J Community Genet (2017) 8:173–181
DOI 10.1007/s12687-017-0300-1

ORIGINAL ARTICLE

De-identified genomic data sharing: the research participant perspective

Deborah Goodman[1] · **Catherine O. Johnson**[2] · **Deborah Bowen**[3] · **Megan Smith**[4] · **Lari Wenzel**[5] · **Karen Edwards**[1]

Received: 6 September 2016 / Accepted: 24 March 2017 / Published online: 5 April 2017

Abstract Combining datasets into larger and separate datasets is becoming increasingly common, and personal identifiers are often removed in order to maintain participant anonymity. Views of research participants on the use of de-identified data in large research datasets are important for future projects, such as the Precision Medicine Initiative and Cancer Moonshot Initiative. This quantitative study set in the USA examines participant preferences and evaluates differences by demographics and cancer history. Study participants were recruited from the Northwest Cancer Genetics Registry and included cancer patients, their relatives, and controls. A secure online survey was administered to 450 participants. While the majority participants were not concerned about personal identification when participating in a genetic study using de-identified data, they expressed their concern that researchers protect their privacy and information. Most participants expressed a desire that their data should be available for as many research studies as possible, and in doing so, they would increase their chance of receiving personal health information. About 20% of participants felt that a link should not be maintained between the participant and their de-identified data. Reasons to maintain a link included an ability to return individual health results and an ability to support further research. Knowledge of participants' attitudes regarding the use of data into a research repository and the maintenance of a link to de-identified data is critical to the success of recruitment into future genomic research projects.

Keywords De-identification · Genomic research · Participant views · Data linkage · Precision Medicine

✉ Deborah Goodman
goodmand@uci.ed

[1] Department of Epidemiology, University of California, Irvine Hall, Irvine, CA 92697, USA

[2] Department of Public Health, University of California, Irvine, CA, USA

[3] School of Public Health, University of Washington, Seattle, WA, USA

[4] Center for Statistical Consulting, University of California, Irvine, CA, USA

[5] Department of Public Health, University of California, Irvine, CA, USA

Introduction

The practice of creating large databases has become increasingly common by adding or combining research participants' data into larger repositories. This enables genome-wide association study type analyses to be done with adequate sample sizes. This idea of combining datasets into larger and separate datasets, called data sharing, is often not in the original consent process of the smaller studies. Therefore, in order to maintain participant anonymity, personal identifiers are removed to de-identify the data and a coded link may be kept between a participant's data and his/her identity to allow for possible future clinical updates, longitudinal epidemiologic studies, or the return of individual research results.

Views of research participants regarding the contribution of their de-identified data to these large research datasets are critical to the success of future projects, such as the Precision Medicine Initiative and Cancer Moonshot Initiative. Previous qualitative studies using focus groups (Oliver et al. 2011) and phone interviews (Jamal et al. 2014; Lemke et al. 2010) have yielded somewhat conflicting results. These studies have found that most research participants agree that the benefits of data sharing outweigh the potential risks (Oliver et al. 2011). However, they are concerned that their personal information

or identity may be extracted from their genomic data (Jamal et al. 2014; Lemke et al. 2010) and have misgivings about sharing their data with for-profit companies or other researchers (Lemke et al. 2010). Participant views on the maintenance of a link with de-identified data are not known.

This quantitative study measures the preferences regarding the addition of data to a research repository and views about the oversight and sharing of de-identified genomic data in a group of research participants from the Participants Issues and Expectations Project (PIP) (Collins and Varmus 2015). In addition, we will investigate whether these views differ by factors including participant demographics, history of cancer, or participant status (cases vs controls vs relatives).

Materials and methods

The source population for the Participant Issues and Expectations Project (PIP) was the set of individuals (n = 3909) enrolled in the Northwest Cancer Genetics Registry (NWCGR) in 1999, part of a national network designed to specialize in the study of inherited predisposition to cancer. The NWCGR has been described elsewhere (Condit et al. 2015). This included people with cancer recruited from Western Washington (n = 2027), first-degree relatives of cases (n = 451), controls who were recruited from a random population sample from W. Washington (n = 527), and people who self-referred in response to community awareness efforts and included both people with and without cancer (n = 904 total; 340 with cancer). Self-referrals with cancer were grouped with cases, and those without were grouped with the controls. Letters, including informed consent, were sent by US mail in 2013 inviting the enrolled individuals to take the online, confidential survey. Up to three invitations were sent to participants at approximately two-week intervals.

