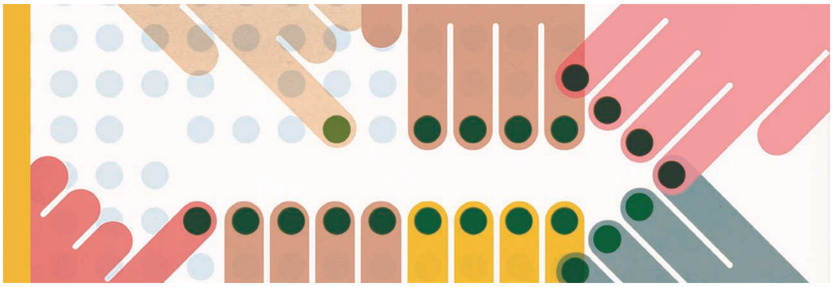
(Re)Using Crowdsourced Health Data: Perceptions of Data Contributors

Andy Alorwu and Aku Visuri, University of Oulu

Niels van Berkel, Aalborg University

Simo Hosio, University of Oulu

// Open data are often contributed by various governments and public sector actors. An increasingly popular way to collect large bespoke data sets is crowdsourcing. In this article, we explore crowdsourced open data as enablers of future software solutions. //



OPEN DATA HAVE been predominantly propelled by governments and public organizations, sharing meaningful data sets on which others can build.[1] As personal digital technologies have rapidly proliferated, most of us now produce a constant stream of data daily—data which can be used to build different types of novel software solutions and services.[2] A particularly interesting class of such data are health related, now collected pervasively by fitness trackers, wearable sensors, and increasingly, even our smartphones through built-in health-monitoring tools.[3] Such data are typically under the control of corporations, which provide the infrastructure, e.g., smartphones, fitness trackers, and social media software, through which such data are generated.[2] In this context, *open health data* (*OHD*) refer to any type of publicly accessible health-related data.[4,5]

Coined as the act of outsourcing a job to an undefined group of people in the form of an open call, crowdsourcing has become a primary means of collecting high-quality data from people at scale.[6] Crowd workers are people who perform crowdsourcing tasks for a fee on crowdsourcing platforms. Recently, researchers have begun to explore crowdsourcing as a tool to collect bespoke OHD as input for digital health software solutions. The perceptions of data donors are crucial to their data donation decision making but remain critically underexplored. An understanding of user perceptions is needed to help the software community take steps to alleviate data donor fears and concerns.

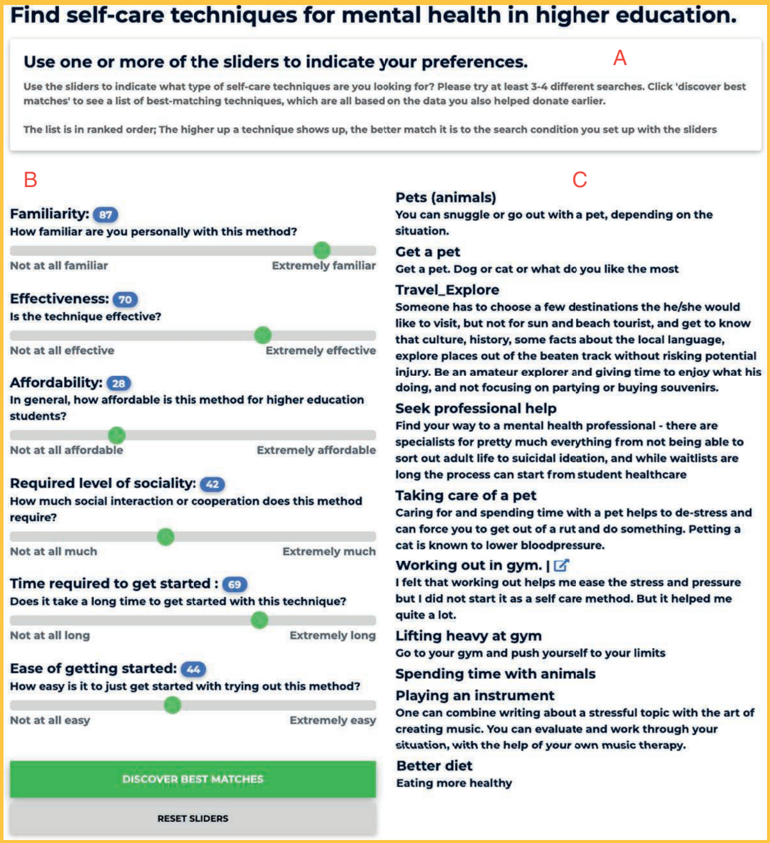
In this article, we present an online study where we invited crowd workers from a popular online crowdsourcing platform to first contribute data to a decision-support system on mental health self-care, and then to explore the system and take an online questionnaire about their perceptions of such OHD reuse in a follow-up study (see "How We Conducted the Study"). The following are our main contributions:

- We outline and detail concerns people experience in the context of OHD in software systems, including privacy concerns, potential for

Digital Object Identifier 10.1109/MS.2021.3117684
Date of current version: 23 December 2021

HOW WE CONDUCTED THE STUDY

The participants in our study had been invited earlier to contribute and assess self-care techniques for mental health using a public data collection and decision-support tool. At that point, participants were informed that all the data they provide will be used freely in research and as an open data set accessible for anyone online, that is, open health data (OHD). The decision-support tool[S1] can analyze such data and turn them into an interactive exploration interface, which is helpful in finding suitable self-care techniques, as seen in Figure S1. Our study was enabled by Prolific,[S2] a crowdsourcing marketplace used for many academic studies. We used a mixed-methods[S3] approach, which combines both qualitative and quantitative data collection and analysis, providing an opportunity to better understand the depth and breadth of the research problem than either a qualitative or quantitative method alone.



Find self-care techniques for mental health in higher education.

Use one or more of the sliders to indicate your preferences.

Use the sliders to indicate what type of self-care techniques are you looking for? Please try at least 3-4 different searches. Click 'discover best matches' to see a list of best-matching techniques, which are all based on the data you also helped donate earlier.

The list is in ranked order; The higher up a technique shows up, the better match it is to the search condition you set up with the sliders

Familiarity: 87
How familiar are you personally with this method?

Effectiveness: 70
Is the technique effective?

Affordability: 28
In general, how affordable is this method for higher education students?

Required level of sociality: 42
How much social interaction or cooperation does this method require?

Time required to get started : 69
Does it take a long time to get started with this technique?

Ease of getting started: 44
How easy is it to just get started with trying out this method?

Pets (animals)
You can snuggle or go out with a pet, depending on the situation.

Get a pet
Get a pet. Dog or cat or what do you like the most

Travel_Explore
Someone has to choose a few destinations the he/she would like to visit, but not for sun and beach tourist, and get to know that culture, history, some facts about the local language, explore places out of the beaten track without risking potential injury. Be an amateur explorer and giving time to enjoy what his doing, and not focusing on partying or buying souvenirs.

Seek professional help
Find your way to a mental health professional - there are specialists for pretty much everything from not being able to sort out adult life to suicidal ideation, and while waitlists are long the process can start from student healthcare

Taking care of a pet
Caring for and spending time with a pet helps to de-stress and can force you to get out of a rut and do something. Petting a cat is known to lower bloodpressure.

Working out in gym. |
I felt that working out helps me ease the stress and pressure but I did not start it as a self care method. But it helped me quite a lot.

Lifting heavy at gym
Go to your gym and push yourself to your limits

Spending time with animals

Playing an instrument
One can combine writing about a stressful topic with the art of creating music. You can evaluate and work through your situation, with the help of your own music therapy.

Better diet
Eating more healthy

FIGURE S1. The decision-support system's interface developed to suggest mental health self-care techniques to participants. (a) The general instructions, (b) criteria for finding self-care techniques, (c) and suggested self-care techniques based on the selected criteria.

