Public Preferences regarding Data Linkage for Research: A Discrete Choice Experiment comparing Scotland and Sweden

Mary P Tully (mary.tully@manchester.ac.uk)
https://orcid.org/0000-0003-2100-3983
Cecilia Bernsten
Uppsala Universitet
Mhairi Aitken
Newcastle University Business School
Caroline Mary Vass
The University of Manchester https://orcid.org/0000-0002-6385-2812

Research article

Keywords: Preferences, discrete choice experiment, linked data

Posted Date: December 10th, 2019

DOI: https://doi.org/10.21203/rs.2.11635/v2

Version of Record: A version of this preprint was published on June 16th, 2020. See the published version at https://doi.org/10.1186/s12911-020-01139-5.

Abstract

Objective: There are increasing examples of linking data on healthcare resource use and patient outcomes from different sectors of health and social care systems. Linked data are generally anonymised, meaning in most jurisdictions there are no legal restrictions to their use in research conducted by public or private organisations. Secondary use of anonymised linked data is contentious in some jurisdictions but other jurisdictions are known for their use of linked data. The publics' perceptions of the acceptability of using linked data is likely to depend on a number of factors. This study aimed to quantify the preferences of the public to understand the factors that affected views about types of linked data and its use in two jurisdictions. Method: An online discrete choice experiment (DCE) previously conducted in Scotland was adapted and replicated in Sweden. The DCE was designed, comprising five attributes, to elicit the preferences from a representative sample of the public in both jurisdictions. The five attributes (number of levels) were: type of researcher using linked data (four); type of data being linked (four); purpose of the research (three); use of profit from using linked data (four); who oversees the research (four). Each DCE contained 6 choice-sets asking respondents to select their preferred option from two scenarios or state neither were acceptable. Background questions included socio-demographics. DCE data were analysed using conditional and heteroskedastic conditional logit models to create forecasts of acceptability. Results: The study sample comprised members of the public living in Scotland (n=1,004) and Sweden (n=974). All five attributes were important in driving respondents' choices. Swedish and Scottish preferences were mostly homogenous with the exception of 'who oversees the research using linked data', which had relatively less impact on the choices observed from Scotland. For a defined 'typical' linked data scenario, the probability (on average) of acceptance was 85.7% in Sweden and 82.4% in Scotland. Conclusion: This study suggests that the public living in Scotland and Sweden are open to using anonymised linked data in certain scenarios for research purposes but some caution is advisable if the anonymised linked data joins health to non-health data.

Background

The process of producing linked datasets is defined as "the bringing together from two or more different sources, data that relate to the same individual, family, place or event" [1]. There are increasing examples of linking data on healthcare resource use and patient outcomes from different sectors of health and social care systems and potentially with data outside of health (linked data). For example, data from hospital admissions can be linked with medical records held by general practitioners (GPs) or more generally linking data from hospital medical records with national mortality data. Such linked data has been used to investigate the association between diabetes and cancer [2] or understand the hazards of discontinuing certain medications after an acute myocardial infarction [3]. It is also possible to link data from the health care sector with data from other sources, such as social care or education [4]. This broadens the types of research questions that can be addressed; for example, data from GP records linked with government data have been used to investigate the relationship between epilepsy diagnoses and social deprivation [5].

Research using linked data generally uses anonymised data, where data have been converted so that individuals can no longer be identified within the final dataset [6]. This process of producing anonymised linked data can occur before anonymization or by using a common identifier across the datasets comprising the linked data. In many jurisdictions, there are no legal restrictions to the use of such anonymised linked data, even in the absence of explicit patient consent. One example jurisdiction, Sweden, is famous for the extent of its use of national registries of anonymised linked data, and the universal unique personal identity number that makes data linkage comparatively straightforward [7, 8]. Despite the legality of using anonymised linked data, use of such data for purposes other than the original use for which the data were collected ('secondary use') has become contentious in some jurisdictions. Public objections have resulted in the failure of national data science initiatives in England [9] and Australia [10], for example.

There is also evidence of heterogeneity in views about the use of anonymised linked data within countries. Evidence suggests that some people are willing to give a general consent for using anonymised linked data but others are content not to be asked for consent provided that the data are used in studies that have been reviewed and approved by an ethics committee [11]. Published systematic reviews indicate the majority of the existing evidence on the preferences of the public about using anonymised linked data comes from the United Kingdom (UK) and particularly Scotland [12, 13]. Relatively little empirical research has been conducted to understand the preferences of the public from jurisdictions such as Sweden, which has widespread collection and use of anonymised linked data [7, 8]. Kodate conducted secondary analysis of media articles published in Swedish newspapers between 1995 and 2005 and identified that the media made frequent calls for improving the quality assurance systems underpinning the use and reporting of data from national registries [14].

Three published systematic reviews have summarised the literature that aims to understand the attitudes of members of the public to the use of linked data [12, 13, 15]. The individual studies identified in these systematic reviews were largely conducted in a single country, or a single region of a country, and there was a paucity of studies making a comparison between jurisdictions. There is some evidence that preferences may differ between jurisdictions. The European Commission's Special Eurobarometer on Data Protection in 2011, reported that 66% of people living in Sweden were unconcerned about unnecessary disclosure of personal information compared with 19% of people living in the UK [16]. Furthermore, the majority (63%) of people living in Sweden were unconcerned about the secondary use of data compared with 20% of people living in the UK.

There are published examples of quantitative studies designed to collate views of the public about using anonymised linked data exist but these studies have traditionally used opinion-based survey or Likert-style agreement questions [13, 17-19], which are limited in their ability to identify the factors driving preferences. Questions which require rating or ranking may reveal individuals' order of preference, but they are not able to quantify differences in magnitude of preference. For example, people may clearly prefer apples to oranges and on a traditional survey question may give apples all five marks and oranges only three marks, or rank apples before oranges in a list. From these rating, you can elicit the order of

preference, but it is not always possible to understand how much better apples are than oranges, or how many oranges an individual would exchange for an apple. Similarly, it does not reveal when increases in apples are no longer satisfying to an individual; perhaps when the individual has ten apples, they would prefer an orange.

