

To share or not to share – understanding individuals' willingness to share biomarkers, sensor data, and medical records

Ruben L. Bach ^a, Henning Silber ^b, Frederic Gerdon ^a, Florian Keusch ^c, Matthias Schonlau^d and Jette Schröder ^b

^aMannheim Centre for European Social Research (MZES), University of Mannheim, Mannheim, Germany; ^bGESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany; ^cSchool of Social Sciences, University of Mannheim, Mannheim, Germany; ^dDepartment of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

ABSTRACT

Technological advances in the recent past made it possible for researchers to collect and analyze large amounts of health data at unprecedented scale and speed. For example, fitness trackers and smartwatches produce steady flows of information on individuals' health. Biomarker data and medical records allow to study individuals at new levels of granularity. The COVID-19 pandemic has highlighted that access to such data for health research and evidence-based public policy decision-making is essential. However, having access to data depends on individuals' willingness to share their data with others. In this paper, we analyze the factors that may affect the probability of individuals to share their biomarker, health, and sensor data using German survey data and a survey experimental vignette design. We study the impact of data type, recipient, and research purpose on respondents' willingness to share their data as well as the effects of respondents' own medical and data sharing history. Overall, participants' willingness to share biomarker data was higher than the willingness to share other data types. Moreover, those who had shared data before were more willing to do so again. In addition, natural language processing analysis of textual responses capturing respondents' motives to share their data shows that individuals do understand how valuable their data is for researchers. However, results also underscore that addressing concerns about the protection of data need to be taken seriously. Emphasizing the value of data shared for research and their purpose may help to increase trust and willingness to share data.


ARTICLE HISTORY

Received 4 August 2023
Accepted 2 April 2024

KEYWORDS

Data sharing willingness; sensor data; open-ended survey responses

CONTACT Ruben L. Bach  r.bach@uni-mannheim.de  Mannheim Centre for European Social Research (MZES), University of Mannheim, A5, 6, Mannheim 68159, Germany

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/1369118X.2024.2351439>.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Introduction

Collecting and analyzing large amounts of health data at unprecedented scale, speed, and granularity has become easier than ever before. For example, fitness trackers and smart-watches produce steady flows of information on individuals' health. Biomarker data and medical records allow researchers to study individuals at new levels of granularity. The COVID-19 pandemic has highlighted that access to such data for health research and evidence-based policy-making around the world provides benefits (Pandit et al., 2022). However, there are also concerns about the data quality and therefore the usefulness of 'digital epidemiology' (Klingwort & Schnell, 2020). Moreover, novel data require new ethical considerations (Mello & Wang, 2020). Crucially, having access to data depends on individuals' willingness to share their data with others (Aitken et al., 2016). To evaluate data quality and to appropriately realize the full potential of digital health research, we need to understand the data-generating processes. Questions about who is and who is not included in the data and what reasons are behind non-participation need to be answered.

Through a survey experiment, we investigate differences in acceptance between specific data types, recipients, and purpose of use, which are decisive factors in individuals' judgments of the acceptability of data sharing according to contextual integrity (CI) theory (Nissenbaum, 2010). In addition, our study systematically evaluates the underlying reasons for respondents' choices through open-ended text data. Our setup builds on survey experimental vignette designs developed to study individuals' willingness to share data (Gerdon et al., 2020; Karwatzki et al., 2017; Keusch et al., 2019; Silber et al., 2022). We implemented a preregistered survey experiment in the German Internet Panel (GIP), a probability-based online panel of the German population, in May 2022. We asked respondents to indicate how willing they were to share biomarkers, sensor data, and medical records with public health agencies, university researchers, and private companies. We study the purpose of data use as a context factor and the influence of individuals' characteristics on their data sharing decisions. Moreover, we asked respondents to explain why they were (un)willing to share their data using an open-ended question. This allows us to get a nuanced understanding of respondents' willingness to share their data beyond the CI-specific factors and individual characteristics that previous literature has looked at. Through an unrestricted answer format, we obtain information about individuals' concerns but also motivations when deciding whether to share their data or not. That is, by combining a deductive experimental approach with an inductive analysis of the concerns and motivations that respondents have, we contribute to the growing literature on data sharing decisions and willingness to share data.

The main research question of this study is: How do contextual factors and personal characteristics influence individuals' willingness to share their health data, and what are the underlying reasons for their decisions? In the next section, we present our theoretical framework in more detail.

Theoretical framework

Nissenbaum's (2010) theory of *contextual integrity* (CI) provides a context-specific view on the appropriateness of data flows. According to CI, data flows take place when

information about an individual is transferred from one party to another, for example, when moving a patient's health data between doctors and insurance or when transferring data to a research institution. For judging the appropriateness of a data flow, CI postulates that the contextual parameters of the data flow need to be defined. Based on such a description of a data flow, one can check whether the data flow is in line with context-specific privacy norms, i.e., appropriate. To arrive at a meaningful description of data flows, CI suggests defining the following contextual parameters: the data type, the actors (data sender, subject, and recipient), and the transmission principles (Nissenbaum, 2010). The transmission principles represent the prerequisites of the data flow, which could be, e.g., the data subjects' consent.

Societal challenges such as COVID-19 have shown that data flows can serve different purposes. For instance, while digitally stored health data can be used to tailor health-related recommendations to individuals, they could also be used for public health management and research. Nissenbaum highlights the relevance of purpose as one of the defining constituents of context (Nissenbaum, 2019).

Following CI, individuals differentiate their evaluations of data flows concerning the outlined contextual parameters. In the next section, we turn to specific factors that were found to affect the sharing of health data for personal and public benefits.

Previous research

Previous literature documents that individuals care about who has access to health data (e.g., Aitken et al., 2016; Hutchings et al., 2020; Skovgaard et al., 2019; Stockdale et al., 2019). The literature suggests that willingness to share health data for public benefit purposes might be given under certain conditions, but that it is crucial to keep the data of the individuals safe from misuse. In the following, we review previous findings on the CI context parameters for the specific case of health data.

Different *data types* may strongly vary in their perceived sensitivity and the required effort to share the data (e.g., Aitken et al., 2016; Grande et al., 2022; Habich-Sobiegalla & Kostka, 2023; Skovgaard et al., 2019). Here, we focus on biomarkers, medical records, and health-related sensor data from smart devices. Silber et al. (2022), for example, found differences in the willingness to share data between these three data types and argued that variation in the *effort needed* to share the data explains this variation. Sharing biomarkers, such as blood or urine samples, requires more effort compared to, for example, medical records, which are already recorded at health insurance companies and doctor's offices. Sensor data, for example, from smartwatches, are collected passively, and the individual whose data are concerned can share them with little effort. Furthermore, a clear understanding of the data that is transferred seems to increase individuals' acceptance to do so (e.g., Breuer et al., 2023; Gomez Ortega et al., 2023), which might be particularly challenging for linking survey with other types of data (see, e.g., Sloan et al. (2020)).

