potential for biases and possible concerns related to accuracy.¹⁰

Communicating with prospective research participants about the data they voluntarily contribute to a digital health study is not easy. From considerations about data collection to data processing, management, and sharing, communicating the details about study data can require a fair amount of time and effort to meet the various literacy needs of

prospective participants.¹¹ In the context of digital health, data may include a variety of personally sensitive information, such as participant heart rate, geographic location, known medications, levels of anxiety, and transcripts from virtual sessions with a personal therapist.⁴ To promote public involvement in science and demonstrate respect for research participants, this study investigated methods of communicating about data processes and related risks by using information visualization tools.^{12,13} Specifically, data flow diagrams were used to present prospective research participants with possible data management risks that were represented as points, pathways, regions, and levels of risk on the diagrams (ie, minimal, maximal).

Participants remarked on how reviewing the data visuals helped them to realize questions about data management and sharing procedures, which they did not recognize when just reviewing the text-based information. While a few participants felt that the data visuals emphasis on risk was not necessary given their perceived low risk of the digital health study, other participants valued how the highlighted features directed their attention, whether that was a bold line, a red flag, or a highlighted region of the diagram.¹² Researchers could use the differing opinions among prospective participants as an opportunity to promote conversation about how people perceive study risks and what steps they would want researchers to take, if the risks were to increase suddenly. For instance, a research coordinator could ask a prospective participant to reflect on the data management risks emphasized in a data flow diagram: “What questions to you have about how data are collected, stored or shared? What do you think might concern other participants, who are like you?” Thereby, inviting them to share their opinion and any possible questions they might have about the warnings presented on the visualization.

Warning signs need to be clear, concise and easy to interpret, because they reflect how the research team demonstrates respect toward study participants. Fundamentally, providing additional information about the data risks in a study adds time and effort to the prospective participants’ experience of the informed consent process.^{9,14,15} Reviewing this information must be valuable from their standpoint.¹⁶ Our analysis surfaced several recommendations about how to do just that: Participants asked for straight lines, a limited range of colors, concise captions, and interactions that direct people toward additional details about a particular warning. The feedback presented in the results offer principles for representing risky points, pathways, regions, and levels on a visualization, but the method of inviting prospective participants to participate in the critique of consent materials is also a useful step before launching a study. Community-based design is not common in the development of consent materials, but our analysis demonstrates the value that it can offer.^{11,17} Involving people in the early stages of study development can generate useful ideas for communicating with prospective participants later and cultivate trust.

Our analysis also found that people need information about how study risks are evaluated and by whom (or what). When asked to consider the possibility that DFDs and warning signs were generated by the researchers, an IRB, a panel of peers, or a LLM, participants wanted researchers to generate the DFD and use the LLM to support formatting and readability. Participants shared various concerns about all the options ranging from worry about researcher bias to not

being able to trust the information provided by the LLM as factual and relevant. While participants raised numerous concerns, our analysis found some consensus that people would appreciate a multi-phased approach to constructing and validating the data visuals presented in a consent process. Such an approach might involve the following steps:

- 1) Access an LLM to highlight key data management processes for a specific study technology.
- 2) Poll the research team and a panel of peers to evaluate and enhance the LLM drafted materials.
- 3) Submit the data visuals to the IRB for review along with the other study communication materials.

In practice, researchers need ways of communicating with prospective participants (and other interested parties) the details of such a multi-faceted study risk evaluation process.¹⁰ In the future, this could be in the form of a simple visual badge appearing alongside an information visualization, like the signifiers of organic and gluten-free foods at the grocery store. A critical limitation of this study is that most of our participants report having at least some college experience and many have achieved graduate-level degrees, which implies that they have had access to training and resources that many other prospective participants in digital health research have not. Future research should specifically investigate strategies for highlighting study risk that promote data literacy and access for diverse populations.

As new technologies, like AI and augmented reality, are integrated into digital health systems, researchers and the public need better ways of understanding the risks associated with such integrations.^{5,18,19} This study raised several considerations about how to communicate with people about the data collection, processing, and sharing considerations associated with a research study.^{13–15,20} While our participants shared ideas about how this might happen, facilitating the type of communication modeled by our study is not practical for most digital health studies. To translate these insights about presenting study data risks into practice, funding agencies might consider taking steps to: (1) promote data literacy guidelines for developing informed consent materials, (2) develop technology to help researchers and IRBs navigate the emerging policy landscape associated with digital health technologies, and (3) lead the creation of infrastructure and support systems through funding agencies to help researchers collaborate around science communication. These investments of time, influence, and money can help researchers and the public to collaborate around the emergent risks related to digital health systems.^{11,16,17,21}

Conclusion

Dataflow diagrams were identified as being useful tools to convey complex data-sharing processes and potential privacy risks. Therefore, DFDs may be useful as a supplement to the traditional informed consent process, if designed well, and provide participants with additional information related to the data-sharing process. Future research should explore designing DFDs to be more participant-centered and should consider incorporating the study activities into the design. Future studies should include quantitative assessments of acceptability, accessibility, understanding, and risk

identification and gathering more qualitative information from participants with more diverse backgrounds.

Acknowledgments

We acknowledge the support from Michael Higgins who assisted with participant recruitment for this study.