Discrete choice experiments (DCEs) are a stated preference method which aim to elicit and quantify the preferences of a sample of the population for a specified service or product described by a set of characteristics (attributes and levels) [20]. A DCE takes account of the opportunity cost when making such choices as people have to make trade-offs between the attributes when they choose their preferred scenario from a set of hypothetical scenarios called 'choice-sets' [20]. In each choice-set, the respondent is presented with options described by the same attributes in varying amounts (levels) [21]. Under the theory of random utility maximisation (see Technical Appendix), it is assumed that individuals choose the option which would provide them with the most value ('utility') and thus the choices made reveal both their preferences and the relative importance of each attribute when making their choice [22]. With the estimates of utility, is also possible to calculate the expected return and forecast from the collected data to estimate the probability of an individual choosing a particular scenario over another. DCEs are increasingly used to quantify people's preferences for health goods, services and interventions, where normal markets rarely exist [23-26]. Although the number of studies in health economics is growing, applications started in the 1990s, making the approach relatively new compared to other question formats. Consensus-based guidelines on best research practice have been produced by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) for researchers seeking to use DCEs to quantify preferences in a healthcare context [27-29]. The aim of this study was to identify and compare the factors most influential in shaping public preferences, in two exemplar jurisdictions, for the type and use of linked data in a research context.

Methods

This study used an online survey to field a DCE to compare the preferences of a sample of the public representing two jurisdictions (Scotland and Sweden). The Scottish DCE was conducted in 2016 [30]; this study replicated that DCE in Sweden, to enable comparisons between the two jurisdictions.

The DCE design and analysis are reported in line with published guidance [20, 29]. The online survey comprised four sections: an initial page of narrative introducing key concepts, and rationale for the sharing and use of anonymised linked data; questions about attitude to sharing and use of anonymised linked data; the choice-sets that formed the DCE; and socio-demographic questions.

Conceptualising the choice question

The DCE used two 'unlabelled' alternatives to present scenarios describing the type and use of anonymised linked data. Respondents were asked to select which, if any, of the two alternatives was their

preferred option. Respondents could also indicate if neither of the two alternatives was acceptable, allowing respondents the option to 'opt-out'. Figure 1 shows an example choice-set.

Attribute and level selection

Each alternative scenario was described with five attributes and plausible levels (see Table 1). For all but one attribute, there were four levels; the remaining attribute (research purpose) had three levels as there was no meaningful fourth level. Detail of the identification and generation of the five attributes and their levels has been published previously in relation to the original Scottish DCE [30]. Briefly, the attributes were chosen as the most important characteristics of sharing and using linked data of concern to the public, based on qualitative research [31] and a systematic review of the literature on public attitudes to linked data [12]. The levels were chosen to represent a range of actual or potential variations in these attributes and set to be within realistic and meaningful ranges to represent how linked data could be potentially shared and used. The wording of the attributes and levels was refined through iterations and engagement with members of an existing public involvement panel (the Farr Institute Scotland Public Panel).

Experimental design

There were 768 (4^4 x 3^1) unique profiles possible from the chosen attributes and levels, which could create 294,528 different combinations for the choice-tasks. To reduce this unmanageable number of potential alternatives, a main effects design was generated using Sawtooth Software [32] with each respondent allocated to one of 40 blocks each containing six choice-sets. The paired alternatives selected by the software were reviewed to remove irrational or implausible choice sets. Pilot testing in Scotland revealed that using twelve choice-sets resulted in respondent fatigue, hence each respondent was presented with six choice-sets in a random order in the main survey [30].

Survey design and piloting

The DCE was embedded as part of an online survey, as described earlier. The Swedish DCE used the same design as that used previously in Scotland, with appropriate changes for the different organisational health care systems. Forward and backward translation was conducted by an independent organisation and validated by bilingual members of the research team. Each survey was tested, using qualitative piloting in interviews in each country, with a convenience sample of 20 people of a variety of ages and gender. The aims of the qualitative pilot were to ensure respondents understood the instructions and the language used, and to test how they interacted with the survey and how long they took to complete it. Minor changes were made to the ordering and wording of some questions in Scotland for improved clarity [30] and these were carried forward to the Swedish survey where no additional changes were needed.

Study population and sampling frame

The relevant study population for this study were adult (18 years and over) members of the public from two selected example jurisdictions (Scotland and Sweden). Scotland was chosen as an exemplar because National Health Service (NHS) Scotland is a publicly funded health care system that has the capacity to share and use anonymised linked data. Sweden was chosen as a comparator because the use of linked data is relatively more common, with large national registries integrating health and other social data used to answer a range of research questions. The two jurisdictions have comparable universal healthcare coverage by either national (the NHS in Scotland) or local (county councils in Sweden) providers, respectively.

For a DCE, the required sample size depends on the number of choice-sets, the number of alternatives in a choice set, and the number of levels attached to an attribute [33]. Given these characteristics, and the objectives to explore preference heterogeneity and compare the responses between Scotland and Sweden, a sample of 1,000 respondents from each country was deemed more than sufficient for this study. In this DCE, the power calculation for sample size suggested by Orme would indicate a minimum sample size of 167 [33]. This power calculation, however, does not make allowances for investigations into preference heterogeneity nor the difference in preferences between Sweden and Scotland. A published review of sample sizes in DCEs found that, out of 505 healthcare DCE studies, only six had sample sizes of over 1,000 [34].

The DCE was sent to a sample of adult members of the public in the two countries (Scotland and Sweden). The sample was identified using an international market research company, Ipsos [35] (called Ipsos Mori in the UK), who provide members of online panels [36]. Participants were members of the Ipsos international panel (called i-Say), who had volunteered to take part in regular market research surveys. Panellists received regular invitations from Ipsos to participate in surveys and were free to decide whether to complete any individual survey. Panellists were selected at random, and invited to take part via an email, with quotas set on key demographic variables, namely, age, gender, and working status, with the aim of achieving a sample of 1,000 people in each country who were representative of the population for these criteria. The Scottish survey was conducted in August 2016 and the Swedish survey in June 2017; both were live for 14 days, until the quotas were filled.

Screening questions were used at the start of the survey, based on the attributes and levels. For example, respondents were asked which of the levels in the attribute “the purpose of research” was closest to their view, along with the option to select that “data linkage should not be permitted under any circumstances”. Respondents selecting the latter were routed-out and did not complete the DCE [30]. It was hypothesised that these respondents would always select the opt-out option. Removing them from the sample ensured that DCE respondents did not fundamentally object to data linkage and thus allowed an investigation of the nuanced public preferences for conducting research with linked data.