Development of the PIP survey

The purpose of the survey was to document the range and frequency of occurrence of concerns and expectations regarding participating in human research studies, including genomic and family studies. Detailed methods for this study, including the survey instrument, have been published previously (Goodman et al. 2016). Briefly, the survey instrument had a total of 22 questions, divided into six general topic areas: decision to participate in research; relationship between researchers and participants; re-consent and broad consent; return of results; use and security of de-identified data; and family communication of health issues. The types of response categories included either yes/no/not sure options, Likert-scales (e.g., 5-point scales rating agreement, likelihood, or importance of the statement with a sixth "don't know" or "it depends" option), or categorical responses. We report here the results of the items related to the use of a research respository and security of de-identified data.

The survey was confidential, not anonymous, using a unique identification number for each participant, and participants were free to skip any questions that they did not wish to answer. Each participant was provided an individual URL to access the secure survey instrument. Links to each individual survey and the participant's NWCGR data were retained to reduce the length of the survey and to allow us to utilize existing demographic and other participant information.

All study procedures were approved by the University of Washington's Human Subjects Division, and also by the University of California, Irvine Institutional Review Board. All participants provided informed consent prior to participation.

Statistical analysis

Responses to all questions were summarized using frequency distributions and compared by participant type (i.e., cases, controls, relatives). For ordered data, ordinal logistic regression was used to assess the association between participant characteristics and attitudes regarding the addition of their personal data into a bio-repository or the maintenance of a link to their de-identified data. In this approach, several cumulative logits are modeled using all possible cut points of the dependent variable, and a single summary odds ratio (OR) and 95% confidence interval describing the relationship between the dependent and independent variable is estimated. Comparisons were adjusted for age, gender, and education as appropriate.

For non-ordered responses, multinomial logistic regression was used to test for differences among groups of participants in their preferences pertaining to the person or agency responsible for maintaining a link to de-identified data and reasons for maintaining a link to de-identified data (nominal-dependent variables). A relative risk ratio (RRR) comparing the likelihood of each answer choice to that of a reference choice, along with a corresponding 95% confidence interval and p value are obtained.

Participant characteristics examined for all questions included status (cases vs controls vs relatives), age (years), gender (male vs female), marital status (not currently partnered vs partnered), diagnosis of cancer at baseline (yes vs no), diagnosis of cancer at follow-up (yes vs no), and stage of cancer for cases reported at either baseline or follow-up (stages 3 and 4 vs stages 1 and 2).

R version 3.3.0 was used for all analyses; the polr function (MASS package) and mlogit function (mlogit package) were used to implement the regression models (R Core Team 2015). A p value ≤ 0.05 was considered statistically significant for all tests. Sample sizes varied slightly by question since participants were allowed to skip any question they did not wish to answer.

Results

As shown in Table 1, about half of the 450 research participants were cases (n = 228), one third were controls (n = 155), and the remainder were relatives (n = 67), compared to the original registry distribution of 50.7% cases, 27.3% controls, and 22% relatives. Overall, the average age was 63.6 years, and the majority of participants were white (94.7%) and well educated, with over 60% having a college degree. Among those participants with cancer at the initial enrollment into the parent study (baseline), melanoma was the most frequent cancer type (29.5%), followed by thyroid cancer (18.3%), and breast cancer (15.5%). Thirty-five research participants without cancer at enrollment into a parent study reported a cancer at the time of this survey (follow-up).

Addition of data to a research repository

When asked about factors related to the decision to make personal data available to a research repository, the majority of participants indicated that it is very important or somewhat important that their samples and information be available to as many research studies as possible, and that by sharing their data with a research repository, they would increase the chance that they would learn health information about themselves (Table 2). Younger participants were significantly more likely to feel these were important factors in their decision to add their data to a research repository compared to older participants (OR = 0.97; 95% C.I. = 0.95–0.98 and OR = 0.97; 95% C.I. = 0.96–0.99, respectively). Compared to those without cancer, participants with a history of cancer at baseline were significantly more likely to want their information and samples available to as many research studies as possible (OR = 1.56; 95% C.I. = 1.04, 2.33) (data not shown). Equally as important to most participants was that their privacy and information be protected, and the importance of this increased slightly with increasing age (OR = 1.05; 95% C.I. = 1.03–1.07). Stratification by gender, marital status, stage of cancer, or subject type (case, control, or relative) did not materially alter these findings (data not shown).

