(Re)using Crowdsourced Health Data

Perceptions of Data Contributors

Alorwu, Andy; Visuri, Aku; van Berkel, Niels; Hosio, Simo Johannes

Published in:
IEEE Software

DOI (link to publication from Publisher):
10.1109/MS.2021.3117684

Publication date:
2022

Document Version
Accepted author manuscript, peer reviewed version

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

A. Alorwu
University of Oulu

A. Visuri
University of Oulu

N. van Berkel
Aalborg University

S. Hosio
University of Oulu

***Abstract*—Open data is often contributed by various governments and public sector actors. An increasingly popular way to collect large bespoke datasets is crowdsourcing. In this work we explore crowdsourced open data as an enabler of future software solutions. We recruited participants from an online paid crowdsourcing platform to provide open mental health related data that was used to create an interactive data-driven decision support system for self-care. We then invited a sub-sample of 80 participants back to explore the tool that was created using their own data and to provide a rich account of perceptions on issues around such health data reuse in software. Our results unfold a range of different perceived threats and opportunities in using crowdsourced data to enable software solutions, and our work contributes a topical case study and discussion toward the use of crowdsourced data in an open fashion.**

OPEN DATA has been predominantly propelled by governments and public organizations sharing meaningful datasets for others to build on [1, 2]. As personal digital technologies have rapidly proliferated, most of us now produce a constant stream of data daily – data that can be used to build different types of novel software solutions and services [3, 4]. A particularly interesting class of such data is health-related data, collected now pervasively by fitness trackers, wearable sensors, and increasingly even our smartphones through built-in health monitoring tools [5]. Such data are typically under the control of corporations that provide the infrastructure, *e.g.*, smartphones, fitness trackers, social media software, through which such data is generated [4]. In this context, *Open Health Data* (OHD) refers to any type of publicly accessible health related data [6, 7]. Coined as the act of outsourcing a job to an undefined group of people in the form of an open call [8], *crowdsourcing* has become a primary means of collecting high-quality data from people at scale [9]. 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 remains critically under-explored. An

understanding of user perceptions is needed to help the software community take steps to alleviate data donor fears and concerns. In this paper, 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. Our main contributions are:

1) We outline and detail concerns people experience in the context of Open Health Data in software systems, including privacy concerns, potential for misuse of data, and the accuracy of data.
2) We identify a range of perceived threats and opportunities in using crowdsourced data to enable software solutions, outlining opportunities for future work.
3) We highlight the importance of the involvement of public and societal stakeholders in software development efforts that rely on open data as they command public trust.

Data collection

We invited 80 of the participants who donated data for the decision support system back to participate in an online questionnaire study. All participants were fluent in English, students in a higher education institution, and had completed at least 50 tasks on Prolific, making them 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 that was built using their own OHD.

How we conducted the study

Participants of our study had earlier been invited 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 data they provide will be used openly in research and as an open dataset accessible for anyone online: Open Health Data. The decision support tool can analyse such data and turn it into an interactive exploration interface that is helpful in finding suitable self-care techniques, as seen in Fig. 1. While the tool itself is out of scope of this paper, more details can be found in [10]. Our study was enabled by Prolific [11] - a crowdsourcing marketplace that has been used for many academic studies. We used a Mixed Methods approach, a method that 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 qualitative or quantitative method alone [12].

The average completion time of the study was 17.34 min, and participants were compensated £3.00 per contribution, making the mean hourly wage well above typical crowd work standards.

Participants' age ranged between 20-54 (M=26.26 years, SD=5.72 years, 52 males, 28 females). 83.75% (N=67) of participants considered themselves knowledgeable with technology. Further, 83.75% (N=67) stated that they collect health-related data about themselves regularly.

Data donation and trust

Donating data towards 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*

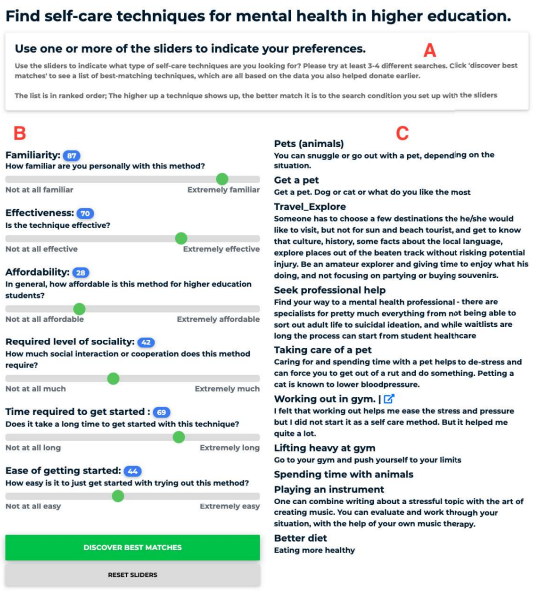

Figure 1. Interface of the Decision Support System developed to suggest mental health self-care techniques to participants [A. General instructions B. Criteria for finding self-care techniques C. Suggested self-care techniques based on the selected criteria]