Analysis

Choice data from the DCE were analysed using discrete choice models. All attributes were categorical and were dummy coded relative to a base level (Table 1) that was deemed to be the ‘worst’. The primary

analysis estimated the preferences from each sample of respondents from the two countries separately using a conditional logit model. To further compare data between Scotland and Sweden, a pooled conditional logit model was estimated with interaction terms between dummy variables that identified the respondent's nationality (1=Scottish) and each attribute level. To account for differences in scale, a pooled heteroskedastic conditional logit model with these same interactions was also estimated [37, 38]. The scale parameter was allowed to vary by the respondent's nationality. In order to identify the scale term, preferences over one attribute must be restricted to be equal across countries. This attribute (purpose of the research) was selected based on statistically insignificant interaction terms in the pooled conditional logit model. All analyses were completed using Stata 13 [39].

The probability of an individual finding a specific scenario acceptable was calculated by estimating the expected observable utility of an alternative and comparing it with expected utility of another. A 'typical' linked data scenario was defined as university researchers or health service staff using linked health records for general public benefit, the profit is invested in public services and the process is overseen by the relevant public services. Two scenarios were then specified as: best-case (the most risk averse scenario, where only university researchers use linked data from health records for the benefit of people whose data are being used, there is no profit made and the process is overseen by a non-governmental body) and worst-case (where university researchers, health service staff, government and commercial researchers use health data linked to social care, education, employment and private sector data for research with any purpose, where the profit is kept by those carrying out these research who also oversee the process). Investigations into preference heterogeneity were conducted using a split sample analysis and comparing the probabilities of scenarios being acceptable.

Results

A total of 1,978 respondents completed the survey and were included in the analysis (Table 2). An additional 974 respondents (461 in Scotland and 513 in Sweden) started the survey but were routed-out at the initial questions (presented in the order shown in Table 3) because they stated that sharing or using linked data was unacceptable under any conditions.

The results of the conditional logit model (Table 4) suggested that all attribute levels were statistically significant ($p<0.01$). The positive coefficients indicated that all levels were preferred, relative to the 'worst' level of each attribute that were used as a 'base level'. The absolute values of the estimated coefficients cannot be interpreted as pure 'preference weights' because these estimated values measure *relative* preference [27] and, therefore, only the relative sizes of changes across levels, in each country, have meaningful interpretations.

The positive interaction terms between levels and nationality could suggest that Scottish respondents have stronger preferences over using linked data than those living in Sweden. However, the inflated coefficients for Scotland may be driven by differences in scale (choice consistency). Table 5 shows the results of the heteroskedastic conditional logit model with interaction and scale terms. These results show that the scale term was statistically significant at the 10% level (p=0.052) which can be interpreted to indicate that the Scottish sample were, on average, more 'consistent' in their decision making when making choices. The estimated error term (the variance of the unobservable element of utility) was smaller in the Scottish sample relative to the Swedish sample. The statistically significant and positive constant term also suggests that, all else being equal, respondents preferred their data not to be used or linked.

The attribute for the source of data being linked aligned with *a priori* expectations as respondents preferred fewer sources of data being linked in both countries. On average, respondents generally preferred different types of health records being linked together rather than health records linked with public social, education or employment records, although all of these scenarios were preferable to linkage with private sector records (the base case). Respondents in both countries preferred that nobody profits rather than the profits go to those carrying out the work. However, people in both countries, on average, preferred profit to be invested into public services or shared with the public than nobody profiting at all.

For the attributes 'who does the research', 'type of data being linked' and 'profit-making', average preferences between the two countries were relatively homogenous. Similarly, the constant term did not significantly differ between the groups indicating respondents in one country were no more or less likely to state that neither scenario was acceptable. The most prominent difference in preferences between Scotland and Sweden were for the attribute 'oversight' (see Table 5). Although, on average, respondents in both countries preferred external oversight of the research, the respondents living in Sweden seemed to value this as being more important, as it had a larger impact on their choice-making.

For the typical data linkage scenario, the probability of acceptance, on average, was 85.7% in Sweden and 82.4% in Scotland. In the 'best-case' scenario of data linkage, the probability, on average, of a Swedish person accepting the alternative was estimated to be 75.0% compared with 72.1% in Scotland. In the 'worst-case', the probability of the scenario being acceptable on average, was 35.0% in both countries. Figure 2 shows the probability of these different scenarios being acceptable in different subgroups of the sample. Differences in acceptability of the 'worst-case' scenario were most prominent when considering gender, with men 50% more likely to find this case of data linkage acceptable compared to women (44% and 27% retrospectively).

An interactive model showing average probability of acceptability in different scenarios is available in the online supplementary materials (see Appendix B). This allows the reader to see the impact of changing

attribute levels on the average probability of acceptability of, for example, the typical scenario. Changing the research attribute to the base level, 'research for any reason' (and keeping all others the same) decreased the probability of acceptance of the typical scenario by 6.1% and 7.1% in Sweden and Scotland respectively. In comparison, changing only the type of data attribute to the base level of 'health data linked to social care, education, employment and private-sector data' decreased the probability of the scenario being accepted by 11.0% and 12.6% in the two jurisdictions.

Discussion

This study quantified the aspects of sharing and using different types of linked data that drove the preferences of members of the public and estimated the potential impact on acceptability of using anonymised linked data. It is the first study to directly compare preferences for the sharing and use of linked data between two countries, showing that there were considerable similarities in average preferences amongst members of the public in Scotland and Sweden. The exception to these common preferences was that people living in Sweden were more influenced by who should have control of 'external oversight' when sharing or using linked data.