Maintaining a link to de-identified data

While about one fifth of respondents felt that there should not be a link between them and their de-identified data, most felt that if such a link was to be maintained, the original researcher would be responsible for its maintenance (Table 3). When asked about the importance of maintaining a link, respondents were divided between an ability to return individual health results and an ability to support further research. The majority of participants agreed that the researchers who use the data share the ethical responsibilities with the original researchers, and results generated from the data should be reported to the repository to allow the return of results. More than half the participants were either not very concerned or not at all concerned about being personally identified when participating in a genetic study using de-identified data, and most participants felt that having their financial identity stolen would be worse

Table 1 Demographics of the research participant (PIP) group

	Total (n = 450)	Cases (n = 228)	Controls (n = 155)	Relatives (n = 67)
Mean (SD) age (years)	63.6 (11.8)	64.3 (11.4)	64.0 (11.5)	60.5 (13.6)
Women	292 (64.9%)	145 (63.6%)	110 (71.0%)	37 (55.2%)
Race				
Asian/Pacific Islander	7 (15.6%)	4 (1.8%)	2 (1.3%)	1 (1.5%)
Black	4 (0.9%)	2 (0.9%)	2 (1.3%)	0
Multi-racial/other	16 (3.6%)	8 (3.5%)	4 (2.6%)	4 (6.0%)
White	423 (94.7%)	214 (93.9%)	147 (94.8%)	62 (92.5%)
Education				
High school or less	40 (8.9%)	19 (8.3%)	13 (8.4%)	8 (11.9%)
Some college	107 (23.8%)	57 (25.0%)	37 (23.9%)	13 (19.4%)
Bachelors degree	276 (61.3%)	126 (55.3%)	105 (67.7%)	45 (67.2%)
Unknown	27 (6.0%)	26 (11.4%)	0	1 (1.5%)
Marital status				
Married/living together	343 (76.2%)	180 (78.9%)	110 (71.0%)	53 (79.1%)
Single	32 (7.1%)	14 (6.1%)	13 (8.4%)	5 (7.5%)
Divorced/separated	44 (9.8%)	19 (8.3%)	20 (12.9%)	5 (7.5%)
Widowed	22 (4.9%)	6 (2.6%)	12 (7.7%)	4 (6.0%)
Unknown	9 (2.0%)	9 (3.9%)	0	0

Table 2 Comparison of participant beliefs regarding the addition of data to a research repository by subject type (cases, controls, or relatives)

	Cases (%)	Controls (%)	Relatives (%)
When making a decision about adding your data to a research repository, how important is it to you that your samples and information are available to as many research studies as possible?			
Very important	55.1	64.7	65.6
Somewhat important	34.2	26.7	21.9
Not very important	8.4	7.3	6.2
Not at all important	2.2	1.3	6.2
When making a decision about adding your data to a research repository, how important is it to you to increase the chance you will learn health information about yourself?			
Very important	40.4	41.3	48.4
Somewhat important	41.3	35.3	26.6
Not very important	15.1	18.7	17.2
Not at all important	3.1	4.7	6.2
When making a decision about adding your data to a research repository, how important is it to you to have a system in place to protect your privacy and information?			
Very important	73.8	69.3	71.9
Somewhat important	18.7	19.3	15.6
Not very important	4.4	8.7	9.4
Not at all important	1.8	2.0	1.6

than having their genetic information stolen. As shown in Table 4, a comparison by participant characteristics found that controls were significantly more likely than cancer cases to agree that all researchers who use data from a research repository have an ethical obligation to return results to the research participants (OR = 1.86; 95% C.I. = 1.08–3.20). In addition, stage of cancer was significantly associated with the belief that the return of results is the ethical responsibility of all investigators using research repository data (OR = 4.05; 95% C.I. = 1.18–13.91). These results did not differ by age, gender, marital status, or history of cancer.