The average completion time of the study was 17.34 min, and participants were compensated £3 per contribution, making the mean hourly wage well above typical crowd work standards. The participants' age ranged between 20 and 54 (mean = 26.26 years; standard deviation = 5.72 years; 52 males; 28 females). The percentage of participants who considered themselves knowledgeable about technology was 83.75% (N = 67). Further, 83.75% (N = 67) stated that they collect health-related data about themselves regularly.

We conducted a thematic analysis,[S4] following a deductive approach to uncover data-contributor perceptions specifically concerning the threats and opportunities of OHD. We analyzed the data with specific questions in mind and coded responses relevant to these questions. Our analysis approach followed a theoretical thematic analysis rather than an inductive one. Given this goal, each segment of data that was relevant to our questions was coded. We used open coding, developing and modifying codes as we worked through the coding process. Two of the authors generated the initial set of codes, simultaneously discussing the coding scheme. The codes were then shared with the other two authors.

(Continued)

HOW WE CONDUCTED THE STUDY (*CONTINUED*)

The authors conducted multiple online meetings to discuss and resolve disagreements with the coding.

Limitations

We acknowledge limitations in our work. Our results originate from Prolific with student participants and as such do not generalize over the entire population. However, it shows an indication of a broader trend as results from marketplaces such as Prolific are valuable to research and produce data with high validity.

References

S1. S. Hosio, J. Goncalves, T. Anagnostopoulos, and V. Kostakos, "Leveraging wisdom of the crowd for decision support," in *Proc. 30th Int. BCS Human Comput. Interaction Conf.*, 2016, pp. 1–12.

S2. S. Palan and C. Schitter, "Prolific.ac — A subject pool for online experiments," *J. Behav. Exp. Finance*, vol. 17, pp. 22–27, Mar. 2018. doi: 10.1016/j.jbef.2017.12.004.

S3. J. Wisdom and J. W. Creswell, "Mixed methods: Integrating quantitative and qualitative data collection and analysis while studying patient-centered medical home models," Agency for Healthcare Research and Quality, Rockville, MD, AHRQ Publication No. 13-0028 -EF, 2013.

S4. V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Res. Psychol.*, vol. 3, no. 2, pp. 77–101, 2006. doi: 10.1191/1478088706qp063oa.

misuse of data, and the accuracy of data.

- We identify a range of perceived threats and opportunities in using crowdsourced data to enable software solutions, outlining opportunities for future work.
- We highlight the importance of the involvement of public and societal stakeholders in software-development efforts, which rely on open data because they command public trust.

Data Collection

We invited back 80 of the participants who donated data for the decision-support system to participate in an online questionnaire study. All of the participants were fluent in English, students at a higher-education institution, and had completed at least 50 tasks on Prolific—a high-quality, online research subject pool and crowdsourcing platform widely used in both academia and industry. Thus, they were fairly experienced crowd workers.

The questionnaire was hosted in Google Forms and contained three distinct stages: background information, brief reactions about the public decision-support system itself (participants could explore the system at this stage through a hyperlink), and finally, an extensive set of questions about OHD reuse. Thus, the final stage was essentially stimulated by their experiences with the public tool, which was built using their own OHD.

Data Donation and Trust

Donating Data Toward Open Data Initiatives

We asked participants what considerations they deemed important when making a decision to donate data for public use, as they did in the first stage of our experiment. To this end, "anonymity" (the state of being unidentifiable) and "how the data will be used" were participants' two most-prominent considerations [87.5% (N = 70) and 88.75% (N = 71), respectively]. These were followed by "imaginable future benefit to myself," with 45% (N = 36), and finally, "perceived societal benefit," with 37.5% (N = 30).

Public, Private, and Societal Stakeholders

We investigated the level of trust people have in various stakeholder organizations building software using OHD [see Figure 1(c)]. To this end, public (governmental agencies and public research organizations) and societal stakeholders (nongovernmental organizations) were considered more trustworthy (50, N = 40; and 48.75%, N = 39, respectively) than private stakeholders (17.5%, N = 14). The participants showed a low level of trustworthiness in private stakeholder organizations (47.5%, N = 38). A majority of the participants were, however, "indecisive or neutral" in their assessments. Not a single participant considered private stakeholders "extremely trustworthy."