The considerable similarities between preferences in the two jurisdictions are curious, given the perceived differences in the use of linked data between the two countries. In Sweden, the use of a unique personal identifier means that the creation and use of registries of data are common [7, 8] but this is less so in Scotland. However, Scotland is distinct from the rest of the UK in that, since the 1970s, all NHS patients have been assigned a Community Health Index number which registers data on address, postcode, GP, date of birth, region of registration and date of death. This is used in all primary health care activities and hospital-based clinical information systems, throughout NHS Scotland, including the Emergency Care Summary. Nevertheless, the lack of a universal personal identifier reduces the ease with which data from different sectors can be linked. Throughout the UK, health data science research is increasing [40] and there is an increasing move towards data use and linkage across sectors where "Data will drive Scotland's next economic revolution" [41]; knowledge of public preferences and acceptability will be vital in ensuring that there is a social license for such work [9]. Therefore, it is reassuring that the preferences in Scotland are so similar to those in a jurisdiction where data use and linkage are perceived as both commonplace and acceptable.

The main differences between the two jurisdictions related to external oversight of data use and linkage. Swedish respondents were more likely to prefer oversight by either the government or the relevant public service, or, to a lesser extent, an independent body (Table 5) than were the Scottish respondents. Swedish national data registries are tightly controlled and their use requires review both by ethics committees and the government organisation Statistics Sweden [8]. In addition, the delivery of healthcare is devolved locally and one study found that Swedes were more likely to want to be involved in local healthcare organisations then were people in England [42]. However, such oversight is not always valued in Sweden. In one study on public perceptions of biobank research, respondents did not trust either the government or county councils to evaluate the risks and benefits of genetic research being proposed [43].

The “best case” scenario chosen in this study was the most risk averse scenario. However, the average probability of this being accepted (over 70% in both jurisdictions) was less than that of the typical scenario, where the average probability of it being accepted was over 80%. This typical scenario was chosen to be similar to much of the health research conducted with linked data, such as that cited earlier [2, 3]. Previous studies have suggested that public benefit is potentially a key condition for acceptability of research using linked data [12, 44, 45]. However, changing this single attribute to the base case (‘research for any reason’) decreased the probability of acceptance of the scenario by only 6-7%. In comparison, changing only the type of data that were linked to the base level (linkage of multiple types of public and private sector data) decreased the probability of the scenario being accepted by almost twice as much. When considered individually in surveys, the purpose of the research, particularly for research involving commercial companies, has been shown to be important for public acceptability of the research [12, 44, 45]. Those studies were limited in their abilities to compare factors that influence acceptability. This study has been able to quantify the differences in acceptability between such factors and this suggests that public benefit may be slightly less important to respondents than the types of data that were being linked.

Multiple types of data linked together, as with the base case, may be of concern to respondents because this is currently an unfamiliar consideration within data linkage. Some members of the public (33%) in the UK are very aware that the NHS uses health data in research, but much fewer (16%) are aware that commercial companies also do so [44]. By extrapolation, it would be expected that even fewer would be aware of the potential for doing research using both types of data linked together. The involvement of the private sector in the use of health data alone is already known to be a controversial topic. Other research has found that 17% of the general public would not accept commercial use of data at all [44] and qualitative studies found that there is a belief in a hidden agenda with commercial companies [31]. Linkage of supermarket loyalty cards with other sources of lifestyle data has been proposed, for example, as a resource for research into obesity [46]. Therefore, there is a need to understand further these preferences of the public with regard to data linkage at such scale, before this takes place.

For both countries, the least preferred option for the profit attribute, compared to the base case, was that nobody would profit. Profit by commercial companies, particularly large amounts of profit, is known to be of particular concern to the public [13]. Many people are also concerned about the British government profiting from selling health data to private healthcare companies [47]. Nonetheless, in this study, the preferred option was that profits be reinvested into the public services, suggesting a very complex and nuanced set of opinions and preferences around the creation and use of profit from the use of health data, particularly when commercial companies are involved. This has been explored in detail by the Wellcome Trust [44] and this study adds to that evidence base by quantifying the preferences. However, it is acknowledged that this study could not tease out the differences in respondents’ preferences about how companies made the profit described in the scenarios. For example, the research mentioned in the scenario could have created profit by sales of a newly developed healthcare product or increased sales of an existing product, either of which could deliver the levels used for the attribute “the purpose of the research”.

The large numbers of respondents in both countries, and their representativeness in observable demographics to the overall national population, were key strengths of the study. These substantial numbers allowed comparisons across subgroups, showing differences in acceptability of the worst-case scenario between men and women and young and old. All respondents were recruited using an internet panel provider to facilitate collecting a large study sample relatively inexpensively. Although DCEs in healthcare are increasingly administered online [48], the limitations of using online surveys and/or internet panels for stated preference studies has not been thoroughly investigated. For instance, respondents to online DCEs are more likely to be computer-literate and, particularly in older age groups, may not be representative of the general population. Their views could possibly be different to those who would have preferred a paper-based method. However, there is some evidence suggesting online health surveys provide good quality data compared with other methods such as postal surveys or telephone interviews [49]. In addition, previous DCEs using more population-representative sampling frames (e.g. from the electoral roll) have resulted in very low response rates, and hence had limited representativeness and generalisability for different reasons [50].

The DCE was restricted to those respondents who did not believe the sharing and use of linked data should be allowed under any circumstances. The choice to limit the population in this way was made to ensure that we could investigate the nuanced opinions of those who agreed, in principle, to sharing and using linked data [30]. Had those people been included, there was a risk that those respondents would have selected the 'opt-out' option in every scenario. When a respondent chooses to opt-out, nothing about their preferences are revealed as there are no trade-offs with the attributes or levels presented in the hypothetical alternatives. For respondents who did not believe the sharing and use of linked data should be allowed under any circumstances, the option to opt-out may have drawn disproportionately from the other alternatives affecting the average and relative choice shares presented in Figure 2 [51]. Table 3 shows, however, that there was not complete opposition to sharing and using linked data. Respondents were not aware that they would be routed-out of the DCE by answering in a particular way, so there is no reason to doubt their answers. In both jurisdictions, only small numbers of respondents did not accept that research should be conducted for one of the purposes presented in the scenarios (the first question). Who the researchers were and who was allowed to profit were much more controversial topics, particularly in Sweden, with most respondents being routed-out at these questions (Table 3). It is possible, therefore, that we were too risk averse in excluding these respondents. In addition, the removal of these individuals limits the generalisability of the study findings. Future research may wish to extend the study sample to include respondents who disagree with the sharing and use of linked data in principle and investigate two-way or higher-order interactions between the attribute levels by incorporating these into the experimental design [28] to understand if certain combinations result in more/less acceptable scenarios. There is also the need for future qualitative investigations in Sweden to understand why so many respondents had concerns regarding sharing and using data, given how commonplace are the use of national registries.