myself" with 45% (N=36) and lastly, *"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 2-C. To this end, public stakeholders (governmental agencies and public research organizations), and societal stakeholders (non-governmental organizations) were considered more trustworthy (50%, N=40 and 48.75%, N=39, respectively) than private stakeholders (17.5%, N=14). Participants showed a low level of trustworthiness in private stakeholder organizations (47.5%, N=38). A majority of participants were, however, *indecisive or neutral* in their assessments. Not a single participant considered private stakeholders as *"extremely trustworthy"*.

Perceived threats and opportunities

We conducted a thematic analysis in line with [13], following a deductive approach to uncover data contributor perceptions specifically concerning threats and opportunities of OHD. We analysed 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 the 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. The authors conducted multiple online meetings to discuss and resolve disagreements with the coding. We structure our qualitative findings around two broad areas of OHD: **threats** (with themes: *privacy and anonymity, abuse and misuse of data, inaccurate data*) and **application areas** (with themes: *scientific research, health data analytics, improved disease prevention, novel health and software solutions, and unknown/future health problems*).

Threats of health data reuse

Privacy and anonymity of personal information was of particular importance to participants. The existing threat of *"de-anonymization"* (P4) was mentioned, and how it has *"harmed the reputation of hundreds of people"* (P4) after data about them was hacked. Some participants could not foresee any harm in collecting health data, *"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 sums it up as the *"end of privacy"* (P1). Participants also expressed concern of becoming potentially *"targeted"* (P33) if their health information becomes *"accessible by health insurance companies"* (P3).

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). Participants also recognized the sensitivity of health data, *"health data is exceptionally sensitive"* (P11) and how it could *"end up in the wrong hands and used inappropriately"* (P11), *"used*

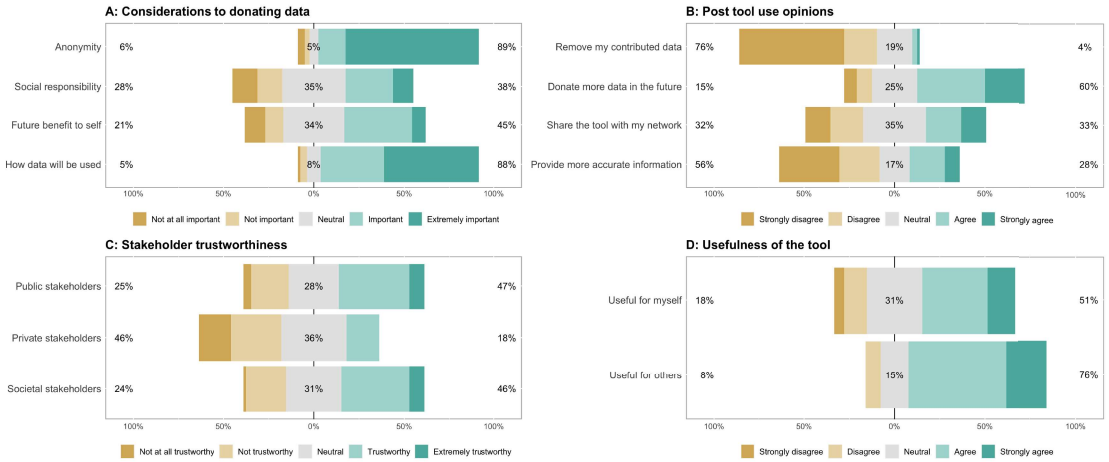

Figure 2. Likert scale responses

differently than was intended" (P41), or *"use the information in a menacing way"* (P53) which could have *"serious consequences for the individual"* (P11). Others also highlighted how malicious users could take advantage of such data to *"create a tool which targets vulnerable people"* (P26) and *"use it for gains to themselves"* (P53).

Some 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 open health data

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 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 open 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 which isn't available to a particular group 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 vast amount of data, analysis could be made *"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).

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, creation of new knowledge, and fostering the detection and prevention of previously unknown diseases. This is partially in line with related work [3, 5], 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 contributed themselves.