As shown in Table 5, older participants were less likely to believe that no link should be maintained between their identity and their de-identified data (RRR = 0.97; 95% C.I. = 0.94–0.99). In addition, older participants and controls (vs cases) were more likely to endorse the importance of maintaining a link to de-identified data to allow participation in future studies rather than being able to have their personal health information returned to them (RRR = 1.02; 95% C.I. = 0.94–1.00, RRR = 1.82; 95% C.I = 1.13–2.91, respectively). In contrast, participants with cancer at baseline were significantly less likely to prioritize the maintenance of a link to de-identified data to allow participation in future studies rather than obtaining personal health information (RRR = 0.58; 95% C.I. = 0.38, 0.90) (Table 6). No significant differences were seen by gender, stage of cancer, or marital status.

Sharing research data

Research participants were most comfortable sharing their research data with other researchers at the same university and non-profit organizations such as The American Cancer Society, followed by researchers at other universities in the USA (Table 7). Participants were least comfortable sharing their data with any researcher who requests the information and for-profit, private organizations such as pharmaceutical companies. Older age was significantly associated with a decreased willingness to share research data with non-profit organizations and any researcher who requests the information (OR = 0.96; 95% C.I. = 0.94, 0.99 and OR = 0.98; 95% C.I. = 0.97–1.00, respectively). No significant differences in willingness to share research data were seen when responses were stratified by participant type (i.e., cases, controls, relatives), gender, marital status, or history, or stage of cancer (data not shown).

Discussion

This study found that over half of research participants were not concerned about the risk of personal identification from the addition of their de-identified genetic data to a shared database, and most participants felt that having their financial identify stolen would be worse than having their genetic information stolen. This is consistent with other studies that have found that while participants expressed concern about maintaining privacy if their personal information was added to a bio-repository, and even felt that it was inevitable their confidentiality would be breeched, the majority expressed that the benefits of pooled data outweigh the potential risks to their privacy (Oliver et al. 2011; Jamal et al. 2014; Lemke et al. 2010; McCarty et al. 2011). Consistent with others (Trinidad et al. 2010), this population expressed that while it was important that their privacy and information be protected, having their information available to many research studies is important to them. It has been shown that participants underestimate the extent to which bio-banked data is shared (Kaufman et al. 2009; McCarty et al. 2007; Ormond et al. 2009; Okun and Schultz 2003) and it is possible that this contributes to their favorable attitude toward using their pooled data.

This study demonstrated that the decision to make personal data available to a research repository is influenced by the participants' desire to acquire health information about themselves. This study did not specifically ask whether "acquiring health information" meant that research participants expected the return of their individual research results or a summation of research findings at the conclusion of the study. However, because most participants agreed that a link to de-identified data should be maintained to allow the return of their individual health results, we can extrapolate that they in fact expect

Table 3 Comparison of participant beliefs regarding a link to their de-identified data by subject type (cases, controls, or relatives)

	Cases (%)	Controls (%)	Relatives (%)
If a link to your de-identified data is maintained somewhere, who should be responsible for maintaining that link?			
The original researcher whose study I participated in	42.3	40.9	53.0
A federal coordinating center (e.g., NIH)	20.3	21.5	13.6
An independent, non-governmental, and non-profit coordinating center	16.2	17.4	16.7
There should be no link between me and my de-identified data in a research repository	21.2	20.1	16.7
In your opinion, if a link is maintained, what is the most important reason to maintain that link?			
The ability for personal health results or health information to be returned to me	43.2	30.9	40.9
The ability to support future research by being invited to participate in a new research study that will need additional information from me	45	56.6	47
There should never be a link	11.7	12.5	12.1
If a link is maintained, the researchers who use the data from the research repository have the same ethical obligations to return results to me as the original investigators			
Strongly agree/agree	83.5	74.3	83.3
Neutral	10.7	15.5	10.6
Strongly disagree/disagree	5.8	10.1	6.1
If a link is maintained, a system should be in place that allows researchers who use this data to report results to the research repository, so that decisions about returning results can be made			
Strongly agree/agree	83.9	81.0	81.0
Neutral	12.4	14.3	15.9
Strongly disagree/disagree	3.7	4.8	3.2
How concerned are you that you would be personally identified (by someone other than the researchers) if you participated in a study involving de-identified genetic data?			
Very concerned	17.2	11.8	11.9
Somewhat concerned	30.0	26.1	37.3
Not very concerned	36.6	43.8	31.3
Not at all concerned	16.3	18.3	19.4
Which of these statements do you agree with the most?			
Having my financial identity stolen would be worse than having my genetic information stolen	69.2	65.4	53.7
Having my genetic information stolen would be worse than having my financial identity stolen	3.1	3.3	3.0
Having my genetic information stolen and having my financial identity stolen would be equally bad	36.7	31.4	43.3