Concerning data recipients, trust, e.g., in scientific institutions, is an important component for individuals' data sharing decisions (Hutchings et al., 2020). The public sees potential for inappropriate data use by both public agencies and commercial actors (Aitken et al., 2016) and both are distrusted by parts of the population (Hutchings et al., 2020). Commercial actors may be perceived as acceptable data recipients if a public

benefit is prioritized over profit and if a public value is generated (Aitken et al., 2016). For health-related sensor data, Gerdon et al. (2020) found that public agencies are less accepted recipients than private companies.

A third factor that may affect individuals' willingness to transfer data is the *purpose*. Beyond private benefits, various public benefits may arise from health data use. For example, swift access to high-quality health data is essential for fighting pandemics (Pandit et al., 2022). Previous research suggests a temporary effect of the COVID-19 pandemic on the increased acceptance of health data use for different health-related purposes (Gerdon et al., 2020; Goetzen et al., 2022; Jörling et al., 2023). Social duty seems to predict increased willingness to share data for a public benefit (Skatova & Goulding, 2019).

Furthermore, respondent characteristics may matter for the willingness to share data (see, e.g., Grande et al., 2022; Hutchings et al., 2020; Karampela et al., 2019). For example, an individual's own *medical history* may affect their data sharing choices. In line with the privacy calculus (Culnan & Armstrong, 1999), rational-choice-type cost-benefit calculations may make individuals with a documented medical history become more willing to share their data when the benefit is personal.

Hypothesis development

We build our hypothesis development on the aforementioned theory of CI and insights from prior research on health data sharing. CI's emphasis on the appropriateness of data sharing, contingent upon adherence to context-specific norms, guides our investigation into individuals' willingness to share health-related data.

Hypothesis 1 (H1): The willingness to share biomarker data is lower than the willingness to share data of the two other data types.

This hypothesis arises from the consideration that biomarker data, requiring significant effort to collect and share, is anticipated to be less readily shared. This aligns with Silber et al. (2022)'s argument that variation in the effort needed to share the data explains some variation between data types.

Hypothesis 2.1 (H2.1): The willingness to share data is higher if the data recipient is a public agency compared to a private company.

This hypothesis reflects the critical role of trust in the data recipient. Public agencies, perceived as prioritizing public interest, are assumed to engender higher trust as data recipients due to the better fit of context, following CI.

Hypothesis 2.2 (H2.2): The willingness to share data is higher if the data recipient is a university research center compared to a public health agency.

This suggests that universities, with their strong association with research and public benefit, are viewed as highly trustworthy recipients. It also emphasizes the importance of the recipient's context fit in data sharing decisions.

Hypothesis 3 (H3): Respondents are more willing to share data if the purpose of data sharing brings a public benefit compared to a personal benefit.

This hypothesis addresses the purpose behind data sharing, underscoring a societal inclination towards altruism and the collective good in the context of health data against the background of the COVID-19 pandemic.

Hypothesis 4 (H4): Public recipients are more accepted for public benefit purposes, whereas private recipients are more accepted for personal benefit purposes.

This hypothesis proposes a nuanced interaction between the type of data recipient and the perceived purpose of data sharing due to context fit, reflecting CI's emphasis on context-specific norms.

Hypothesis 5.1 (H5.1): Less healthy respondents are less likely to share their medical records for a public benefit than for personal benefit since they may hope to get advice on their health condition.

This suggests that individuals with a documented medical history may prioritize immediate personal benefits in their data sharing decisions, a reflection of the privacy calculus model.

Hypothesis 5.2 (H5.2): There will be a spill-over effect on the other two data types so that less healthy respondents are also less likely to share biomarkers or their sensor data for a public benefit than for a personal benefit.

This hypothesis builds on the logic of H5.1 and extends it to additional data types.

Hypothesis 6.1 (H6.1): Respondents who have previously shared a specific data type are more willing to share that data type again.

This suggests that past positive experiences with data sharing can increase one's propensity to share data, reflecting learned trust and familiarity.

Hypothesis 6.2 (H6.2): The more data people have shared previously, the more likely they will share data of any given type.

This builds on the logic of H6.1 and extends it to the broader context of data sharing behavior.

Through these hypotheses, embedded in the theoretical framework of CI and previous research, we aim to understand the nuanced factors influencing health data sharing preferences. This approach not only aligns with CI's emphasis on context-specific norms but also seeks to enrich the discourse on privacy and data ethics in the digital era by providing empirical insights into individuals' data sharing behaviors.

Research methods

To test our hypotheses, we implemented a preregistered survey experimental vignette study (see https://osf.io/pygx5?view_only=f8e78bf869324c94b942d8a1598fe85d for the preregistration report) in the German Internet Panel (GIP) (Blom et al., 2015). Our study featured a $3 \times 3 \times 2$ factorial design (Auspurg & Hinz, 2015). The experiment manipulated the data type (medical history, biomarkers, sensor data), the recipient (public health agency, university, private company), and the purpose (public benefit, personal recommendation). Each participant received one randomly chosen vignette per data type, i.e., three out of the 18 vignettes (see the Online Appendix for the wording). The

order of the vignettes was randomized among respondents to avoid systematic bias due to order effects. Each vignette included one of the three treatments regarding the recipient and one of the two treatments regarding purpose. After the first vignette that a respondent received, we asked respondents to explain why the specific answer category was selected using an open-ended question (see below for details). We use the experimentally varied dimensions of data type, recipient, and purpose to test our hypotheses H1, H2.1, H2.2, H3, and H4.

Table 1 shows the dimensions and levels of the vignette study. The structure and the text of the vignettes were inspired by previous research (Gerdon et al., 2020; Silber et al., 2022). An example of a vignette is given here:

Medical records provided by health insurance companies may be used to assess the health status of individuals. With an individual's consent, these data are transmitted to a German public health agency. This public health agency uses these data to detect the spread of infectious diseases in the population at an early stage and develop solutions to contain them. The public health agency ensures that the data is secure, anonymous, and protected from misuse.

After each vignette, respondents were asked to indicate their willingness to transmit their data for this purpose on a five-point rating scale ('1 – very unlikely', '2 somewhat unlikely', '3 – neither unlikely nor likely', '4 – somewhat likely', '5 – very likely') using the following question: 'What about you, how likely or unlikely is it that you would consent to your health information being shared for this purpose?' Following this closed-ended question, respondents were prompted to elaborate on their response using an open-ended question: 'Can you please tell us why you answered that it is [response to closed-ended question] that you would share your health information in the situation described before?' This question was only asked after the first vignette a respondent saw. Respondents typed their textual answers in an answer box below the question.

Table 1. Dimensions, levels, and wording of the vignettes.