On the future use of open health data

We found that despite donating data as 'open', participants still wanted to have a say on how the data is eventually used. Specifically by *who*, for *what*, and *where*. This exemplifies an unconscious perception of still owning the donated data despite having given away the rights to the data. Such perceptions may foreshadow a deeper divide between user attitudes towards OHD donation and the use of such OHD for future software development by private stakeholder entities such as pharmaceutical and insurance companies to which participants expressed an aversion [3]. Considering the poor trust of our participants towards 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 down the line. The concern is understandable given the challenge in predicting the long-term effects on privacy [3], potential abuse of data [14], and the additional risk of one's identity being revealed if two or more personal data are combined irrespective of being anonymized [15]. 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.

Participants' position on anonymity due to privacy concerns may present roadblocks for public health software solutions that require identifiable user data [7]. The de-identification 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 digital anonymity and privacy of their personal data donated towards open data initiatives. Also, threats of privacy, de-anonymization, commercial use of OHD, and abuse or misuse of OHD by entities such as insurance companies as mentioned by our study participants are in line with [3, 6, 14, 15].

The benefits of opening up and sharing health-related data are extensive, as it may provide access to rare data that is critical to the development of software solutions that bring understanding of specific diseases and offers a means to improve long-term care conditions (including self-care) in line with previous studies [5, 16]. Our respondents expressed a strong focus on research, disease diagnosis and prevention, and development of health-related software solutions using OHD similar to [1, 5, 14].

Our results also highlight that crowd-workers seem to be interested in donating their health data toward open data initiatives, which can be considered 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 [4] 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* [4]. 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 [4]. In this regard, the software industry could benefit from this data as it is critical to the improvement of its processes and is an important resource for AI-based software solutions [5]. 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.

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.

Conclusion

Our study investigated crowd-workers' perceptions towards the reuse of their OHD in software solutions. In order 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 towards 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," *Information systems management*, vol. 29, no. 4, pp. 258–268, 2012.
2. J. Linåker and P. Runeson, "Collaboration in open government data ecosystems: Open cross-sector sharing and co-development of data and software," in *International Conference on Electronic Government*. Springer, 2020, pp. 290–303.
3. M. J. Bietz, C. S. Bloss, S. Calvert, J. G. Godino, J. Gregory, M. P. Claffey, J. Sheehan, and K. Patrick, "Opportunities and challenges in the use of personal health data for health research," *Journal of the American Medical Informatics Association*, vol. 23, no. e1, pp. e42–e48, 2016.
4. A. Poikola, K. Kuikkaniemi, O. Kuittinen, H. Honko, A. Knuutila, and V. Lähteenoja, "Mydata—an introduction to human-centric use of personal data," *Finnish Ministry of Transport and Communications, 55p*, vol. 8, 2020.
5. S. Dolley, "Big data's role in precision public health," *Frontiers in public health*, vol. 6, p. 68, 2018.
6. P. Kostkova, H. Brewer, S. de Lusignan, E. Fottrell, B. Goldacre, G. Hart, P. Koczan, P. Knight, C. Marsolier, R. A. McKendry *et al.*, "Who owns the data? open data for healthcare," *Frontiers in public health*, vol. 4, p. 7, 2016.
7. 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," *Journal of Public Health Management and Practice*, vol. 21, no. 5, pp. E1–E7, 2015.
8. J. Howe, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.
9. A. Parameswaran, A. D. Sarma, and V. Venkataraman, "Optimizing open-ended crowdsourcing: the next frontier in crowdsourced data management," *Bulletin of the Technical Committee on Data Engineering*, vol. 39, no. 4, p. 26, 2016.
10. S. Hosio, J. Goncalves, T. Anagnostopoulos, and V. Kostakos, "Leveraging wisdom of the crowd for decision support," in *Proceedings of the 30th International BCS Human Computer Interaction Conference 30*, 2016, pp. 1–12.
11. S. Palan and C. Schitter, "Prolific. ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.
12. J. Wisdom and J. W. Creswell, "Mixed methods: integrating quantitative and qualitative data collection and analysis while studying patient-centered medical home models," *Rockville: Agency for Healthcare Research and Quality*, 2013.
13. V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research*

in psychology, vol. 3, no. 2, pp. 77–101, 2006.

14. 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," *Journal of medical ethics*, 2019.
15. L. Sweeney, A. Abu, and J. Winn, "Identifying participants in the personal genome project by name (a re-identification experiment)," *arXiv preprint arXiv:1304.7605*, 2013.
16. 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 journal of rare diseases*, vol. 14, no. 1, pp. 1–15, 2019.