Perceived Threats and Opportunities

We structured our qualitative findings around two broad areas of OHD: threats (with themes of privacy and anonymity, abuse and misuse of data, and inaccurate data) and application areas (with themes of scientific research, health data analytics, improved disease prevention, novel health and software solutions, and unknown/future health problems).

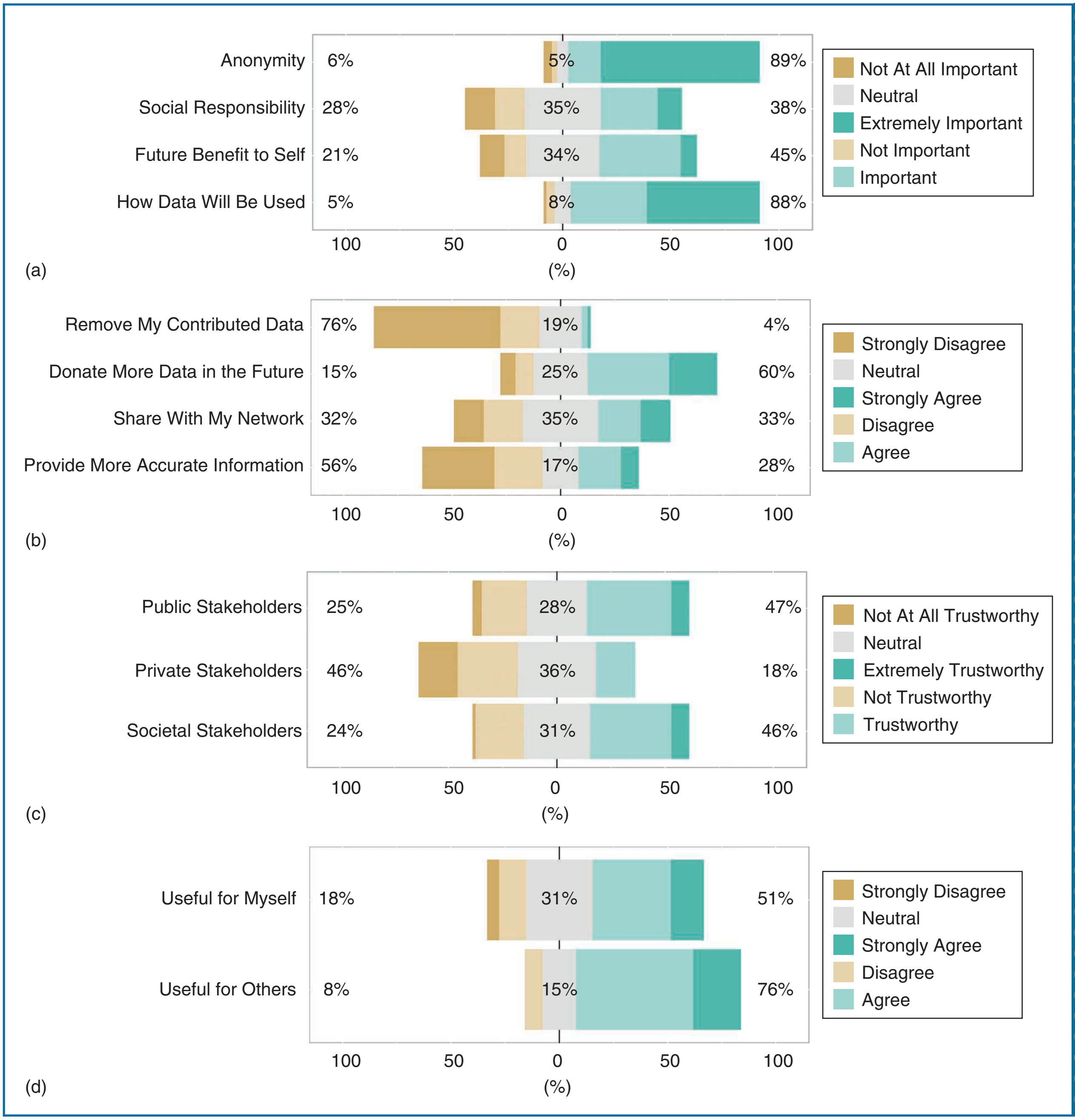


FIGURE 1. The Likert-scale responses to questions concerning (a) aspects to consider prior to donating data, (b) user opinions about their data after using the designed tool, (c) stakeholder trustworthiness, and (d) the perceived usefulness of the tool.

Threats of Health Data Reuse

The privacy and anonymity of personal information was of particular importance to participants. Participant four (P4) mentioned the existing threat of "deanonymization" and how it has "harmed the reputation of hundreds of people" after data about them was hacked. Some of the participants could not foresee any harm in collecting health data, stating "I don't think it would be harmful to collect health data anonymously," (P10) and "I don't see any threats if the person remains anonymous" (P31). Others, however, were outspoken about the negative impact of open data reuse overshadowing its benefits, "The risks might outweigh the benefits, in my opinion" (P11). One respondent summed it up as the "end of privacy" (P1). The participants also expressed concern about becoming potentially "targeted" (P33) if their health information becomes "accessible by health insurance companies" (P3).

The participants highlighted various potential abuses and misuses of their health data. Health data abuse was perceived as a "threat," (P67) especially in situations of "improper use" (P32) where individuals or companies "use it as their own" (P67) and "sell the data" (P68). The participants also recognized the sensitivity of health data: "health data are exceptionally sensitive" (P11); and how it could "end up in the wrong hands, used inappropriately" (P11); "used differently than was intended (P41);" or "use the information in a menacing way" (P53); resulting in "serious consequences for the individual" (P11). Others also highlighted how malicious users could take advantage of such data to "create a tool that targets vulnerable people" (P26) and "use it for their own gain" (P53).