Author contributions

Brian J. McInnis, PhD (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision, Visualization), Ramona Pindus, BS (Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization), Daniah H. Kareem, BSPH (Conceptualization, Formal analysis, Investigation, Validation, Visualization), Julie Cakici, PhD, RN (Formal analysis, Validation), Daniela G. Vital, MPH, BSPH (Project administration, Supervision), Eric Hekler, PhD (Conceptualization, Funding acquisition, Methodology), and Camille Nebeker, EdD, MS (Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision)

Funding

This work was supported by the National Cancer Institute of the National Institutes of Health under award number R01CA244777, Patient-Centered Outcomes Research Institute (PCORI) Award (ME-2020C3-21310) and the National Science Foundation [2124975]. The research presented in this paper is solely the responsibility of the author(s) and does not necessarily represent the views of the sponsors.

Conflicts of interest

None declared.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1. Nebeker C. mHealth research applied to regulated and unregulated behavioral health sciences. *J Law Med Ethics*. 2020;48:49-59. <https://doi.org/10.1177/1073110520917029>
2. Hekler E, Tiro JA, Hunter CM, et al. Precision health: the role of the social and behavioral sciences in advancing the vision. *Ann Behav Med*. 2020;54:805-826. <https://doi.org/10.1093/abm/kaaa018>
3. Rothstein MA. Big data, surveillance capitalism, and precision medicine: challenges for privacy. *J Law Med Ethics*. 2021;49:666-676. <https://doi.org/10.1017/jme.2021.91>
4. Chikwetu L, Miao Y, Woldetensae MK, et al. Does deidentification of data from wearable devices give us a false sense of security? A systematic review. *Lancet Digit Health*. 2023;5:e239-47-e247. [https://doi.org/10.1016/S2589-7500\(22\)00234-5](https://doi.org/10.1016/S2589-7500(22)00234-5)
5. Sero D, Zaidi A, Li J, et al. Facial recognition from DNA using face-to-DNA classifiers. *Nat Commun*. 2019;10:2557. <https://doi.org/10.1038/s41467-019-10617-y>
6. Crossfield SSR, Zucker K, Baxter P, et al. A data flow process for confidential data and its application in a health research project. *PLoS One*. 2022;17:e0262609. <https://doi.org/10.1371/journal.pone.0262609>
7. Quisel T, Foschini L, Signorini A, et al. Collecting and analyzing millions of mHealth data streams. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery 2017:1971-1980.
8. Murchie KJ, Diomedes D. Fundamentals of graphic design—essential tools for effective visual science communication. *FACETS*. 2020;5:409-422. <https://doi.org/10.1139/facets-2018-0049>
9. Jun GT, Ward J, Morris Z, et al. Health care process modelling: which method when? *Int J Qual Health Care*. 2009;21:214-224. <https://doi.org/10.1093/intqhc/mzp016>
10. Appenzeller A, Hornung M, Kadow T, et al. Sovereign digital consent through privacy impact quantification and dynamic consent. *Technologies*. 2022;10:35. <https://doi.org/10.3390/technologies10010035>
11. Nyirenda D, Sariola S, Kingori P, et al. Structural coercion in the context of community engagement in global health research conducted in a low resource setting in Africa. *BMC Med Ethics*. 2020;21:90. <https://doi.org/10.1186/s12910-020-00530-1>
12. Ibrahim R, Yen S. Formalization of the data flow diagram rules for consistency check. *IJSEA*. 2010;1:95-111. <https://doi.org/10.5121/ijsea.2010.1406>
13. Abujarad F, Alfano S, Bright TJ, et al. Building an informed consent tool starting with the patient: the patient-centered virtual multimedia interactive informed consent (VIC). *AMIA Annu Symp Proc* 2018. 2017;2017:374-383.
14. Abujarad F, Peduzzi P, Mun S, et al. Comparing a multimedia digital informed consent tool with traditional paper-based methods: randomized controlled trial. *JMIR Form Res*. 2021;5:e20458. <https://doi.org/10.2196/20458>
15. Kraft SA, Constantine M, Magnus D, et al. A randomized study of multimedia informational aids for research on medical practices: implications for informed consent. *Clin Trials*. 2017;14:94-102. <https://doi.org/10.1177/1740774516669352>
16. O' Sullivan L, Feeney L, Crowley RK, et al. An evaluation of the process of informed consent: views from research participants and staff. *Trials*. 2021;22:544. <https://doi.org/10.1186/s13063-021-05493-1>
17. Spector-Bagdady K, De Vries RG, Gornick MG, et al. Encouraging participation and transparency in biobank research. *Health Aff (Millwood)*. 2018;37:1313-1320. <https://doi.org/10.1377/hlthaff.2018.0159>
18. Card AJ, Harrison H, Ward J, et al. Using prospective hazard analysis to assess an active shooter emergency operations plan. *J Healthc Risk Manag*. 2012;31:34-40. <https://doi.org/10.1002/jhrm.20095>
19. Knapp P, Martin-Kerry J, Moe-Byrne T, et al. The effectiveness and acceptability of multimedia information when recruiting children and young people to trials: the TRECA meta-analysis of SWATs. *Health Soc Care Deliv Res*. 2023;11:1-112. <https://doi.org/10.3310/HTPM3841>
20. Tait AR, Voepel-Lewis T, Levine R. Using digital multimedia to improve parents' and children's understanding of clinical trials. *Arch Dis Child*. 2015;100:589-593. <https://doi.org/10.1136/archdischild-2014-308021>
21. Nebeker C, Gholami M, Kareem D, et al. Applying a digital health checklist and readability tools to improve informed consent for digital health research. *Front Digit Health*. 2021;3:690901. <https://doi.org/10.3389/fgdth.2021.690901>