Statistical Disclosure Control: An Interdisciplinary Approach to the Problem of Balancing Privacy Risks and Data Utility

Michael Comerford

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Computing Science

College of Science and Engineering
University of Glasgow

December 2014

Abstract

The recent increase in the availability of data sources for research has put significant strain on existing data management work-flows, especially in the field of statistical disclosure control. New statistical methods for disclosure control are frequently set out in the literature, however, few of these methods become functional implementations for data owners to utilise. Current workflows often provide inconsistent results dependent on *ad hoc* approaches, and bottlenecks can form around statistical disclosure control checks which prevent research from progressing. These problems contribute to a lack of trust between researchers and data owners and contribute to the under utilisation of data sources.

This research is an interdisciplinary exploration of the existing methods. It hypothesises that algorithms which invoke a range of statistical disclosure control methods (recoding, suppression, noise addition and synthetic data generation) in a semi-automatic way will enable data owners to release data with a higher level of data utility without any increase in disclosure risk when compared to existing methods. These semi-automatic techniques will be applied in the context of secure data-linkage in the e-Health sphere through projects such as DAMES and SHIP.

This thesis sets out a theoretical framework for statistical disclosure control and draws on qualitative data from data owners, researchers, and analysts. With these contextual frames in place, the existing literature and methods were reviewed, and a tool set for implementing *k*-anonymity and a range of disclosure control methods was created. This tool-0set is demonstrated in a standard workflow and it is shown how it could be integrated into existing e-Science projects and governmental settings.

Comparing this approach with existing workflows within the Scottish Government and NHS Scotland, it allows data owners to process queries from data users in a semi-automatic way and thus provides for an enhanced user experience. This utility is drawn from the consistency and replicability of the approach combined with the increase in the speed of query processing.

Acknowledgements

I am grateful to the Economic and Social Research Council for funding this PhD studentship. I would like to thank my supervisors Joe Sventek, who agreed to step in and supervise a social scientist (and aspiring programmer), and Paul Lambert, who persevered for the duration - their support, advice and encouragement was invaluable. Thanks also go to John Watt and Susan McCafferty, formerly of the National e-Science Centre, for their help and guidance.

In addition, I would like to thank colleagues from the Scottish Government, the University of Edinburgh and National Records of Scotland, especially Ellen Lynch who was a particular source of encouragement: without her help I could not have gained access to interview respondents, data for case studies and the necessary resources to run our ESRC Festival of Social Science event. I would also like to add a personal thanks to Robert McCaa for the opportunity to work with the Minnesota Population Center.

I have also been extremely lucky to enjoy the support of my family throughout my studies. The comforts of home and time spent with my mum, dad, sister, brother and Molly have kept me relatively sane. I would also like to express my appreciation to my grandparents John and Jackie Ward for always encouraging me to question everything. Finally, to my wife Holly, thanks for not letting me join the merchant navy and encouraging me to continue studying. Your love has enabled me to endure the low points and celebrate the highs and I am forever grateful.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction – The Data Deluge and the Expansion of Data Sources

It would be remiss not begin this introduction without the obligatory nod to the vast quantity of data available to researchers, and the rapid increase in potential sources in recent years. The literature and commentators often refer to the volume of electronic data available for analysis as ‘big data’ (Lynch, 2008), the data ’deluge’ (Economist, 2010), the data ’explosion’ (Microsoft, 2013) and even a data ’tsunami’ (Argonne National Laboratory, 2012). The growth in data and opportunities for novel analysis has been recognised across disciplines. In the health field (which formed the original backdrop for this research), the case for e-Health and health informatics to harness these data resources has been championed by both academic and government organisations alike (see Silber (2003); Mansell (2012); Pagliari et al. (2007); Blaya et al. (2010)). This enthusiasm has been supported by projects to promote access to electronic health records (including (Scottish Health Informatics Programme (SHIP); Ford et al., 2009)) as well as cross-sectoral data-linkages to explore health outcomes from multiple angles simultaneously (McCafferty et al. (2010) for example).