Dimension	Levels	Wording	Hypotheses tested*
Data type	Medical records	Medical records provided by health insurance companies may be used to assess the health status of individuals.	H1
	Biomarkers	Blood samples may be used to assess the health status of individuals.	
	Sensor data	Sensors installed on smartphones, smartwatches, and other wearable devices collect data that may be used to assess the health status of individuals.	
Recipient	Public health agency	With the consent of an individual, these data are transmitted to a German public health agency.	H2.1 & H2.2 & H4
	University	With the consent of an individual, these data are transmitted to a German university.	
	Private company	With an individual's consent, these data are transmitted to a German private company.	
Research purpose	Public benefit	This [recipient] uses this data to detect the spread of infectious diseases in the population at an early stage and develop solutions to contain them.	H3 & H4
	Personal recommendation	This [recipient] uses these data to provide people with personalized recommendations to protect themselves from infectious diseases.	

Note: Structure of the vignettes: '[DATA TYPE]. [RECIPIENT]. [RESEARCH PURPOSE]. The [RECIPIENT] ensures that the data is secure, anonymous, and protected from misuse.' See the Online Appendix for the original wording in German.

*Hypotheses H5.1, H5.2, H6.1, and H6.2 were tested with survey items collected after the vignette experiment (see paragraph 'Other measures').

Other measures

In addition, to test hypotheses H5.1 and H5.2, we included a survey question that measured respondents' medical history using a multiple-choice format. Respondents were asked to indicate, for each of 15 diseases, whether they were ever diagnosed with it. Furthermore, we included a multiple-choice question on previous data sharing (blood, sensor data, medical records, and other data) to test our hypotheses H6.1 and H6.2. Per our preregistered design, we created indices for the two multiple-choice questions as simple sum scores. An attention check asking how carefully a person had read the vignettes was used to assess whether the results are robust regarding self-reported attentiveness during the experiment.

Transparent changes

Overall, we did not deviate in meaningful ways from the preregistered design. We made one small adjustment regarding the analysis (see Section 6). We made minor changes to the wording of hypotheses H3, H5.1, and H5.2 without altering the hypothesized relationships or their directions. Furthermore, we renamed the attention check, which was incorrectly called a 'manipulation check' in the preregistration report.

Per our preregistration, we fielded the study in May 2022 in the 59th wave of the GIP, a probability-based online panel survey of the German adult population (Blom et al., 2015). A total of 3,870 out of 5,907 invited panelists participated in our study between May 1, 2022, and May 31, 2022. Overall, 48.1% of our sample reported to be female, 51.8% to be male, and .1% reported another gender. 5.6% of all respondents were 26 years and younger, 30.2% between 27 and 46 years, 34.0% between 47 and 61 years, and 30.1% were 62 years and older. The majority of the sample (53.9%) had a high educational level (ISCED 5–8), followed by 32.6% with a medium educational level (ISCED 3–4). The remaining 13.6% of the sample reported a low educational level (ISCED 1–2). For details on the recruitment and sampling strategy of the GIP, see the Online Appendix.

As per our preregistration, we did not remove respondents with item nonresponse, including break-offs, from our analyses but dropped missing values on a pairwise basis. Missing values were not imputed. Approximated power analysis using an ANOVA design with repeated measures and within-between interaction using the software G*Power (Faul et al., 2009) with input parameters *effect size* = 0.1, *α-error probability* = 0.05, *power* = 0.80, *number of groups* = 18, *number of measurements* = 3, and *nonsphericity correction* = 1 suggests a sample size of 648 respondents. Thus, our sample exceeds the suggested sample size by a factor of six.

Statistical analysis

We apply a series of linear multilevel random effects models and linear regression models to analyze the closed-ended survey data using the *lme4*-package in R (Bates et al., 2015; R Core Team, 2022). As described in Section 5, each respondent rated a total of three vignettes, resulting in a total of 11,610 (3,870 respondents * three vignettes) data points. To account for the correlation of responses within respondents, we use multilevel random effects models. When focusing on one data type only, we use linear regression as each respondent answered only one vignette per data type (3,870 data points per respondent).

For specificity and clarity, we test each set of hypotheses as reported below with a separate model, which allows us to isolate and interpret the impact of individual variables.

Our preregistration specified four models. Model 1 contains the main effects only. We include predictors for data type, recipient, and purpose. This model tests hypotheses H1, H2.1, H2.2, and H3. Model 2 adds interaction effects between recipient and purpose, and tests hypothesis H4. In Model 3, we add a summary variable capturing respondents' medical history and its interaction with purpose to Model 1 to test H5.1 and H5.2. Differing from the preregistration, we do not add respondents' previous data sharing experiences to Model 3. Instead, we specify a separate Model 4, which contains main effects as in Model 1 plus information on respondents' previous data sharing experiences to test H6.1 and H6.2. Model 5 adds socio-demographic control variables to Model 1.

As per our robustness analysis, we estimate all models twice: once with all respondents and once excluding 101 respondents who indicated that they did not read the vignettes carefully (i.e., respondents who chose the two lowest values on a seven-point scale from 'not carefully at all' to 'very carefully'). Results without the inattentive respondents are shown in the Online Appendix (Tables A.5, A.6, and A.7). They do not meaningfully differ from the results shown here.

Regarding the inductive analysis of the response to the open-ended question, we opted for automated analysis using a natural language processing (NLP) approach. Specifically, we use the Bidirectional Encoder Representations from Transformers (BERT) language model (Devlin et al., 2018). BERT is a pretrained language model that can easily be adapted or fine-tuned for downstream tasks like text classification, topic modeling, and stance detection. Using a pre-trained, transformer-based approach like BERT often outperforms more traditional approaches, such as bag-of-words approaches in typical NLP tasks (see, e.g., Gasparetto et al., 2022; Gweon & Schonlau, 2022). We use the *Simple Transformers*-library in Python for this task (Rajapakse, 2022).

We analyze the textual responses by grouping them into categories that summarize concerns and motivations that users may have when sharing their data. One human coder with domain knowledge selected a random subsample of $n_{total} = 1,577$ textual responses (about $n = 500$ per data type) and inductively developed a coding scheme based on them (Table 2). Of all responses, 90 (about 2% of the full sample) did not make any sense as respondents put in a single character only or a meaningless response, such as 'No'. These responses are part of the 306 responses with the code 'unclear'. Note that, at this stage, our codes do not differentiate between a positive or a negative evaluation as these are inferred from the closed-ended survey data.

We randomly selected 60% ($n = 945$) of all $n = 1,577$ manually annotated responses as training data to fine-tune a pre-trained BERT model for German ('bertbase-german-cased', see <https://huggingface.co/bert-base-german-cased>). Pretrained language models typically work best when fine-tuned on a downstream task. We use another 20% ($n = 316$) of the manually annotated responses as validation data, and the final 20% ($n = 316$) as test data to evaluate the performance of our model on data unseen in the training and fine-tuning process. Last, we apply the fine-tuned BERT model to classify all 3,870 textual responses (including the manually annotated $n = 1,577$ responses). Details of the training and fine-tuning, including model performance on unseen test data, are shown in the Online Appendix. We feed raw textual responses without further preprocessing, cleaning, or editing to BERT's WordPiece tokenizer, the tokenizer used by the Simple Transformers library.