the return of their individual genetic results. Currently, the de-identification of data limits the ability of researchers to return individual results to participants (Budimir et al. 2011), and which, if any, individual genetics results should be returned to research participants is highly debated (Budimir et al. 2011). It is important that researchers understand that the return of individual results is an important motivation for data sharing among participants and research participants must be properly educated about the likelihood that personal results will be returned. It is also important to consider whether the responsibility for the return of results lies with the researcher who has analyzed the data or with the registry that has provided the data. In the cancer registry from which this study population is drawn, individual results were not returned to participants because the testing was not conducted in a Clinical Laboratory Improvements Amendments (CLIA)-approved laboratory.

While others have shown an increased willingness to share personal data with increased age (Trinidad et al. 2010), this study found a direct association between age and the importance of protecting privacy and information when deciding to allow data for a research repository. It is possible that our result is due to the narrow age range in our study population. Most participants in this study agreed that a link to de-identified data should be maintained to support further research. Consistent with greater privacy concerns, there was a

Table 4 Ordinal logistic regression results comparing participant views regarding responsibility of maintaining link to de-identified data by subject type, gender, and history and stage of cancer

	Odds ratio	95% C.I.	*p* value
If a link is maintained, the researchers who use the data from the research repository have the same ethical obligations to return results to me as the original investigators			
Subject type			
Control vs case	**1.86**	**1.08, 3.20**	**0.025**
Relative vs case	1.31	0.61, 2.83	0.492
Gender	0.60	0.34, 1.05	0.072
Cancer at baseline	**0.53**	**0.32, 0.88**	**0.014**
Cancer at follow-up	0.64	0.38, 1.06	0.084
Stage of cancer at follow-up	**4.05**	**1.18, 13.91**	**0.026**
Age	1.00	0.98, 1.02	0.747
Marital status	0.69	0.37, 1.31	0.256
If a link is maintained, a system should be in place that allows researchers who use this data to report results to the research repository, so that decisions about returning results can be made			
Subject type			
Control vs case	1.10	0.61, 1.97	0.56
Relative vs case	1.48	0.70, 3.15	0.308
Gender	0.71	0.40, 1.26	0.242
Cancer at baseline	0.93	0.54, 1.58	0.775
Cancer at follow-up	0.80	0.47, 1.38	0.430
Stage of cancer at follow-up	1.14	0.25, 5.23	0.868
Age	1.01	0.99, 1.03	0.377
Marital status	1.02	0.54, 1.93	0.954

Figures in bold indicate statistical significance

Table 5 Multinomal logistic regression results comparing participant views of reasons for maintaining link to de-identified data and the responsible entity by subject type, gender, age, marital status and history and stage of cancer

If a link to your de-identified data is maintained somewhere, who should be responsible for maintaining that link?

	Comparison groups*	RRR	95% C.I.	*p* value
Subject type				
Control vs case	2 vs 1	1.02	0.57, 1.82	0.950
	3 vs 1	1.07	0.57, 1.98	0.838
	4 vs 1	0.89	0.49, 1.62	0.710
Relative vs case	2 vs 1	0.44	0.18, 1.07	0.070
	3 vs 1	0.72	0.31, 1.66	0.445
	4 vs 1	0.52	0.22, 1.18	0.118
Gender	2 vs 1	0.88	0.49, 1.56	0.653
	3 vs 1	0.87	0.48, 1.60	0.661
	4 vs 1	0.98	0.55, 1.74	0.938
Cancer at baseline	2 vs 1	1.05	0.61, 1.80	0.860
	3 vs 1	0.80	0.46, 1.42	0.448
	4 vs 1	1.17	0.67, 2.02	0.580
Cancer at survey	2 vs 1	1.09	0.63, 1.91	0.749
	3 vs 1	0.78	0.44, 1.39	0.398
	4 vs 1	1.06	0.60, 1.85	0.843
Stage of cancer	2 vs 1	0.96	0.20, 4.55	0.955
	3 vs 1	0.58	0.10, 3.24	0.530
	4 vs 1	1.27	0.33, 4.92	0.731
Age (years)	2 vs 1	0.98	0.96, 1.01	0.152
	3 vs 1	0.98	0.96, 1.00	0.098
	4 vs 1	**0.97**	**0.94, 0.99**	**0.005**
Marital status	2 vs 1	1.42	0.66, 3.04	0.374
	3 vs 1	1.09	0.46, 2.58	0.836
	4 vs 1	0.68	0.26, 1.74	0.416