This vast increase in data sources has also contributed to concerns over the role of data in society, and the potential threats to privacy that are introduced as technology advances in realms such as social media (Szongott et al., 2012), the digital economy (Tene and Polonetsky, 2012) and medicine (Steinbrook, 2008). These concerns are not new however, for example Garfinkel’s discussion of the creation of a National Data Center in the US in the 1960s raises similar concerns about potential attacks on personal privacy (Garfinkel, 2000). Solove reaches even to more abstract concept of privacy that has always been described as under attack, citing Nelson (2002) Solove notes that “[p]rivacy, it seems, is not simply dead. It is dying over and over again” (Solove, 2008, 5).

Therefore, the enthusiasm for data has been tempered by the practicalities of providing access, and the potential threats to the rights of citizens. The overarching difficulties of realising the potential of these new data sources have proven to be technical, ethical and sociological as each discipline contributes to its own growing literature, for example; on the affordances of privacy preserving techniques (Agrawal and Srikant, 2000), new legal governance frameworks for emerging data sources (Laurie and Sethi, 2012), or the sociology of surveillance (Lyon, 2003). However, little attention has been paid to the interface between these spheres and the technical operationalisation of ethical frameworks for example.

In parallel to the data deluge, has been the development of e-Science, which since its inception in the UK in 1999, has widened in scope from its early focus on grid computing and nuclear physics to the more general "application of computer technology to the undertaking of modern scientific investigation, including the preparation, experimentation, data collection, results dissemination, and long-term storage and accessibility of all materials generated through the scientific process" (Bohle, 2013a). The e-Science approach is an interdisciplinary one, as it combines multiple disciplines with computer science to facilitate cutting-edge research workflows or make current workflows more efficient. As such, an e-Science approach could be used to interrogate the interconnected spheres of technology and ethics in the big data and privacy context. This could potentially provide new workflows or enhance existing ones. Especially since e-Science approaches have already been applied with positive effect for data access and data management problems in other fields (Doherty et al., 2010).

Returning to the data themselves, in the wider UK research context, they are drawn from myriad different sources, including a rise in administrative data being made available for research purposes. Administrative data are routinely collected by government and other organisations for the purposes of providing some form of service or function; they were not collected for research purposes in the first instance. These data have advantages for both data owners and analysts. The cost of primary data collection can be prohibitive and this limits the number of studies that can successfully obtain funding. A clear example of this cost reduction principle in practice is the review of the UK Census. The "Beyond 2011" programme evaluated the potential for replacing the UK paper census with data drawn, at least in part, from administrative sources. This is already an approach taken by other countries including Germany, the Scandinavian countries, the Netherlands, Switzerland, Austria and Israel (Office for National Statistics, 2014, 6). In their final report the project recommended moving to an online census, but the direction of travel was laid out so that more research into the potential for administrative data to replace the census was also recommended (Office for National Statistics, 2014).

In addition, UK research funding councils have placed great emphasis on the secondary use of data. The Economic and Social Research Council (ESRC), for example, makes available funding specifically for this type of data initiative[1]. Furthermore, in 2013 the ESRC announced funding for four administrative data research centres[2] (ADRC). The Administrative Data Task-force that recommended the establishment of the ADRCs cited the benefits of reduced costs for data production, increased efficiency through data re-use and the faster production of policy relevant research (Administrative Data Taskforce (Technical Group), 2013). Both the approach of the ESRC and the Beyond 2011 programme also resonate with the open data and transparency agenda that has gained traction with governments and public bodies. Through open government initiatives, such as those discussed in Janssen et al. (2012), an increasing number of datasets are being offered to the public.