Table 2. Coding scheme for the open-ended question. The counts, N, indicate the frequency distribution of the codes in the manually annotated data.

Category	Definition	Example response	N
(Lack of) trust	Responses either mention that respondents trust data to be safe and protected or a lack of trust and concerns about misuse.	<i>Concerns about misuse of data.</i>	243 (15.4%)
Additional conditions	Responses that mention under which circumstances they would or would not share their data.	<i>I reject the disclosure of personal data in general. However, should it involve epidemic dimensions – then I would act cooperatively.</i>	34 (2.2%)
Missing information	More information needed to judge whether data are safe and protected.	<i>I am not sure where my personal health information ends up and what it is used for.</i>	24 (1.5%)
Data protection and privacy	Responses mention data protection and privacy.	<i>If it helps, I am for it. I think privacy is greatly exaggerated anyway.</i>	311 (19.7%)
Purpose of collected data	Response emphasizes or questions usefulness of data collection and use.	<i>Don't need a recommendation from the health department. See chaos during COVID-19.</i>	505 (32.0%)
Recipient	Responses mention issues related to data recipient, e.g., competence and use of data by recipient.	<i>Since a university is serious and value is added through research.</i>	142 (9.0%)
Other	Responses that do not match the other categories but convey a meaningful response.	<i>I hope for a saving because duplicate examinations could be avoided.</i>	134 (8.5%)
Unclear	Responses that are off-topic and/or do not make sense.	<i>That's why.</i>	184 (11.7%)
Total			1,577 (100.0%)

Note: Categories contain both negative and positive responses regarding the topic(s) mentioned. The original text of responses is in German. English translation by authors.

Results

Table 3 shows mean willingness ratings to share data for each of the 18 vignettes. Overall, willingness ranges from 2.11 (willingness to share sensor data for personal recommendations with a private company) to 3.14 (willingness to share biomarkers for personal

Table 3. Means of willingness ratings for each of the 18 vignettes.

Data	Purpose	Recipient	Mean	Median	SD	Lower	Upper	N
Biomarker	Personal recommendations	Public health agency	2.86	3.00	1.37	2.75	2.96	639
Biomarker	Personal recommendations	Private company	3.11	4.00	1.36	3.00	3.21	638
Biomarker	Personal recommendations	University	2.92	3.00	1.34	2.81	3.02	637
Biomarker	Public benefit	Public health agency	2.95	3.00	1.36	2.85	3.06	639
Biomarker	Public benefit	Private company	3.14	4.00	1.35	3.03	3.24	638
Biomarker	Public benefit	University	3.04	3.00	1.32	2.93	3.14	636
Medical records	Personal recommendations	Public health agency	2.74	3.00	1.35	2.64	2.85	639
Medical records	Personal recommendations	Private company	3.00	3.00	1.38	2.89	3.10	638
Medical records	Personal recommendations	University	2.77	3.00	1.31	2.67	2.87	637
Medical records	Public benefit	Public health agency	2.87	3.00	1.35	2.76	2.97	639
Medical records	Public benefit	Private company	2.94	3.00	1.34	2.83	3.04	638
Medical records	Public benefit	University	2.91	3.00	1.32	2.80	3.01	636
Sensor data	Personal recommendations	Public health agency	2.45	2.00	1.33	2.34	2.55	639
Sensor data	Personal recommendations	Private company	2.11	2.00	1.22	2.02	2.21	638
Sensor data	Personal recommendations	University	2.63	2.00	1.31	2.53	2.73	637
Sensor data	Public benefit	Public health agency	2.84	3.00	1.39	2.73	2.95	639
Sensor data	Public benefit	Private company	2.42	2.00	1.30	2.32	2.52	638
Sensor data	Public benefit	University	2.89	3.00	1.31	2.79	2.99	636

Note: SD = Standard deviation; '1 – very unlikely' to '5 – very likely'.

recommendations with a private company). The difference between the two vignettes of more than one on a five-point scale is quite striking, supporting the argument that respondents' willingness to share data depends on the specific circumstances of the respective data sharing context. In these two extreme cases, it seems that the data type (sensor data vs. biomarkers) has a particularly strong impact on respondents' average willingness ratings. The overall mean across all experimental conditions is 2.81, which is slightly below the midpoint of the rating scale (3.0).

Regression models

Figure 1 plots the estimated coefficients from the various multilevel regression models described in Section 6. Coefficients do not seem to differ in meaningful ways between the various model specifications. Moreover, coefficients remain robust regarding the addition of socio-demographic control variables (Model 5). Table A.2 in the Online Appendix shows full regression model specifications. Moreover, excluding respondents who indicated a low attentiveness value from all models estimated (see Section 6) does not change results substantially (see Tables A.5, A.6, and A.7 in the Online Appendix).

Turning to our hypotheses, we find that the willingness to share biomarkers (the reference category) is the highest among all three data types. It seems that sharing sensor data is by far the least accepted. Thus, we do not find support for H1, which suggests that biomarkers would result in the lowest willingness to share data.

Regarding recipients, H2.1 and H2.2 stated that respondents would be more willing to share data with a public agency compared with a private company and be more willing to share data with a university compared with a public agency. Our results do not support these hypotheses. Respondents' willingness to share data does not seem to depend on the recipient as relevant coefficients do not differ from zero in meaningful ways.

In H3, we expected that respondents would be more willing to share data if the purpose of data sharing provides a public benefit compared to a personal benefit. Our data shows that respondents are, on average, more willing to share data for a public benefit than for personal recommendations. We do not find evidence, however, that there is an interaction between the recipient and the purpose (H4). Private companies seem to be as accepted for public benefit purposes as public agencies and universities. Likewise, universities and public agencies seem to be equally accepted as private companies for private recommendations.

Respondents' medical history seems to influence their willingness to share data. Overall, those who have been diagnosed with more medical conditions seem to be more willing to share their data. Other than hypothesized in H5.1 and H5.2, there is no interaction between respondents' medical history and the purpose of the data transmission, however. Since our hypotheses H5.1 and H5.2 specified effects of respondents' medical history per data type, we show results of linear regression analysis separately for each data type in Figure 2. Our conclusion that there is no interaction of respondents' medical history with the purpose of the data transmission remains unchanged.