Figures in bold indicate statistical significance

*Groups compared: 1 = The original researcher whose study I participated in

2 = A federal coordinating center

3 = An independent, non-governmental, and non-profit coordinating center

4 = There should be no link between me and my de-identified data in a research repository

small but significant inverse association between age and the belief that a link should never be maintained. However, increasing age was also associated with the belief that a link should be maintained to support further research. Others have shown that social volunteer motivation increases with age (Okun and Schultz 2003) and this may explain the increased altruistic view related to the necessity of data linkage to allow for future studies. Inclusion of a wide age range of study participants is necessary to improve the generalizability of research findings. An awareness of differences in willingness to share data related to participant age is important to recognize. Data-sharing plans and recruitment efforts should be tailored by age to provide appropriate education about the protection of privacy and information while recognizing an altruistic motivation for research participation.

While these participants favored data sharing in order to make as much information as possible available for research, this altruistic motivation was significantly greater among those with a history of cancer. It may be that participants previously affected by cancer feel a more urgent need for future research compared to an unaffected population. Although others have shown that altruism is not a motivation among cancer patients who participate in clinical trials (Daugherty et al. 1995), we found that participants with a history of cancer were significantly more likely to endorse maintaining a link to data in order to support future research rather than to gain personal health results. It is likely that this difference is explained by that fact that our study participants were enrolled from a cancer registry and were not participating in a clinical trial for cancer treatment.

Others have shown that it is important to research participants that they feel they have control over which researchers have access and use their data (Jamal et al. 2014; Lemke et al. 2010). This study found that other than the original researcher,

Table 6 Multinomial logistic regression results comparing participant views of reasons for maintaining link to de-identified data and the responsible entity by subject type, gender, age, marital status and history, and stage of cancer

In your opinion, if a link is maintained, what is the most important reason to maintain that link?

	Comparison groups*	RRR	95% C.I.	*p* value
Subject type				
Control vs case	2 vs 1	**1.82**	**1.13, 2.91**	**0.013**
	3 vs 1	1.45	0.70, 2.98	0.314
Relative vs case	2 vs 1	1.23	0.65, 2.31	0.527
	3 vs 1	1.14	0.44, 2.93	0.792
Gender	2 vs 1	1.17	0.74, 1.85	0.495
	3 vs 1	1.32	0.66, 2.64	0.438
	4 vs 1	0.98	0.55, 1.74	0.938
Cancer at baseline	2 vs 1	**0.58**	**0.38, 0.90**	**0.014**
	3 vs 1	0.34	0.04, 3.31	0.354
Cancer at survey	2 vs 1	0.84	0.27, 2.56	0.759
	3 vs 1	0.34	0.04, 3.31	0.354
Stage of cancer	2 vs 1	0.84	0.27, 2.56	0.759
	3 vs 1	0.34	0.04, 3.31	0.354
Age (years)	2 vs 1	**1.02**	**1.00, 1.04**	**0.057**
	3 vs 1	**0.97**	**0.94, 1.00**	**0.053**
Marital status	2 vs 1	1.04	0.55, 1.97	0.900
	3 vs 1	1.22	0.43, 3.50	0.711

Figures in bold indicate statistical significance

*Groups compared: 1 = The ability for personal health results or health information to be returned to me

2 = The ability to support future research by being invited to participate in a new research study which will need additional information from me

3 = There should never be a link

Table 7 Comparison of responses by subject type (cases, controls, or relatives) to the question: Which groups would you be comfortable sharing your research data?