In areas like health research where datasets are likely to contain particularly sensitive personal data, data controllers have taken a more cautious approach to data access, however access to secondary care records and clinical trials data is becoming much more common. The National Health Service (NHS) in Scotland have the Electronic Data Research and Innovation Service (eDRIS) which provides researchers with data such as the Scottish Morbidity Record as well as the technical infrastructure for data-linkage and a secure-setting for data analysis. Similarly in England, initiatives have been undertaken to increase information sharing and access to data for research. For example, the 'care.data' programme (NHS England) primary care data from GP records could become available to researchers.

Further to this, serious attention has been paid to the creation and curation of digital health records and the infrastructure needed to store and analyse them. Projects such as the Secure Anonymised Information Linkage (SAIL) Databank provide the technical infrastructure to house, link and analyse large volumes of data using new methodologies and computing resources (Ford et al., 2009; Lyons et al., 2009). Similarly the governance literature and procedures have been reviewed and updated to cope with new data sources. In this vein, the Scottish Health Informatics Programme (SHIP) published an updated governance framework for working with health related data (Laurie and Sethi, 2012).

Governance procedures, like those of SHIP, and data management processes have attempted to keep pace with the expansion of data sources. However, there is some lag between the two. This gap is also subject to shifts in societal attitudes to privacy and research. As has been said, data privacy is not a new phenomena but with the vast quantities of digital data that are created everyday it has recently occupied a more central position in the social conciousness than in previous years(Halstuk and Chamberlin, 2006). As this phenomena is not new, the techniques and approaches to protecting privacy and confidentiality have developed over time, particularly through the workflows of government statisticians. For

[1] Details of the Secondary Data Analysis Initiative at http://www.esrc.ac.uk/research/skills-training-development/sdai/

[2] The announcement can be found at http://www.esrc.ac.uk/news-and-events/press-releases/28673/the-big-data-family-is-borndavid-willetts-mp-announces-the-esrc-big-data-network.aspx

example, consider the discussion in Dalenius and Reiss (1982) of data swapping to prevent disclosure in microdata releases and tabular outputs. What is new is the expectation that data controllers should provide access to data and be able to handle multiple queries from different users at scale and still fulfil their responsibility to protect the privacy of data subjects. In order to achieve this, the methodologies and workflows of data controllers need to be adapted.

Statistical disclosure control (SDC), defined as a"set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations" (Elliot et al., 2005a, 12), provides data owners and controllers with the tools to fulfil their obligations in terms of privacy and confidentiality for their data subjects, as well as in terms of data access for their users. Measures of disclosure risk range greatly in their complexity from simple thresholds to complicated and highly configured statistical operations such as those proposed by Li and Li (2009) that utilise information theory.

However, despite the body of literature on SDC, including measures of risk and data utility, it is still not well understood how data controllers operationalise their own value or ethical policies by translating privacy requirements into practical applications and how these decisions impact on the research process. Rules for data privacy are often stated, for example the use of thresholds for low cell counts in data tables, however the principles that underpin such rules are not often stated.

In the context of the gaps identified in the literature above the research aims are stated below. It should also be noted that in common with the majority of the projects discussed above, the focus of this thesis will be the UK and, in particular, the Scottish context (for reasons of data access discussed in Chapter 3), however comparisons have been drawn with other countries where the contrast adds to the discussion. For example, the wider European context, as well as that of the US and Canada which are included in Chapter 2. The Scottish context was also chosen because of the pragmatic motivations that underpin this thesis. This research was funded through a specific call by the Data Management through e-Social Science project[3] (DAMES) and the Scottish Health Informatics Programme[4] (SHIP), as such it was the funders intention that this research would contribute to their work on the collation, management, dissemination and analysis of personal data, specifically in the area of systems for mitigating unauthorised disclosure.

1.1 Research Aims

The aim of this research is to shed light on the practices of SDC by examining the theoretical understanding of risk, specially risks to privacy and confidentiality alongside the responsibil-

[3] See http://www.dames.org.uk/ for details of the project

[4] see http://www.scot-ship.ac.uk/

ities of data controllers. This will attempt to address the gap in the literature by constructing a framework for SDC as an interplay of scientific and value judgements in the mitigation of disclosure risks.