Regarding respondents' data sharing history (H6.2), we find that those who have shared data before are more willing to share data (again) (Figure 1). Splitting up our data by data type (Figure 2), we find that those who have shared a specific data type are more willing to share this specific type of data again, providing support for hypothesis

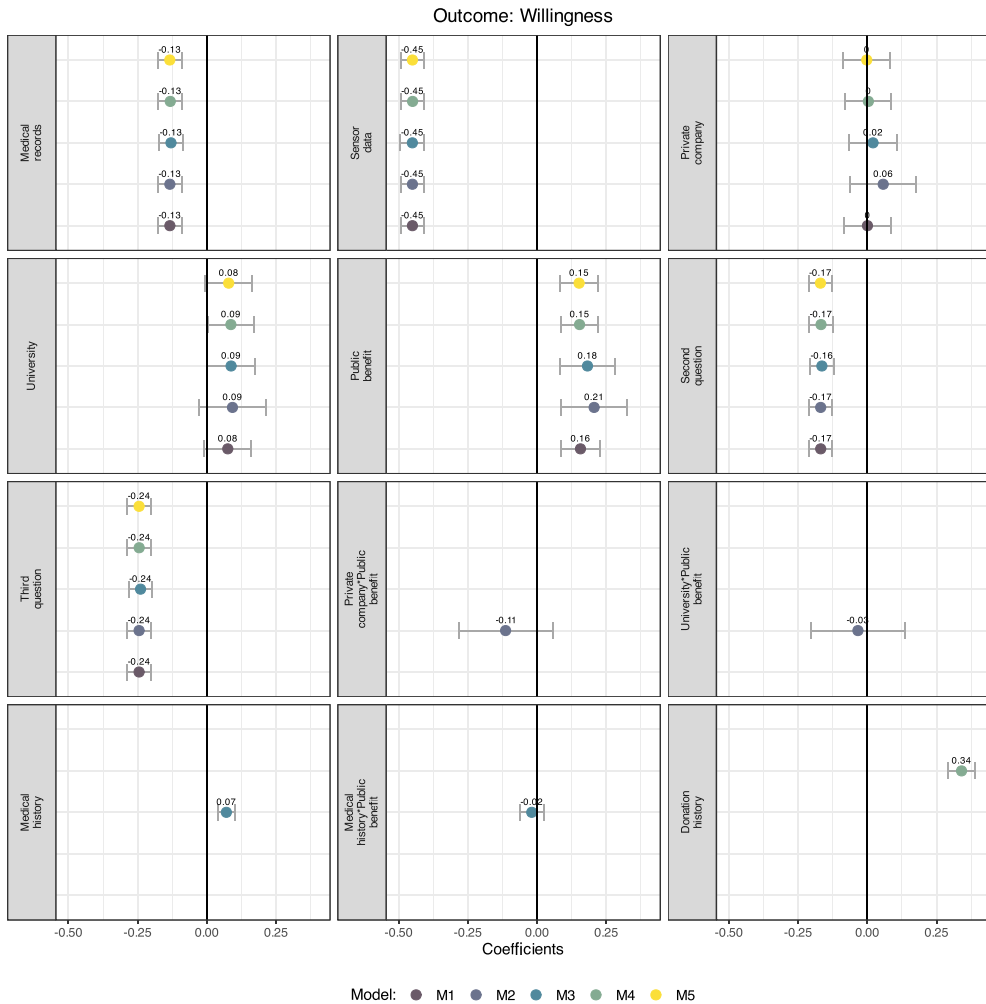


Figure 1. Point estimates and 95% confidence intervals of the multilevel linear regressions predicting willingness to share health data, by predictor and model. Reference categories: Biomarkers (data type), public health agency (recipient), private recommendations (public benefit), first question (question order). M5 contains sociodemographic variables (age, gender, citizenship, education) as control variables. Estimates of socio-demographic characteristics are not shown in the figure. See Table A.2 in the Online Appendix for full model specifications and results.

H6.1. Interestingly, there seems to be an interaction of data sharing history with data type. The coefficient of having shared medical records before is about twice the size of the coefficient of having shared blood before. Having shared sensor data before has an even stronger effect. Those who have shared sensor data before are more than one scale point more willing to share sensor data than those who have not done so previously. This is a large effect on a five-point scale.

Before we turn to the results of the analyses of the open-ended responses, we highlight two noteworthy findings beyond our preregistered hypotheses. First, there seems to be an interaction of data type and recipient (see top two panels in Figure 2). That is,

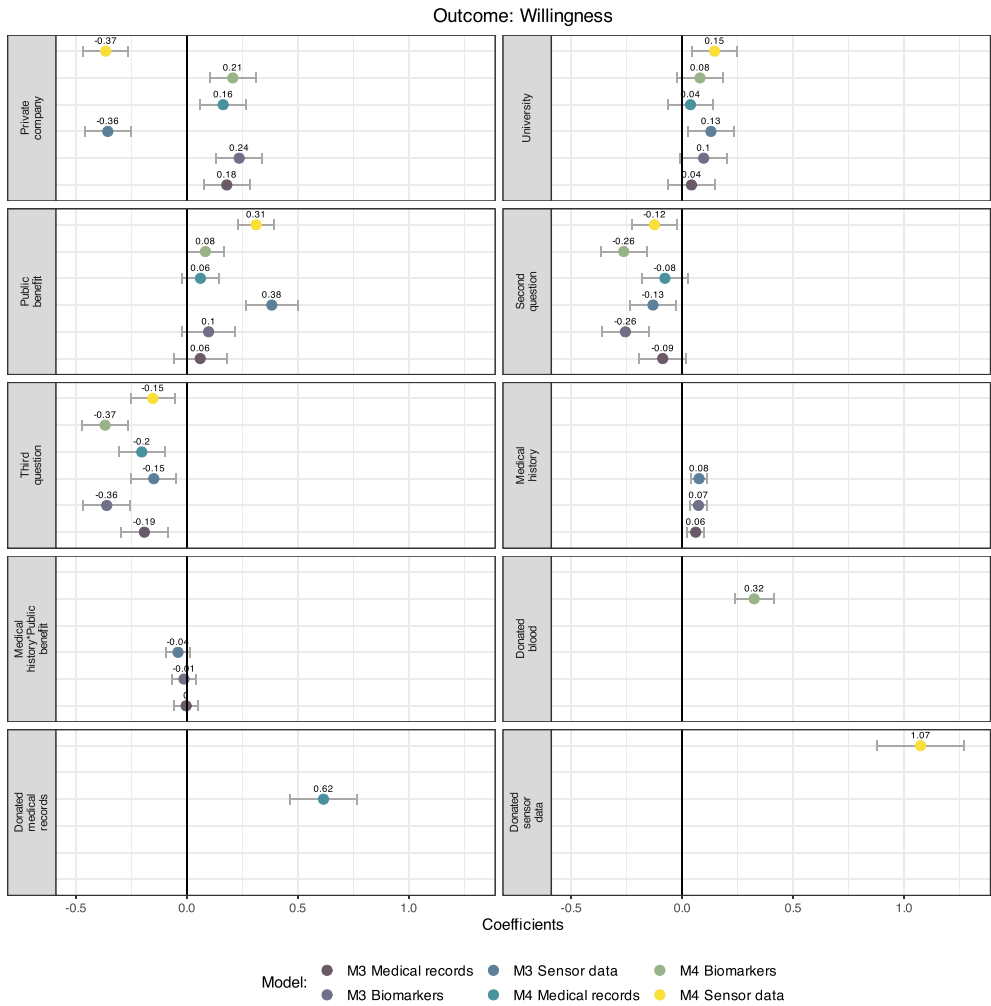


Figure 2. Point estimates and 95% confidence intervals of the linear regressions predicting willingness to share health data, by predictor, model, and data type. Reference categories: Public health agency (recipient), private recommendations (public benefit), first question (question order). See Tables A.3 and A.4 in the Online Appendix for full model specifications and results.