Conclusions

This study suggests that the public living in Scotland and Sweden are open to using anonymised linked data in certain scenarios for research purposes but some caution is advisable if health data are linked to non-health data. Despite the use of linked data for national registries being more common in Sweden, replicating the Scottish DCE there has revealed substantial similarities in the preferences of the public in the two jurisdictions. Given the considerable investment in data intensive health research and the use of linked data in Scotland, this suggests there may be considerable value in further comparative work between the two countries. In particular, further research to understand the reasons underpinning public concerns around data linkage and sharing in relation to experiences in Sweden might be valuable to inform the development of future practices in Scotland.

Abbreviations

GP – General Practitioner

DCE – discrete choice experiment

NHS – National Health Service

UK – United Kingdom

Declarations

Ethics approval and consent to participate

The study was discussed with the Ethics Committee of the University of Manchester and the ethical vetting process in Sweden. Their opinions were that the study collected non-controversial, non-sensitive and non-personal data, which would be anonymous when passed to the researchers. Ipsos complies with a number of relevant industry quality standards including ISO 27001:2005, the international standard for security of personal information.

Consequently, it did not require formal research ethics approval.

Consent for publication

Not applicable

Availability of data and material

An interactive model has been made available as supplementary material in Appendix B.

Competing interests

The authors declare that they have no competing interests

Funding

This study was funded by The Farr Institute. The Farr Institute is supported by a 10-funder consortium: Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the Medical Research Council, the NIHR, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates), and the Wellcome Trust, (MRC Grant No: MR/K006665/1).

Authors' contributions

MPT: designed and conducted DCE in Sweden, designed and conducted comparative study and wrote the manuscript

CB: conducted qualitative pilot work in Sweden and validated translation and contributed to the writing of the manuscript

MA: designed and conducted DCE in Scotland, contributed to design of comparative study and contributed to the writing of the manuscript

CV: Conducted the data analysis and contributed to the writing of the manuscript.

Acknowledgements

Gareth McAteer, Clive Fostock and Sara Davidson at Ipsos Mori, Scotland, for helping with the practical aspects of data collection in Sweden

References

1. Holman CDAJ, Bass JA, Rosman DL, Smith MB, Semmens JB, Glasson EJ, Brook EL, Trutwein B, Rouse IL, Watson CR *et al*: **A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system**. *Australian Health Review* 2008, **32**(4):766-777.
2. Williams R, van Staa TP, Gallagher AM, Hammad T, Leufkens HGM, de Vries F: **Cancer recording in patients with and without type 2 diabetes in the Clinical Practice Research Datalink primary care data and linked hospital admission data: a cohort study**. *BMJ Open* 2018, **8**(5):e020827.
3. Boggon R, van Staa TP, Timmis A, Hemingway H, Ray KK, Begg A, Emmas C, Fox KAA: **Clopidogrel discontinuation after acute coronary syndromes: frequency, predictors and associations with death and myocardial infarction–a hospital registry-primary care linked cohort (MINAP-GPRD)**. *European Heart Journal* 2011, **32**(19):2376-2386.
4. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford DV, Brown G, Leake K: **The SAIL databank: linking multiple health and social care datasets**. *BMC Medical Informatics and Decision*

Making 2009, **9**(1):3.

5. Pickrell WO, Lacey AS, Bodger OG, Demmler JC, Thomas RH, Lyons RA, Smith PEM, Rees MI, Kerr MP: **Epilepsy and deprivation, a data linkage study**. *Epilepsia* 2015, **56**(4):585-591.
6. Information Commissioner's Office. **Anonymisation: managing data protection risk code of practice**. Date Accessed: 26/10/2017; Available from: https://ico.org.uk/for-organisations/guide-to-data-protection/anonymisation/ Archived by WebCite® at http://www.webcitation.org/6uVKws7MH.
7. Emilsson L, Lindahl B, Köster M, Lambe M, Ludvigsson JF: **Review of 103 Swedish Healthcare Quality Registries**. *Journal of Internal Medicine* 2014, **277**(1):94-136.
8. Ludvigsson JF, Almqvist C, Bonamy A-KE, Ljung R, Michaëlsson K, Neovius M, Stephansson O, Ye W: **Registers of the Swedish total population and their use in medical research**. *European Journal of Epidemiology* 2016, **31**(2):125-136.
9. Carter P, Laurie GT, Dixon-Woods M: **The social licence for research: why care.data ran into trouble**. *J Med Ethics* 2015, **41**(5):404-409.
10. Garrety K, McLoughlin I, Wilson R, Zelle G, Martin M: **National electronic health records and the digital disruption of moral orders**. *Soc Sci Med* 2013, **101**:70-77.
11. Damschroder LJ, Pritts JL, Neblo MA, Kalarickal RJ, Creswell JW, Hayward RA: **Patients, privacy and trust: patients' willingness to allow researchers to access their medical records**. *Soc Sci Med* 2007, **64**(1):223-235.
12. Aitken M, de St. Jorre J, Pagliari C, Jepson R, Cunningham-Burley S: **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* 2016, **17**(1).
13. Hill EM, Turner EL, Martin RM, Donovan JL: **"Let's get the best quality research we can": public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study**. *BMC Med Res Methodol* 2013, **13**:72.
14. Kodate N: **Events, Public Discourses and Responsive Government: Quality Assurance in Health Care in England, Sweden and Japan**. *Journal of Public Policy* 2010, **30**(3):263-289.
15. Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC: **Written informed consent and selection bias in observational studies using medical records: systematic review**. *BMJ* 2009, **338**.
16. TNS Opinion & Social. **Special Eurobarometer 359. Attitudes on Data Protection and Electronic Identity in the European Union**. Brussels 2011 Date Accessed: 10/06/2019. Available from: http://ec.europa.eu/public_opinion/archives/eb_special_359_340_en.htm#359
17. Buckley BS, Murphy AW, MacFarlane AE: **Public attitudes to the use in research of personal health information from general practitioners' records: a survey of the Irish general public**. *J Med Ethics* 2011, **37**:50-55.
18. Medical Research Council. **The Use of Personal Health Information in Medical Research** 2007 Date Accessed: 10/06/2019. Available from: https://www.mrc.ac.uk/documents/pdf/the-use-of-personal-health-information-in-medical-research-june-2007/