As these judgements fundamentally affect the quality and utility of the data provided to researchers; and this utility is exclusively discussed in terms of its analytical utility when compared with the raw data, (i.e. does published data provide the same results when subject to statistical tests? For example, see Purdam and Elliot (2007)), it is also the aim of this research to broaden the discussion on operationalised disclosure control to encompass the practical experience of users and data controllers. The aim here is to show why it the user experience is relevant and how it can be operationalised in the development of SDC workflows.

Equipped with these more conceptual instruments, it is also the aim of this research to provide a practical example of this type of development work in the form of a software tool-kit that facilitates the SDC processes of data controllers. The explicit aim here is to provide a tool-kit that enhances current workflows, can be integrated into existing systems and is compatible with the virtual research environments typical of e-Science projects.

1.2 Thesis Structure

Chapter 2, provides a theoretical, legal and ethical framework for this thesis to ground future arguments on disclosure control as an interdisciplinary undertaking. This is done first by exploring the rules that govern the collection, disclosure and use of personal data in a legal sense (section 2.1) by examining the data protection legislation in the UK, the EU and the US with particular reference to the use of data for health research. Having established the rules, section 2.2 seeks to scratch under the surface of these rules and provide some perspective on privacy and the concerns that underpin those rules. This is done with reference to Garfinkel, Nissenbaum and Solove. In section 2.3, a framework for disclosure risk as a meta-risk arising through the processes of reflexive modernity is developed drawing on the work of Ulrich Beck. In this section Beck also poses the interdisciplinary challenge for the estimation of risk that is pursued throughout this thesis. Section 2.4, continues the theoretical work by attempting to unpick the philosophical reasoning that lies beneath the fears of technology, unauthorised disclosure and invasions of privacy, through Heidegger's work on enframing, acceptance and disclosure. Moving back from the theoretical to the practical, section 2.5 discusses proposed changes to EU regulations that demonstrate a reflexive shift in the perceived risk to privacy that new and emerging data practices reveal and the impact this could have on future research. Chapter 2 concludes with a discussion of the public and private spheres of data disclosure and different context specific ways in which individuals actions often differ from the abstract concerns about privacy that are expressed earlier in the chapters narrative.

Chapter 3 details the interdisciplinary, pragmatic-sequential research design that frames how this research was conducted as well as setting out the e-Science methodological context within which this research takes place (section 3.2). Specific research methods are described in sections 3.3 – interviews, and section 3.4 – the case study. Also in this chapter *k*-anonymity as a method for disclosure control is introduced (section 3.5) and the use of quantitative analyses is described as a measure of data utility (3.6). Lastly, the development of the software tool-kit used in later empirical chapters, NIAH, is discussed in section 3.7.

In Chapter 4, a review of the literature on statistical disclosure and disclosure control is presented. Section 4.1 provides a brief insight into the historical development of disclosure control since the 1960s and the early publication of microdata in the US as well as discussion during the late 1980s and 1990s in the UK regarding Census outputs. Sections 4.2, 4.3 and 4.4 cover the literature on the measurement of disclosure risk, methods for disclosure control and approaches to data utility. These three sections are interwoven with a case study approach that provides an example of the methods discussed in the literature using the Scottish Health Survey dataset.

Chapter 5 sets out the argument for using the framework developed in Chapter 2 to expand the scope of the literature on statistical disclosure control to include a further perceived risk — that data are made available as part of society's reflexive process of introspection and risk estimation, yet the controls put in place to balance the meta-risk of statistical disclosure create their own set of risks that might preclude the efficient use of these data. To mitigate these new risks, this Chapter makes the case for the user experience in statistical disclosure workflows. In particular this is done by exploring issues of data access (section 5.1) and the treatment of statistical outputs (section 5.2). These ideas are approached through an interdisciplinary relationship of qualitative concern for the user experience and the technological accordances offered by development in virtual research environments (section 5.3).