respondents seem to be less willing to share sensor data with a private company than with a public health agency. For the other two data types (biomarkers and medical records), individuals are more willing to share their data with a private company, however. As we observed earlier, respondents seem to be especially concerned regarding the transmission of their *sensor data* in general. This effect seems to be even stronger when the recipient is a private company. Second, there is an interaction of data type and purpose, where sensor data stick out again (Panel ‘Public benefit’ in Figure 2). Respondents seem to be more accepting of sharing sensor data for public benefit than for private recommendations. Thus, the willingness to share sensor data seems to be much more context-dependent than the willingness to share other data types.

Textual responses

The results of the regression analyses do not allow us to understand *why* some respondents were, for example, more hesitant to share one data type than another. The textual responses to the open-ended question help us get a better understanding of respondents' concerns but also their motivations when (un)willing to share their data.

Figure 3 shows the frequency of textual responses, by predicted category and willingness rating. Overall, the 'purpose of collected data' was mentioned most often in the comments. Interestingly, high willingness to share data was strongly associated with mentioning the purpose of collecting the data. Low willingness to share data was associated with mentioning lack of trust, concerns about data misuse, as well as having concerns about data protection and privacy.

Figure 4 shows the frequency of textual responses by category and data type. The impact of the data type that was documented by our regression analyses in Section 7.1 seems to be also present in respondents' motivations and concerns found in the textual responses. Figures A.1, A.2, and A.3 in the Online Appendix split up textual responses by willingness rating and data type. These figures show that respondents more often *lack* trust and have *concerns* about data protection and privacy when asked about sensor data than for the other two data types. At the same time, they seem to be comparably more convinced of the benefits of sharing biomarkers and medical records.

Discussion

Individuals' willingness to share their health data with researchers and government agencies is a prerequisite for evidence-based policy-making. Using a survey experiment,

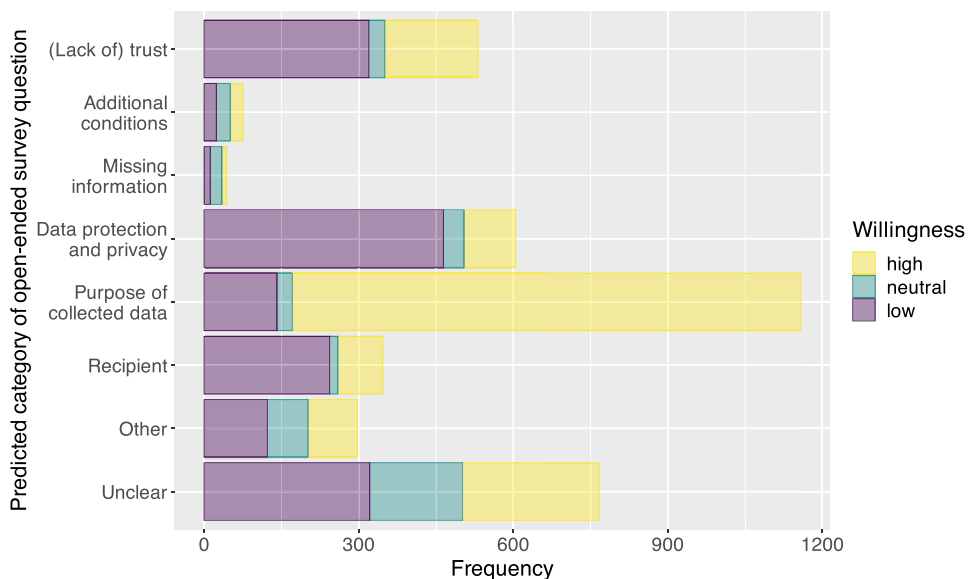


Figure 3. Number of textual responses, by category and willingness rating. *High* corresponds to values '5 – very likely' and '4 – somewhat likely', *neutral* to '3 – neither likely nor unlikely', and *low* to '2 – somewhat unlikely' and '1 – very unlikely'.

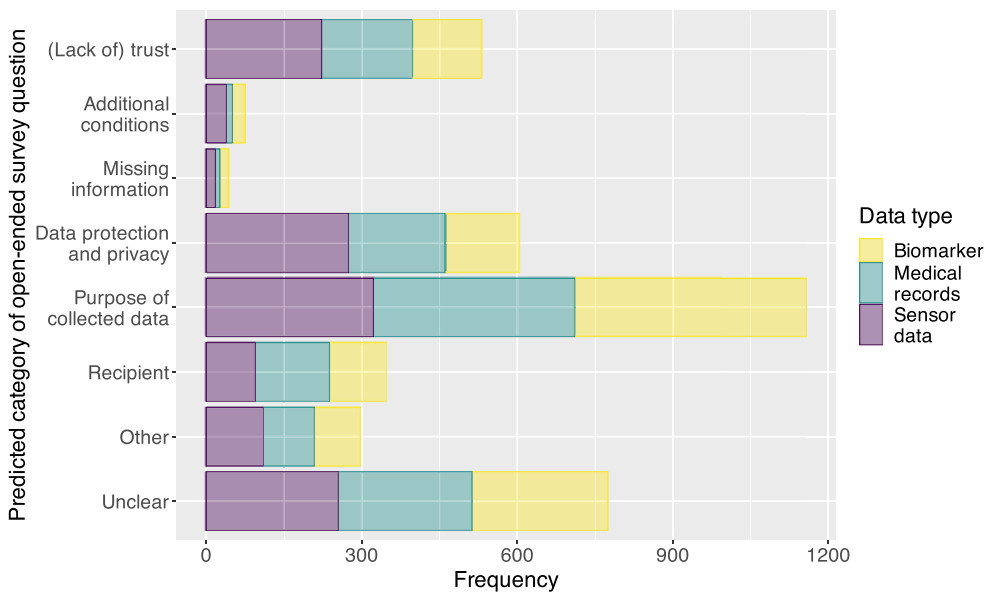


Figure 4. Number of textual responses, by category and data type.

we provide novel empirical evidence for what factors influence the individual data sharing decision-making process. We find that data type and purpose of data use are major drivers of the data sharing decision; individuals show the highest willingness to share biomarkers, followed by medical records and sensor data, and they are more willing to share their data if they will be used for public benefit compared to personal recommendations. Textual answers provided fine-grained insights into respondents' motivations and concerns when being asked to share their data. They indicate that perceived problems with trust and privacy are indeed important factors particularly when individuals decide about sharing sensor data. [Figure 3](#) underlines the argument that respondents do understand how valuable their data are for researchers as many of those who are willing to share their data emphasize the purpose of the data. The textual responses underscore that addressing concerns of individuals about the protection of their data needs to be taken seriously. Crafting privacy policies that are easy to understand is essential to build trust among respondents to share their data for research. Moreover, emphasizing the purpose of shared data may boost respondents' willingness to do so.