	Cases (%)	Controls (%)	Relatives (%)
Other researchers at the same university			
Yes, please share	93.0	89.5	91.0
No, don't share	1.3	3.3	1.5
Not sure	5.7	7.2	7.5
Researchers at other universities in the USA			
Yes, please share	84.1	76.2	86.6
No, don't share	2.7	6.8	1.5
Not sure	13.3	17.0	11.9
US government researchers (for example, researchers at the National Institutes of Health)			
Yes, please share	74.0	64.2	76.1
No, don't share	9.0	13.9	13.4
Not sure	17.0	21.9	10.4
Researchers at universities in other counties			
Yes, please share	47.3	48.0	50.7
No, don't share	22.8	27.3	23.9
Not sure	29.9	24.7	25.4
Non-profit organizations (for example, The American Cancer Society)			
Yes, please share	90.2	84.9	91.0
No, don't share	3.1	5.3	1.5
Not sure	6.7	9.9	7.5
For-profit, private organizations (for example, pharmaceutical companies)			
Yes, please share	24.6	18.8	16.4
No, don't share	46.9	51.7	50.7
Not sure	28.5	29.5	32.8
Any researcher who requests the information			
Yes, please share	12.8	12.1	14.9
No, don't share	54.9	63.8	50.7
Not sure	32.3	24.2	34.3

participants were most comfortable sharing their pooled data and information with other researchers at the same university and non-profit organizations, followed by researchers at other universities in the USA. Also consistent with others (Hoeyer 2012; Steinsbekk et al. 2013; Caulfield et al. 2014), research participants were least comfortable having their information shared with for-profit, private organizations. Possible reasons for this varying level of comfort with data sharing may be due to a mismatch participants feel between their altruistic motivations and the financial incentives of the private organizations (Trinidad et al. 2010), they feel that for-profit companies are not interested in the public good and they question if the for-profit company may use their data in a morally problematic manner (Lemke et al. 2010). It has also been noted that participants feel that the involvement of for-profit industries with data sharing may limit the return of individual results and decrease public available of the health benefits resulting from private research (Caulfield et al. 2014). In this population, divorced or separated participants were significantly more likely to view the need for a federal coordinating center, rather than the original researcher, to maintain the link to their de-identified data. Views concerning the sensitivity of genetic data have been previously shown to vary by marital status: divorced, widowed, separated, or never married participants are significantly more likely to endorse restricted rather than public data access (Shabani et al. 2014).

It is important to improve research participant trust by increased education regarding the scientific benefits of sharing data with commercial recipients, the safeguards commercial users have in place to protect the participants' privacy, as well as the contribution of some industry leaders to major scientific advances using shared data, including bringing therapies and products to market (Haddow et al. 2007). Future research is needed to understand the aspects of sharing data with for-profit companies that impacts participant trust.

This study has several limitations. While perceptions about sharing genetic data have been shown to differ by race and education (McGuire et al. 2011), this study population was limited to a highly educated and mostly white group of older adults, and we were unable to evaluate these possible variations in beliefs. Although 35 participants in the control group reported cancer at baseline and 35 participants without cancer at baseline reported cancer at the time of the survey, we still found significant differences between cases and controls in attitude regarding maintaining a link to de-identified data. While some differences were seen by stage of cancer, these findings were based on small numbers and further conclusions are not possible. We were unable to evaluate preferences related to the use of de-identified genomic data by cancer site because of small numbers for most cancer types. Finally, because this study includes participants from a long-standing research population, selection bias may positively influence their attitudes about data sharing and de-identified data.

The numbers of research repositories are steadily increasing and the scientific benefits of open data sharing are enormous. This, however, must be weighed against the ethical concerns posed by the use of shared data. Knowledge of research participants' attitudes regarding the addition of their data into a research repository and the maintenance of a link to their de-identified data are critical to the success of recruitment into future genomic research studies. Participant beliefs regarding data sharing and de-identification vary by participant characteristics, and these differences are critical to consider for successful recruitment for future research. Additional studies are also needed to compare attitudes regarding the use of large-scale, shared databases and the associated confidentiality concerns among all stakeholders, including the research participants, investigators, and funding agencies in order to achieve common ground and successfully advance genomic research.

Acknowledgments The authors wish to thank the individuals enrolled in the NWCGR for their ongoing participation in and contribution to cancer research. They also acknowledge and thank Lesley Pfeiffer, Anne Renz, Joan Scott, and David Kaufmann for their work contributing to the earlier stages of this project.