In Chapter 6, an attempt is made to take the ideas developed in Chapter 5 and deploy them through the development of a software tool-kit, NIAH. This includes the use of a small qualitative pilot study that captures the data users perspective which is used to incorporate aspects of the user experience into the develop work (section 6.2). The new tool-kit is compared with existing approaches to the SDC workflow that are in use by the Scottish Government and NHS Scotland (section 6.3), before the Chapter concludes with a preliminary exploration of the potential for integrating NIAH with existing systems and the types of virtual research environment discussed in Chapter 5 (section 6.4).

In keeping with the pragmatic objectives and research design, Chapter 7 demonstrates how the tool-kit and the approach discussed in the preceding Chapters can be applied to data in a ‘real-world’ context. Section 7.1 shows how the Scottish Government have adapted the NIAH tool-kit and integrated it into their own systems allowing analysts to use it through a simple web interface. In addition, section 7.2 returns to the case study approach first used in Chapter 4 and details the use of NIAH in an SDC workflow using real, ‘raw’ data from the field of education.

The final Chapter in this thesis, draws the discussions of the above chapters to a close and concludes whether, and to what extent, the challenges set out in Chapter 1 have been addressed. In particular, section 8.1 pays attention to the interdisciplinarity of the overall project and the contribution to the existing literature on disclosure control. Section 8.2 sets out the potential spaces this research creates and what future work might take place as a result.

Chapter 2

Privacy & Confidentiality: A legal, ethical and theoretical understanding

This chapter looks at the underlying ethical issues associated with large data-sets on individuals primarily through Beck's sociology of risk and Heidegger's concept of 'enframing', while acknowledging the literature on privacy and the public's fear of the invasion of privacy. It attempts to draw together a theoretical grounding as to why societies go to great lengths to preserve personal privacy in research data, as well as recognising the interdisciplinary collaboration required to make meaningful advances in the field of disclosure control and real-world solutions to privacy concerns. The theoretical and ethical basis for this research in statistical disclosure control will be explored. In doing this some of the tensions between the public good and personal privacy will be exposed. From this introduction the focus shifts to legal frameworks that seek to protect the right of citizens in not having their personal data disclosed unlawfully; this will provide a contemporary backdrop from which to explore the fears and insecurities that underpin these laws and their relation to statistical disclosure and the use of personal data for research. From there the work of Ulrich Beck and Martin Heidegger will be used to delve deeper into the theoretical concepts that characterise the privacy concerns of citizens in the modern world where data are growing in scope and size alongside the analytical technologies to harness them. Once this groundwork is established it will be used to frame discussions about changes to the legislative frameworks set out in the beginning as well as explore the apparent juxtaposition in the attitudes of individuals toward data sharing and linkage in the public and private sectors.

The position of this research is that it is not enough for societies to have rules on privacy, and statistical disclosure control as a result, without understanding the reasons that underpin those rules. In order to ensure the rules keep pace with the views of a society it is important to capture the fears of the society (see the discussion at the end of section 2.1) as well as harnessing philosophical literature that can be used in an attempt to explain these fears. In the pursuit of this aim, Beck's sociology of risk is used injunction with the further philosophical depth of Heidegger's concern that technology has changed how people relate to each other as human beings. This will provide the necessary foundations to understand, and take account of, the public good versus personal privacy, and disclosure risk versus data utility balance in contemporary debates on anonymisation and statistical disclosure control. In addition, this theoretical exploration will position disclosure control within an interdisciplinary context where ethical, mathematical and technological discourses are needed to solve contemporary issues in the field.

The protection of personal data held by institutions is a serious concern for the public and for those charged with the data's protection. This seriousness is drawn from the emphasis society as a whole places on privacy and security. Society's commitment to privacy is apparent in the highest echelons of political institutions; a right to a private life is enshrined in Article 8 of the European Convention on Human Rights (ECHR). Attempts to store and share more and more detailed personal data are met with fierce criticism. This is despite the protestations of governments that this increase in data is in the public interest. As such, public apprehension is well documented; for examples of this in the field of data-linkage and research see Scottish Government (2011a); Aitken (2012). In the proceeding section the legal frameworks mentioned above are discussed in further detail.