These findings are mostly in line with earlier research by Silber et al. (2022) who showed large differences in willingness to share between health data types for cancer research. However, our respondents reported on average lower willingness to share any type of data. This could be attributed to the topic, that is, potentially there might be a higher willingness to share data for cancer research than for research on infectious diseases.

We did not confirm differences in willingness to share by the recipient of the data, which could be due to the context in which the data sharing request was presented. Here, we chose a scenario (containment of an infectious disease) that likely was very salient to respondents, given the pandemic circumstances at the time of data collection.

The exception is sensor data for which we found an interaction with the recipient; individuals reported being less willing to share sensor data with a private company

compared with a public health agency, which is not in line with a previous finding that public agencies tend to be less accepted for health sensor data use (Gerdon et al., 2020). Regarding the norm-based perspective of CI, we may expect that what type of data collection and uses an individual is familiar with could lead to a subjective perception of these uses being 'normal'. However, this does not necessarily mean that individuals desire the current practices. Contextual integrity would look at injunctive norms, that is, what *should* be the case (Nissenbaum, 2010). The use of sensor data by private companies probably is – among our vignettes – one with higher familiarity to many individuals, and at the same time this use is embedded in discourses of excessive commercial data collection (see, e.g., Pew Research Center, 2018). This combination might be the reason for the relatively low willingness ratings. Our results may therefore reflect discontent with currently 'normal' practices.

From a theoretical perspective, the study confirms that the theory of contextual integrity (Nissenbaum, 2010) provides appropriate categories to describe data flows concerning individual willingness to share health data. While our study illustrated that data type and purpose were the most relevant categories regarding willingness to share data, the theory also provides a foundation to go further into explaining these decisions. Specifically, the next step would be to follow the theory of contextual integrity to investigate social norms defining the appropriateness of the different investigated data flows. The textual responses analyzed in the present study provide some insights into potentially relevant norms specific to contextual parameters, such as expectations concerning reasonable purposes to be pursued by different actors. Learning more about these social norms in contrast to actual data sharing decisions might help to explain gaps between *desired* data-related practices and *actual* practices, potentially relating to research on the so-called 'privacy paradox' (e.g., see Dienlin and Trepte (2015)). Such a study could again use a survey with an experimental element to ask for the provision of additional respondent data, while also asking respondents about the appropriateness of different data use practices. Contrasting these decisions and responses along with textual elaborations would be a valuable step towards explaining the aforementioned gaps.

We acknowledge that the results of our study are limited to Germany, and we hope to see researchers assessing the willingness to share health data in other countries with different societal levels of privacy concerns. While we asked for hypothetical willingness to share health data, previous research shows a strong relationship between intent and behavior in vignette experiments (Hainmueller et al., 2015). Future research may also dive deeper into the mechanisms *why* individuals are willing to share data. The literature on survey participation and nonresponse may provide food for thought (see, e.g., Groves & Couper, 2012). Furthermore, questions such as what makes a data recipient trustworthy and how can trust be restored in times of populists' backlash against scientific consensus and scientific institutions require further study (Bellolio, 2022; O'Neill, 2018).

From a practitioner's perspective, the various scenarios can guide researchers in getting an understanding of how each decision in the data collection process can affect the results. As suggested by the empirical results and required by the General Data Protection Regulation for personal data, data collectors need to be clear about which type of data they use for which purpose. The textual responses support, for instance, that respondents are particularly interested in the purpose of data use. The textual responses also contained further valuable information on respondents' concerns and motivations when

being asked to share their data. Such information may be used to guide the design of future studies collecting health data to align with respondents' demands, such that participation rates can be maximized while minimizing participants' concerns about their data at the same time. Thereby, our results are of interest to a broad community of researchers collecting and working with novel data types that may be considered sensitive.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by a postdoctoral fellowship of the German Academic Exchange Service (DAAD) to Ruben L. Bach, by German Research Foundation (DFG) under grant 139943784; "Collaborative Research Center SFB 884 Political Economy of Reforms" (Project A8) to Florian Keusch; and by Volkswagen Foundation under grant "Consequences of Artificial Intelligence for Urban Societies (CAIUS)" to Ruben L. Bach.

Data and code availability statement

Data are deposited at GESIS data archive (<http://dx.doi.org/10.4232/1.14325>) and are publicly available. The hypotheses and the design and sampling plan of this study were preregistered. The preregistration report is available at https://osf.io/pygx5/?view_only=f8e78bf869324c94b942d8a1598fe85d. All code is available at https://github.com/rubac/Willingness_open_ends.

Notes on contributors

Ruben Bach is a research fellow in the Data and Methods Unit of the Mannheim Centre for European Social Research (MZES) at the University of Mannheim, Germany. His research is rooted in computational social science and tackles questions of online news media consumption, AI-guided decision-making, and quality aspects of novel data products in the social sciences.

Henning Silber is a senior researcher at GESIS – Leibniz Institute for the Social Sciences. His research interests include survey methodology, political sociology, and the experimental social sciences.

Frederic Gerdon is a PhD candidate at the University of Mannheim, Germany. His research is concerned with privacy, AI-guided decision-making, and social science research methodology.

Florian Keusch is a Professor of Social Data Science and Methodology in the Department of Sociology at the University of Mannheim, Germany.

Matthias Schonlau is a Professor in the Department of Statistics at the University of Waterloo, Canada. His research interests include survey methodology and natural language processing for open-ended questions.

Jette Schröder is a senior researcher at GESIS – Leibniz Institute for the Social Sciences, Germany. Her research interests include survey methodology, well-being, and family research.