19. Whiddett R, Hunter I, Engelbrecht J, Handy J: **Patients' attitudes towards sharing their health information**. *International Journal of Medical Informatics* 2006, **75**(7):530-541.
20. Lancsar E, Louviere J: **Conducting Discrete Choice Experiments to Inform Healthcare Decision Making**. *PharmacoEconomics* 2008, **26**(8):661-677.
21. Lancaster KJ: **A New Approach to Consumer Theory**. *Journal of Political Economy* 1966, **74**(2):132-157.
22. McFadden D: **Conditional logit analysis of qualitative choice behaviour**. In: *Frontiers in Econometrics.* Edited by Zarembka P. New York: Academic Press INC.; 1974: 105-142.
23. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW: **Discrete Choice Experiments in Health Economics: A Review of the Literature**. *PharmacoEconomics* 2014, **32**(9):883-902.
24. de Bekker-Grob EW, Ryan M, Gerard K: **Discrete choice experiments in health economics: a review of the literature**. *Health Economics* 2012, **21**(2):145-172.
25. Groothuis-Oudshoorn CGM, Fermont JM, van Til JA, Ijzerman MJ: **Public stated preferences and predicted uptake for genome-based colorectal cancer screening**. *BMC Medical Informatics and Decision Making* 2014, **14**(1):18.
26. Wortley S, Tong A, Lancsar E, Salkeld G, Howard K: **Public preferences for engagement in Health Technology Assessment decision-making: protocol of a mixed methods study**. *BMC Medical Informatics and Decision Making* 2015, **15**(1):52.
27. Hauber AB, González JM, Groothuis-Oudshoorn CGM, Prior T, Marshall DA, Cunningham C, Ijzerman MJ, Bridges JFP: **Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force**. *Value in Health* 2016, **19**(4):300-315.
28. Reed Johnson F, Lancsar E, Marshall D, Kilambi V, Mühlbacher A, Regier DA, Bresnahan BW, Kanninen B, Bridges JFP: **Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force**. *Value in Health* 2013, **16**(1):3-13.
29. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, Johnson FR, Mauskopf J: **Conjoint analysis applications in health–a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force**. *Value Health* 2011, **14**(4):403-413.
30. Aitken M, McAteer G, Davidson S, Frostick C, Cunningham-Burley S: **Public Preferences regarding Data Linkage for Health Research: A Discrete Choice Experiment**. *International Journal of Population Data Science* 2018, **3**(1):11.
31. Aitken M, Cunningham-Burley S, Pagliari C: **Moving from trust to trustworthiness: Experiences of public engagement in the Scottish Health Informatics Programme**. *Science and Public Policy* 2016, **43**(5):713-723.
32. Sawtooth Software Inc.: **Sawtooth SSI Web 8.3.8** [program], 2012.
33. Orme B: **Sample size issues for conjoint analysis studies**. In: *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research.* 2nd edn. Madison, Wis: Research Publishers LLC.; 2010: 57-66.

34. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA: **Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide**. *The Patient - Patient-Centered Outcomes Research* 2015, **8**(5):373-384.
35. Ipsos. **Website**. Date Accessed: 10/06/2019; Available from: https://www.ipsos.com/en.
36. Ipsos. **i-Say panels**. Date Accessed: 10/06/2019; Available from: https://social.i-say.com/.
37. Vass CM, Wright S, Burton M, Payne K: **Scale Heterogeneity in Healthcare Discrete Choice Experiments: A Primer**. *Patient* 2018, **11**(2):167-173.
38. Hole AR: **Small-sample properties of tests for heteroscedasticity in the conditional logit model**. *Economics Bulletin* 2006, **3**(18):1-14.
39. StataCorp: **Stata Statistical Software: Release 13** [program]. College Station, TX: StataCorp LP., 2013.
40. HDRUK. **Health Data Research UK**. Date Accessed: 17/06/2019; Available from: https://www.hdruk.ac.uk/.
41. Scottish Enterprise. **Data driven innovation**. Date Accessed: 17/06/2019; Available from: https://www.scottish-enterprise.com/support-for-businesses/develop-products-and-services/data-driven-innovation.
42. Fredriksson M, Eriksson M, Tritter J: **Who wants to be involved in health care decisions? Comparing preferences for individual and collective involvement in England and Sweden**. *BMC Public Health* 2017, **18**(1):18.
43. Kettis-Lindblad Å, Ring L, Viberth E, Hansson MG: **Genetic research and donation of tissue samples to biobanks. What do potential sample donors in the Swedish general public think?** *European Journal of Public Health* 2005, **16**(4):433-440.
44. Wellcome Trust. **The One-Way Mirror: Public attitudes to commercial access to health data** 2016 Date Accessed: 10/06/2019. Available from: https://wellcome.ac.uk/sites/default/files/public-attitudes-to-commercial-access-to-health-data-wellcome-mar16.pdf Archived by WebCite® at http://www.webcitation.org/6uQJbVLrc.
45. Tully MP, Bozentko K, Clement S, Hunn A, Hassan L, Norris R, Oswald M, Peek N: **Investigating the Extent to Which Patients Should Control Access to Patient Records for Research: A Deliberative Process Using Citizens' Juries**. *J Med Internet Res* 2018, **20**(3):e112.
46. Morris MA, Wilkins E, Timmins KA, Bryant M, Birkin M, Griffiths C: **Can big data solve a big problem? Reporting the obesity data landscape in line with the Foresight obesity system map**. *International Journal of Obesity* 2018, **42**(12):1963-1976.
47. Royal Statistical Society. **Royal Statistical Society research on trust in data and attitudes toward data use/data sharing** 2014 Date Accessed: 23/10/2017. Available from: https://www.statslife.org.uk/images/pdf/rss-data-trust-data-sharing-attitudes-research-note.pdf Archived by WebCite® at http://www.webcitation.org/6uQKsMF3z.
48. Soekhai V, de Bekker-Grob EW, Ellis AR, Vass CM: **Discrete Choice Experiments in Health Economics: Past, Present and Future**. *PharmacoEconomics* 2019, **37**(2):201-226.