2.1 Legal Frameworks

Within the UK, aside from the overarching ECHR, personal data is subject to the Data Protection Act 1998 (DPA) which stipulates how data are to be treated by data controllers and processors. The definition of what constitutes personal data is contentious; the Information Commissioner's Office (ICO) have issued guidance that includes an eight step decision tree to help data controllers decide if their data should be considered personal or not. For the purposes of this research, the first two steps provide a relevant definition of personal data. First, if an individual is identifiable from the data, and here identifiable means directly identifiable by name or date of birth, for example. Second, if the data are not directly identifiable does the data 'relate to' an individual. This is a form of indirect identification, that the ICO define as: "Data which identifies an individual, even without a name associated with it, may be personal data where it is processed to learn or record something about that individual, or where the processing of that information has an impact upon that individual" (Information Commissioner's Office, 2012, 9). Defined in this way, personal data can cover anonymised data, which is the focus of this dissertation. The DPA sets out eight data protection principles. Principle 7, in particular, forms part of the basis for this research into statistical disclosure control because of the emphasis it places on the data controllers to ensure data are not disclosed unlawfully.

> DPA Principle 7: “Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data” (Data Protection Act 1998, Schedule 1).

This legal framework that puts the onus on data controllers to process sensitive personal data fairly and prevent the loss of personal data is not unique to the UK. As has been discussed, the right to privacy is a common commitment across Europe through the ECHR. Legislative protection of personal data can also be seen in the United States. Although the US does not have one piece of coverall legislation such as the DPA, the right to privacy and the protection of personal information is upheld through various statutes. In the health field the Health Insurance Portability and Accountability Act 1996 (HIPAA) requires institutions engaged in health care and related activities to prevent the “wrongful disclosure of individually identifiable health information” (US Congress 1996). In Canada, the Personal Information Protection and Electronic Documents Act 2000 (PIPEDA) provides a similar regulatory framework to the UK with an ombudsman responsible for ensuring that data controllers are compliant.

Aside from setting out the legal obligations of the data controllers, all of the above legislative frameworks acknowledge the public versus personal tension in some way. In the UK this is done through a series of exemptions for activities for which the public interest is considered to outweigh the personal. For example, the DPA carries an exemption for the use of data to safeguard national security and an exemption for data used for research and statistical purposes. This public versus personal balance has been tested in court. In England & Wales High Court (2011) the Department of Health (DoH) contested an Information Tribunal’s decision to force the disclosure of abortion statistics. The DoH argued that the disclosure “would create a real risk of patients being identifiable” and would infringe their right to privacy. However, the Court found in the Information Commissioners Office’s (ICO) favour and ordered the DoH to disclose the information. In coming to this decision the Court reviewed the Information Tribunal’s earlier conclusions which included an assessment of the benefits of disclosure in this case. The Tribunal saw seven areas in which the data would serve the public interest:

> “checking compliance with the Abortion Act; enabling public scrutiny of the way abortion law was applied; ensuring accountability in relation to medical practitioners; providing external checks and balances to Department of Health scrutiny; identifying trends; planning healthcare services, including monitoring the rates of foetal abnormalities; and informing public debate” England & Wales High Court (2011).

On the other side of the balance was the potential risk to the patients and the potential harm this could cause. The Tribunal had taken expert evidence from statisticians including guidance from the Office for National Statistics. They concluded "that the possibility of identification by a third party from these statistics was extremely remote" (cited in England & Wales High Court (2011)), and thus the balance was tipped in the public interest's favour.