ORCID

Ruben L. Bach  <http://orcid.org/0000-0001-5690-2829>
 Henning Silber  <http://orcid.org/0000-0002-3568-3257>
 Frederic Gerdon  <http://orcid.org/0000-0003-4442-6698>
 Florian Keusch  <http://orcid.org/0000-0003-1002-4092>
 Jette Schröder  <http://orcid.org/0000-0002-1000-5855>

References

- Aitken, M., de St. Jorre, J., Pagliari, C., Jepson, R., & Cunningham-Burley, S. (2016). Public responses to the sharing and linkage of health data for research purposes: A systematic review and thematic synthesis of qualitative studies. *BMC Medical Ethics*, 17(1), 1–24. <https://doi.org/10.1186/s12910-016-0153-x>
- Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. Sage.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Belloio, C. (2022). An inquiry into populism's relation to science. *Politics*, 02633957221109541.
- Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population. *Field Methods*, 27(4), 391–408. <https://doi.org/10.1177/1525822X15574494>
- Breuer, J., Kmetty, Z., Haim, M., & Stier, S. (2023). User-centric approaches for collecting Facebook data in the 'post-api age': Experiences from two studies and recommendations for future research. *Information, Communication & Society*, 26(14), 2649–2668. <https://doi.org/10.1080/1369118X.2022.2097015>
- Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science*, 10(1), 104–115. <https://doi.org/10.1287/orsc.10.1.104>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dienlin, T., & Trepte, S. (2015). Is the privacy paradox a relic of the past? An in-depth analysis of privacy attitudes and privacy behaviors. *European Journal of Social Psychology*, 45(3), 285–297. <https://doi.org/10.1002/ejsp.2049>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, 13(2), 83. <https://doi.org/10.3390/info13020083>
- Gerdon, F., Nissenbaum, H., Bach, R. L., Kreuter, F., & Zins, S. (2020). Individual acceptance of using health data for private and public benefit: Changes during the COVID-19 pandemic. *Harvard Data Science Review*, Special Issue 1.
- Goetzen, A., Dooley, S., & Redmiles, E. M. (2022). Ctrl-shift: How privacy sentiment changed from 2019 to 2021. *Proceedings on Privacy Enhancing Technologies*, 2022(4), 457–485. <https://doi.org/10.56553/popets-2022-0118>
- Gomez Ortega, A., Bourgeois, J., Hutiri, W. T., & Kortuem, G. (2023). Beyond data transactions: A framework for meaningfully informed data donation. *AI & SOCIETY*, 1–18.
- Grande, D., Mitra, N., Iyengar, R., Merchant, R. M., Asch, D. A., Sharma, M., & Cannuscio, C. C. (2022). Consumer willingness to share personal digital information for health-related uses. *JAMA Network Open*, 5(1), e2144787. <https://doi.org/10.1001/jamanetworkopen.2021.44787>
- Groves, R. M., & Couper, M. P. (2012). *Nonresponse in household interview surveys*. John Wiley & Sons.
- Gweon, H., & Schonlau, M. (2022). Automated classification for open-ended questions with BERT. *arXiv preprint arXiv:2209.06178*.

- Habich-Sobiegalla, S., & Kostka, G. (2023). Sharing is caring: Willingness to share personal data through contact tracing apps in China, Germany, and the us. *Information, Communication & Society*, 26(14), 1–28. <https://doi.org/10.1080/1369118X.2022.2113421>
- Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8), 2395–2400. <https://doi.org/10.1073/pnas.1416587112>
- Hutchings, E., Loomes, M., Butow, P., & Boyle, F. M. (2020). A systematic literature review of health consumer attitudes towards secondary use and sharing of health administrative and clinical trial data: A focus on privacy, trust, and transparency. *Systematic Reviews*, 9(1), 235. <https://doi.org/10.1186/s13643-020-01481-9>
- Jörling, M., Eitze, S., Schmid, P., Betsch, C., Allen, J., & Böhm, R. (2023). To disclose or not to disclose? factors related to the willingness to disclose information to a COVID-19 tracing app. *Information, Communication & Society*, 26(10), 1954–1978. <https://doi.org/10.1080/1369118X.2022.2050418>
- Karampela, M., Ouhbi, S., & Isomursu, M. (2019). Connected health user willingness to share personal health data: Questionnaire study. *Journal of Medical Internet Research*, 21(11), e14537. <https://doi.org/10.2196/14537>
- Karwatzki, S., Dytynko, O., Trenz, M., & Veit, D. (2017). Beyond the personalization–privacy paradox: Privacy valuation, transparency features, and service personalization. *Journal of Management Information Systems*, 34(2), 369–400. <https://doi.org/10.1080/07421222.2017.1334467>
- Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., & Kreuter, F. (2019). Willingness to participate in passive mobile data collection. *Public Opinion Quarterly*, 83(S1), 210–235. <https://doi.org/10.1093/poq/nfz007>
- Klingwort, J., & Schnell, R. (2020). Critical limitations of digital epidemiology: Why COVID19 apps are useless. *Survey Research Methods*, 14(2), 95–101.
- Mello, M. M., & Wang, C. J. (2020). Ethics and governance for digital disease surveillance. *Science*, 368(6494), 951–954. <https://doi.org/10.1126/science.abb9045>
- Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
- Nissenbaum, H. (2019). Contextual integrity up and down the data food chain. *Theoretical Inquiries in Law*, 20(1), 221–256. <https://doi.org/10.1515/til-2019-0008>
- O’Neill, O. (2018). Linking trust to trustworthiness. *International Journal of Philosophical Studies*, 26(2), 293–300. <https://doi.org/10.1080/09672559.2018.1454637>
- Pandit, J. A., Radin, J. M., Quer, G., & Topol, E. J. (2022). Smartphone apps in the COVID19 pandemic. *Nature Biotechnology*, 40(7), 1013–1022. <https://doi.org/10.1038/s41587-022-01350-x>
- Pew Research Center. (2018). *Americans’ complicated feelings about social media in an era of privacy concerns*. Retrieved February 20, 2023, from <https://www.pewresearch.org/fact-tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/>
- Rajapakse, T. (2022). *Simple transformers* [Computer software manual]. <https://simpletransformers.ai/>
- R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software manual].
- Silber, H., Gerdon, F., Bach, R., Kern, C., Keusch, F., & Kreuter, F. (2022). A preregistered vignette experiment on determinants of health data sharing behavior: Willingness to donate sensor data, medical records, and biomarkers. *Politics and the Life Sciences*, 41(2), 161–181. <https://doi.org/10.1017/pls.2022.15>
- Skatova, A., & Goulding, J. (2019). Psychology of personal data donation. *PLoS One*, 14(11), e0224240. <https://doi.org/10.1371/journal.pone.0224240>
- Skovgaard, L. L., Wadmann, S., & Hoeyer, K. (2019). A review of attitudes towards the reuse of health data among people in the European union: The primacy of purpose and the common good. *Health Policy*, 123(6), 564–571. <https://doi.org/10.1016/j.healthpol.2019.03.012>

- Sloan, L., Jessop, C., Baghal, T. A., & Williams, M. (2020). Linking survey and twitter data: Informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics*, 15(1–2), 63–76. <https://doi.org/10.1177/1556264619853447>
- Stockdale, J., Cassell, J., & Ford, E. (2019). “Giving something back”: A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. *Wellcome Open Research*, 3, 6. doi:10.12688/wellcomeopenres.13531.2