Compliance with ethical standards

Funding This research was supported by NIH grant no. R01CA149051 to Karen Edwards (PI), "Identification of Issues and Expectations of Subjects Participating in Genetic Studies of Cancer".

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All study procedures were approved by the University of Washington's Human Subjects Division, and also by the University of California, Irvine Institutional Review Board.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Budimir D, Polasek O, Marusic A et al (2011) Ethical aspects of human biobanks: a systematic review. Croat Med 52:262–279

Caulfield T, Burningham S, Joly Y et al (2014) A review of the key issues associated with the commercialization of biobanks. J Law Biosci 1(1):94–110

Collins FS, Varmus H (2015) A new initiative on Precision Medicine. N Engl J Med 372:793–795

Condit CM, Korngiebel DM, Pfeifer M, Renz AD, Bowen DJ, Kaufman D, Mercer Kollar LM, Edwards K (2015) What should be the character of the researcher-participant relationship? Views of participants in a longstanding cancer genetic registry. IRB: Ethics & Human Research 37(4):1–10

Daugherty C, Ratain MJ, Grochowski E, Stocking C, Kodish E et al (1995) Perceptions of cancer patients and their physicians involved in phase I trials. J Clin Oncol 13(5):1062–1072

Goodman D, Johnson C, Wenzel L, Bowen D, Condit C, Edwards KL (2016) Consent issues in genetic research: views of research participants. Public Health Genomics 19(4):220–228

Haddow G, Laurie G, Cunningham-Burley S, Hunter KG (2007) Tackling community concerns about commercialisation and genetic research: a modest interdisciplinary proposal. Soc Sci Med 64(2): 272–282

Hoeyer K (2012) Trading in cold blood? In: Dabrock P, Taupitz J, Ried J (eds) Trust in biobanking: dealing with ethical, legal and social issues in an emerging field, vol 33. Springer Berlin Heidelberg, Berlin, pp 21–41

Jamal L, Sapp JC, Lewis K et al (2014) Research participants' attitudes towards the confidentiality of genomic sequence information. Eur J Hum Genet 22(8):964–968

Kaufman DJ, Murphy-Bollinger JM, Scott J, Hudson KL (2009) Public opinion about the importance of privacy in biobank research. Am J Hum Genet 85(5):643–654

Lemke AA, Wolf WA, Herbert-Beirne J, Smith ME (2010) Public and biobank participant attitudes toward genetic research participation and data sharing. Public Health Genomics 13(6):368–377

McCarty CA, Garber A, Reeser JC, Fost NC (2011) Study newsletters, community and ethics advisory boards, and focus group discussions provide ongoing feedback for a large biobank. Am J Med Genet A 155A(4):737–741

McCarty CA, Nair A, Austin DM, Giampietro PF (2007) Informed consent and subject motivation to participate in a large, population-based genomics study: the Marshfield Clinic Personalized Medicine Research Project. Community Genet 10(1):2–9

McGuire AL, Oliver JM, Slashinski MJ et al (2011) To share or not to share: a randomized trial of consent for data sharing in genome research. Genet Med 13(11):948–955

Okun MA, Schultz A (2003) Age and motives for volunteering: testing hypotheses derived from socioemotional selectivity theory. Psychol Aging 18(2):231–239

Oliver JM, Slashinski MJ, Wang T et al (2011) Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. Public Health Genomics 15(2):106–114

Ormond KE, Cirino AL, Helenowski IB, Chisholm RL, Wolf WA (2009) Assessing the understanding of biobank participants. Am J Med Genet A 142A(2):188–198

R Core Team (2015). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/; MASS package (code for ordinal logistic regression: Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0–387–95457-0.

Shabani M, Bezuidenhout L, Borry P (2014) Attitudes of research participants and the general public towards genomic data sharing: a systematic literature review. Expert Rev Mol Diagn 14(8):1053–1065

Steinsbekk KS, Ursin LO, Skolbekken JA, Solberg B (2013) We're not in it for the money lay peoples moral intuitions on commercial use of their biobank. Med Health Care Philos 16(2):151–162

Trinidad SB, Fullerton SM, Bares JM et al (2010) Genomic research and wide data sharing: views of prospective participants. Genet Med 12(8):486–495