49. Mulhern B, Longworth L, Brazier J, Rowen D, Bansback N, Devlin N, Tsuchiya A: **Binary Choice Health State Valuation and Mode of Administration: Head-to-Head Comparison of Online and CAPI**. *Value in Health* 2013, **16**(1):104-113.
50. Giles EL, Becker F, Ternent L, Sniehotta FF, McColl E, Adams J: **Acceptability of Financial Incentives for Health Behaviours: A Discrete Choice Experiment**. *PLOS ONE* 2016, **11**(6):e0157403.
51. Campbell D, Erdem S: **Including Opt-Out Options in Discrete Choice Experiments: Issues to Consider**. *The Patient - Patient-Centered Outcomes Research* 2019, **12**(1):1-14.
52. McFadden D, Train K: **Mixed MNL models for discrete response**. *Journal of Applied Econometrics* 2000, **15**(5):447–470.

Tables

Table 1: Attributes and Levels

Attribute	Levels (text variation for Sweden in brackets)
The researchers are:	Only university researchers.
	Only university researchers or NHS staff (researchers employed by a county council).
	Only university researchers, NHS staff or government researchers (researchers employed by a county council or researchers employed by one of the authorities).
	University researchers, NHS staff, government researchers (researchers employed by a county council or researchers employed by one of the authorities) and commercial researchers such as market research organisations or pharmaceutical companies.[a]
The type of data being linked:	Information from your GP (primary care) records being linked with information from your other NHS (county council) health records e.g. hospital records.
	Information from your NHS (county council) health records being linked with information from your social care or education records.
	Information from your NHS (county council) health records being linked with information from your social care or education records, or from your employment and benefits (national health insurance) records.
	Information from your NHS (county council) health records being linked with information from your social care, education, employment, and benefits (national health insurance) records, as well as information collected about you in the private sector e.g. through online shopping accounts. [a]
The purpose of the research:	Research using linked information should only be conducted if it will have direct benefits for the people whose information is being used.
	Research using linked information should only be conducted if it will have general public benefits.
	Research using linked information should be allowed for any reason.[a]
Profit-Making:	Nobody should be allowed to profit from research carried out using linked information.
	Any profit made from research carried out using linked information should be shared with the public.
	Any profit made from research carried out using linked information should be invested into public services.
	Any profit made from research carried out using linked information should be kept by those carrying out the research. [a]
Oversight:	The process should be overseen by the Scottish (Swedish) Government.
	The process should be overseen by a non-governmental independent body (an independent body that is not part of the Swedish Government).
	The process should be overseen by the relevant public service(s); for example, research that uses information from people's health records should be overseen by the NHS (county council).
	The process should be overseen by the organisations undertaking the research. [a]

[a] base level in the analysis

Table 2. Characteristics of the study sample

Characteristic	Scotland N = 1,004 (%)	Sweden N = 974 (%)
Gender		
Male	421 (41.9%)	499 (51.2%)
Female	583 (58.1%)	475 (48.8%)
Age		
18 – 34 years	275 (27.4%)	354 (36.3%)
35 – 54 years	358 (35.7%)	411 (42.2%)
55+ years	371 (37.0%)	209 (21.5%)
Employment [a]		
Working part or full time	557 (55.5%)	627 (64.3%)
Not working	444 (44.2%)	341 (35.0%)

[a] missing data: Scotland (n=3) and Sweden (n=6)

Table 3. Number of respondents routed-out after responding "data linkage should not be permitted under any circumstances" to the initial survey questions

Initial survey questions, in order*	Scotland (n)	Sweden (n)
Purposes of research	67	51
Who are the researchers	104	186
What types of information may be linked	135	128
Management of potential profits	145	140
Arrangements for oversight/monitoring	8	5
Public involvement in data linkage research	2	3
Total	**461**	**513**

*Each question asked respondents which of the levels in Table 1 was closest to their view about the attribute (or that data linkage should not be permitted).

Table 4: Results of the conditional logit model

Attribute and level	Estimate coefficient (standard error)		
	Scotland	Sweden	Country comparison[b]
Researchers:			
University researchers	0.214*** (0.05)	0.168** (0.05)	0.047 (0.07)
University /health service staff	0.500*** (0.05)	0.312*** (0.05)	0.188** (0.07)
University/health service staff/government	0.445*** (0.05)	0.337*** (0.05)	0.108 (0.07)
University/health service staff/government/commercial	Base level[a]		
Data to be linked:			
Primary care linked to other health records	0.918*** (0.05)	0.706*** (0.05)	0.212** (0.07)
Health records linked to social care/education records	0.664*** (0.05)	0.403*** (0.05)	0.261*** (0.07)
Health records linked to social care/education/employment/benefits records	0.407*** (0.05)	0.171*** (0.05)	0.236** (0.07)
Health records linked to social care/education/ employment/benefits records/private sector	Base level[a]		
Purpose:			
Direct benefits for the people whose information is used	0.322*** (0.04)	0.254*** (0.04)	0.068 (0.06)
Research conducted if it will have general public benefits	0.548*** (0.04)	0.430*** (0.04)	0.118* (0.06)
Research for any reason	Base level[a]		
Profit-making:			
Nobody profits	0.326*** (0.05)	0.171*** (0.05)	0.156* (0.07)
Profit shared with the public	0.579*** (0.05)	0.397*** (0.05)	0.182* (0.07)
Profit invested into public services	0.739*** (0.05)	0.506*** (0.05)	0.233** (0.07)
Profit goes to those doing the research	Base level[a]		
Oversight:			
Overseen by independent body	0.346***	0.420***	-0.074

	(0.05)	(0.05)	(0.07)
Overseen by relevant public service	0.265***	0.457***	-0.192**
	(0.05)	(0.05)	(0.07)
Overseen by Government	0.066	0.289***	-0.223**
	(0.05)	(0.05)	(0.07)
Overseen by the organisation undertaking the research	Base level[a]		
Constant	0.886***	0.620***	0.266*
	(0.08)	(0.08)	(0.12)
Number of observations	**18072**	**17532**	**35604**

*p<0.05; **p<0.01; ***p<0.001

[a] Each attribute used categorical levels, which were dummy coded relative to a base level (Table 1) that was deemed to be the 'worst'.