Some of the participants were particularly worried about the quality of donated data. For instance, P14 stated that the data could be "misinformative" because "there could be bad quality creeping in" (P63) and that people could "intentionally provide wrong data" (P38). As such, the "actual legitimacy of the data, how accurate and truthful it is" (P37) comes into question.

Application Areas of OHD

The participants expressed enthusiasm toward how OHD could transform the development of digital health software. Several participants saw promise in using OHD to improve existing solutions: "improving the products of the software companies, keeping them accurate and up to date" (P34) and "to improve the current services provided but also to produce new services for use in public" (P27). Others were more specific in eliciting how OHD could be explored to target various user groups or even target specific health issues, "it can be used to develop [computer] programs targeted to specific groups and problems" (P48).

OHD presents "great opportunities" (P3) for the development of "tools to improve the health of the 'average Joe'" (P12). It would help open up the development of "software and apps with accurate and customized results" (P4) that would provide "detailed and personalized health advice to its users" (P3). OHD has the capacity to "help target unidentified problems or provide novel solutions that were not previously apparent" (P19) or even offer "health support."

We observed that participants put much value in the use of OHD for scientific research. OHD, our participants believe, is critical to "medical scientific studies" (P58) as it has potential to increase the "amount of data for researchers" (P5) based on the premise that "more data will always be helpful in finding answers to scientific questions, particularly if these questions relate to health" (P64). Another participant was of the opinion that diversity in OHD could propel research in previously understudied areas: "there are so many aspects of women's health that go unstudied because of lack of interest/funding. Donating health data is one way to get around this block as it is a relatively inexpensive method for collecting large amounts of data from a diverse group of people" (P52).

The availability of OHD could help "detect health conditions" (P18). Making health data open means that more hands are available to work on such data, increasing the possibility of identifying "patterns to prevent future diseases/health problems" (P3) and also discovering previously "unidentified problems" (P19) that already exist within the population. Thus, collective OHD "could be useful in diagnosing and treating a variety of health issues" (P42).

OHD has the potential to provide insights that were previously unknown. By making health data open, we would unlock the opportunity to conduct various forms of analysis on the data, which could help "build a picture of inequalities in health among certain groups and make it possible to provide a service that isn't available to particular groups of people" (P40). P4 believes OHD could help unveil "more accurate statistics on common symptoms or diseases that people are not publicly ready to talk about." With such a vast amount of data, an analysis could be conducted "based on demographics" (P9) to understand the health of "people at a particular age" (P20) or on even certain diseases to "determine their causes" (P39) as it can be a "very effective way to establish patterns between people with similar health state" (P49).

ABOUT THE AUTHORS

ANDY ALORWU is a doctoral researcher with the Crowd Computing Research Group at the Center for Ubiquitous Computing, University of Oulu, 90570, Finland. His research interests include personal data management and privacy, m-health, and mobile computing. Alorwu received an M.Sc. in information processing science from the University of Oulu, Finland. Contact him at andy.alorwu@oulu.fi.

NIELS VAN BERKEL is an associate professor with the Human-Centered Computing Group at Aalborg University, Aalborg, 9220, Denmark. His research interests include human–computer interaction, and social and ubiquitous computing. van Berkel received a Ph.D. in computer science and engineering from the University of Melbourne, Australia. Contact him at nielsvanberkel@cs.aau.dk.

AKU VISURI is a postdoctoral researcher with the Crowd Computing Research Group at the Center for Ubiquitous Computing, University of Oulu, 90570, Finland. His current research interests focus on ubiquitous computing and quantified-self applications and technologies, and digital health and well-being. Visuri received a Ph.D. in computer science from the University of Oulu, Finland. Contact him at aku.visuri@oulu.fi.

SIMO HOSIO is an associate professor at the Center for Ubiquitous Computing, University of Oulu, 90570, Finland, where he leads the Crowd Computing Research Group, and he is affiliated with the Center for Life Course Health Research. His research interests include crowdsourced well-being solutions, and social and ubiquitous computing. Hosio received a Ph.D. in computer science and engineering from the University of Oulu, Finland. Contact him at simo.hosio@oulu.fi.

Discussion

Our results highlight how OHD is perceived to have broad potential: it was seen as suitable for the creation of health and wellness-related software applications, fuel scientific research, new knowledge, and for fostering the detection and prevention of previously unknown diseases. This is partially in line with related work[3,7,] and a particular strength of our exploration is the fact that participants had "skin in the game" after having explored a software tool created using OHD they themselves contributed.

On the Future Use of OHD

We found that despite donating data as "open," participants still wanted to have a say on how the data are eventually used. Specifically, by whom, for what, and where. This exemplifies an unconscious perception of still owning the donated data despite having given away the rights to them. Such perceptions may foreshadow a deeper divide between user attitudes toward OHD donation and the use of such OHD for future software development by private stakeholder entities such as pharmaceutical and insurance companies, toward which participants expressed aversion.[7] Considering the poor trust of our participants toward private stakeholders, and their highlighting of possible data abuse by insurance companies, it is evident that people are concerned about who might use their data in the future. The concern is understandable given the challenge in predicting the long-term effects on privacy,[7] potential abuse of data,[8] and the additional risk of one's identity being revealed if two or more personal data are combined irrespective of being anonymized.[9] One potential avenue to explore here is for other stakeholders (public and societal) to join the development of software solutions based on OHD as they command more trust among people in using the data for broader societal benefit.