[b] The country comparison model, estimated using pooled data using a condition logit model, included interaction terms between dummy variables that identified the respondent's nationality (1=Scottish) and each attribute level

Table 5: Results of the heteroskedastic conditional logit model

Attribute and level	Estimate coefficient (standard error)	
	Pooled data (Scotland and Sweden)	Interaction terms[b]
Researchers:		
University researchers	0.168** (0.05)	0.001 (0.07)
University /health service staff	0.312*** (0.05)	0.080 (0.08)
University/health service staff/government	0.337*** (0.05)	0.012 (0.08)
University/health service staff/government/commercial	Base level[a]	
Data to be linked:		
Primary care linked to other health records	0.706*** (0.05)	0.014 (0.11)
Health records linked to social care/education records	0.403*** (0.05)	0.118 (0.09)
Health records linked to social care/education/employment/benefits records	0.171*** (0.05)	0.148 (0.08)
Health records linked to social care/education/ employment/benefits records/private sector	Base level[a]	
Purpose:		
Direct benefits for the people whose information is used	0.253*** (0.03)	0[C]
Research conducted if it will have general public benefits	0.430*** (0.04)	0[C]
Research for any reason	Base level[a]	
Profit-making:		
Nobody profits	0.171*** (0.05)	0.086 (0.07)
Profit shared with the public	0.397*** (0.05)	0.057 (0.08)
Profit invested into public services	0.506*** (0.05)	0.074 (0.09)

Profit goes to those doing the research	Base level[a]		
Oversight:			
Overseen by independent body	0.420*** (0.05)	-0.148* (0.07)	
Overseen by relevant public service	0.457*** (0.05)	-0.249*** (0.07)	
Overseen by Government	0.289*** (0.05)	-0.238*** (0.07)	
Overseen by the organisation undertaking the research	Base level[a]		
Constant	0.620*** (0.08)	0.076 (0.11)	
Scale term (if the respondent is Scottish)		0.242	(0.12)
Number of observations	**35604**		

*p<0.05; **p<0.01; ***p<0.001

[a] Each attribute used categorical levels, which were dummy coded relative to a base level (Table 1) that was deemed to be the 'worst'.

[b] Interaction terms indicate the effect of being Scottish on the estimated coefficients.

[c] The estimated model included an interaction term in which the attribute 'purpose' was restricted to zero.

Figures

Which of these scenarios would you prefer:
(Please scroll down and over if necessary to see the entire scenario)

Pair 1 of 6

	SCENARIO 1	SCENARIO 2
The researchers are:	Only university researchers or NHS staff	Only university researchers
The type of data being linked (NB. All data would be anonymous):	Information from your NHS health records being linked with information from your social care or education records.	Information from your NHS health records being linked with information from your social care or education records, or from your employment and
The purpose of the research:	Research using linked information should be allowed for any reason.	Research using linked information should only be conducted if it will have general public benefits.
Profit-making:	Any profit made from research carried out using linked information should be invested into public services.	Any profit made from research carried out using linked information should be shared with the public.
Oversight:	The process should be overseen by the Scottish Government.	The process should be overseen by the relevant public service(s); for example, research that uses information from people's health records should be overseen by the NHS.
	○ Prefer Scenario 1	○ Prefer Scenario 2

Neither are acceptable to me ○

Next

Figure 1

Example of scenario choice

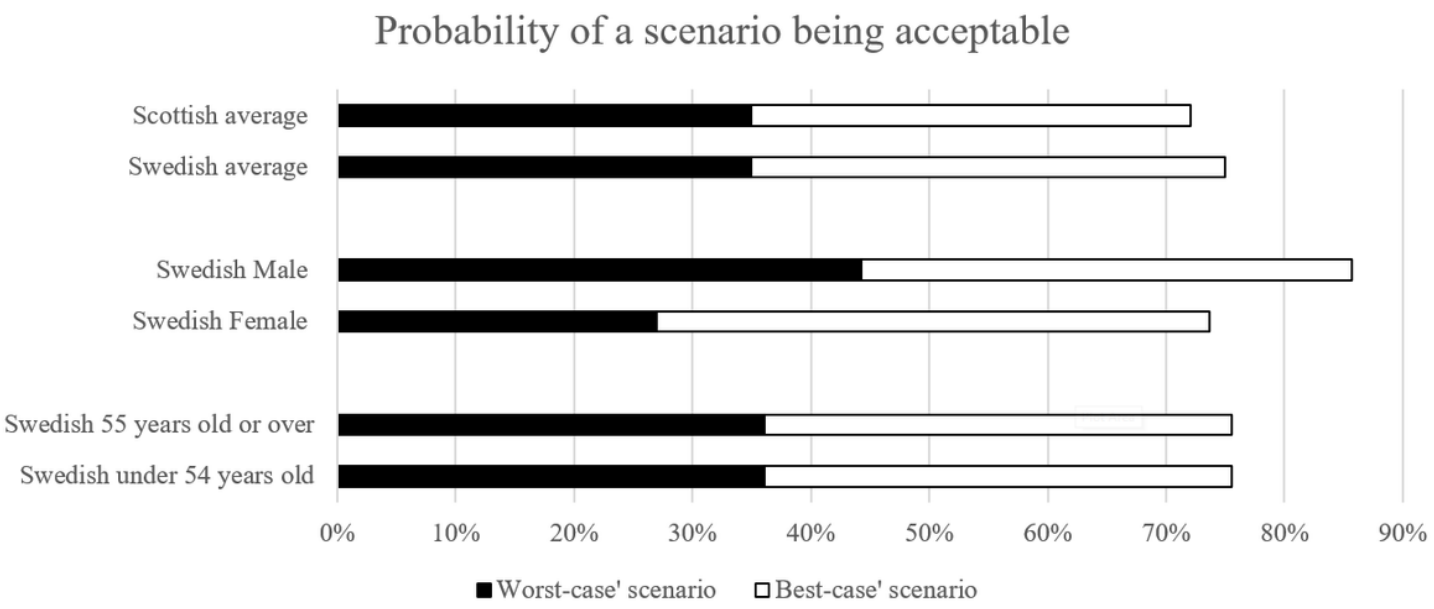

Figure 2

Probability of a scenario describing linked data being acceptable

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TechAppen.pdf