The participants' position on anonymity due to privacy concerns may present roadblocks for public health software solutions that require identifiable user data.[5] The deidentification of user data can limit the ability to analyze the data and target specific (demographic) groups to see what conditions may be prevalent in those groups. More research is needed to unpack and understand user perceptions and expectations regarding the digital anonymity and privacy of their personal data donated to open data initiatives. Also, the threats of privacy, deanonymization, commercial use of OHD, and abuse or misuse of them by entities (e.g., insurance companies) as mentioned by our study participants are in line with previous studies.[4,7–9]

The benefits of opening up and sharing health-related data are extensive as they may provide access to rare data, which is critical to the development of software solutions that bring an understanding of specific diseases, and offer a means to improve long-term care conditions (including self-care), in line with previous work.[3,10] Our respondents expressed a strong focus on research, disease diagnosis and prevention, and the development of health-related software solutions using OHD similar to those of previous studies.[1,3,8]

Our results also highlight that crowd workers seem to be interested in donating their health data toward open data initiatives, which is a promising development. Combining this insight with participants' wishes to retain a degree of control over their data, it is reasonable to assume that new data management models[2] are essential to explore right now.

Toward a New Paradigm

One particularly closely connected movement to participants' hopes about retaining control over their data is MyData.[2] MyData is an emerging, human-centric data management model and set of guidelines that aims to empower people to access, use, manage, and give permissions to their personal data. The sensitivity of personal health makes it a pioneering economic asset class that will affect all aspects of society.[2] In this regard, the software industry could benefit from this data as they are critical to the improvement of their processes and is an important resource for artificial intelligence-based software solutions.[3] MyData, should it take root, can be instrumental in facilitating the creation of future digital health software that use people's health data as core building blocks.

Our study investigated crowd workers' perceptions toward the reuse of their OHD in software solutions. To elicit perspectives in a realistic manner, we presented participants a tool based on their previously contributed health data. Our findings highlight threats and opportunities toward the use of OHD as embedded in future software solutions.

References

1. M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, adoption barriers and myths of open data and open government," *Inform. Syst. Manage.*, vol. 29, no. 4, pp. 258–268, 2012. doi: 10.1080/10580530.2012.716740.
2. A. Poikola, K. Kuikkaniemi, O. Kuittinen, H. Honko, A. Knuutila, and V. Lähteenoja, "Mydata—An introduction to human-centric use of personal data," Ministry of Transport and Communication, 2020. [Online]. Available: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/162405/MyData%20-%20introduction%20to%20human-centric%20use%20of%20personal%20data%203rd%20revised%20edition.pdf?sequence=1
3. S. Dolley, "Big data's role in precision public health," *Frontiers Public Health*, vol. 6, pp. 68, Mar. 2018. doi: 10.3389/fpubh.2018.00068.
4. P. Kostkova et al., "Who owns the data? open data for healthcare," *Frontiers Public Health*, vol. 4, p. 7, Feb. 2016. doi: 10.3389/fpubh.2016.00007.
5. E. G. Martin, N. Helbig, and G. S. Birkhead, "Opening health data: What do researchers want? early experiences with New York's open health data platform," *J. Public Health Manage. Pract.*, vol. 21, no. 5, pp. E1–E7, 2015. doi: 10.1097/PHH.0000000000000127.
6. A. Parameswaran, A. D. Sarma, and V. Venkataraman, "Optimizing open-ended crowdsourcing: The next frontier in crowdsourced data management," *Bull. Techn. Committee Data Eng.*, vol. 39, no. 4, pp. 26–37, 2016.
7. M. J. Bietz et al., "Opportunities and challenges in the use of personal health data for health research," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. e1, pp. e42–e48, 2016. doi: 10.1093/jamia/ocv118.
8. S. Kalkman, J. van Delden, A. Banerjee, B. Tyl, M. Mostert, and G. van Thiel, "Patients' and public views and attitudes towards the sharing of health data for research: A narrative review of the empirical evidence," *J. Med. Ethics*, 2019. doi: 10.1136/medethics-2019-105651.
9. L. Sweeney, A. Abu, and J. Winn, "Identifying participants in the personal genome project by name (a re-identification experiment)," 2013, arXiv:1304.7605.
10. S. Courbier, R. Dimond, and V. Bros-Facer, "Share and protect our health data: An evidence based approach to rare disease patients' perspectives on data sharing and data protection-quantitative survey and recommendations," *Orphanet J. Rare Dis.*, vol. 14, no. 1, pp. 1–15, 2019. doi: 10.1186/s13023-019-1123-4.