Similarly, in the US, HIPAA provides exemptions and alternative access regulations for matters concerned with the public good and states that it "strikes a balance that permits important uses of information, while protecting the privacy of people who seek care and healing" (US Congress 1996). As with the UK legislation, it provides exemptions for public health activity and research, allowing data controllers to disclose personal information without the individual's consent provided the activity meets certain criteria. Unlike the ombudsman systems of the UK and Canada, under HIPAA, violations can be criminally prosecuted and recent cases against workers in the healthcare industry have resulted in custodial sentences for unlawful access of medical records.

This is a brief overview of the legal requirements for the handling of personal data. The fact these regulations exist is testament to the importance placed on the right to privacy and the protection of personal information. However, there is an attempt to capture the balance between the public interest and personal privacy, specifically in England & Wales High Court (2011). This case saw the judiciary balance the specific disclosure risks of data against the possible public benefits. What the legal frameworks do not do is give any reasoning at a theoretical level, as to why these laws are deemed necessary or unpick the calibration of these balances. In the next sections an attempt is made to do this primarily drawing on concepts developed by Beck and Heidegger.

2.2 Perceptions of Privacy and the Public's Fears

Having established that this chapter is something of a theoretical justification for statistical disclosure control and the protection of an individual's privacy, it is important to understand the fears that the public have surrounding the potential risks to privacy that underpin the rules set out above. Before proceeding with the theoretical substance of these sections it is worth acknowledging the literature that documents some of these fears and the perspectives they offer on privacy. Although the dictionary definition of privacy is"the state of being free from the attention of the public"[1], Garfinkel (2000) argues that the term privacy is somewhat limited in its ability to capture the wider concept. For Garfinkel privacy is not simply about 'hiding things' but also encapsulates 'self-possession, autonomy, and integrity' (Garfinkel,

[1] Oxford Dictionary Definition of Privacy http://www.oxforddictionaries.com/definition/english/privacy accessed 08/08/2015

2000, 4). Garfinkel also offers, in his discussion of the 'database nation', a description of the fears attached to the centralisation of data that resonates with the opening-up of data and data-linkage that features in contemporary debates and resembles the fear of a modern digital Panopticon (Betham, 2008). The National Data Center proposed by the US Government in the 1960s would centralise data from the federal agencies into one system, however the fear that one single databank would place too much power in the hands of bureaucrats and reduce individual autonomy was articulated strongly enough to prevent the data centre being built.

In the contemporary context this argument against granting power to a single authority has parallels with the discussions surrounding data linkage, and the ability of organisations to create hybrid data sets on individuals by linking distributed sources. It should be noted that although Garfinkel is critical of the National Data Center, he is also critical of the alternative 'nation of databanks' that took its place, specifically the potential for error. His critique provides an interesting commentary on the role of government and how by not proceeding with central controls they absconded from their responsibility to protect the rights of citizens and allowed data, and decisions about how that data are processed, to be siloed and carried out in secret by private organisations. This critique is worth bearing in mind when we return to the public versus private dichotomy in section 2.6.

The public's fear is not limited to the centralisation of data, but also the collection of data on individuals or surveillance. Nissenbaum takes up this aspect of privacy in her exploration of the concept of public surveillance and discusses more specific cases for concern than Garfinkel (Nissenbaum, 2004). Highlighting three guiding principles that feature in public and legal discourse in the US: "(1) limiting surveillance of citizens and use of information about them by agents of government, (2) restricting access to sensitive, personal, or private information, and (3) curtailing intrusions into places deemed private or personal." (Nissenbaum, 2004, 125) we can see that disclosure control would fit neatly within the second principle. Nissenbaum attempts to move the discourse on from the usual public versus private dichotomy by setting out the argument for contextual integrity, which sets two tests and should either one be not be satisfied then privacy has been undermined. These two tests rest on the ideas of appropriateness, in a particular context should data be revealed, and the distribution of data, does the flow of information conform to the norms associated with that context. These two tests will be revisited in section 2.6.

Also advancing a different perspective on the concept of privacy is Solove (Solove, 2008; Strandburg, 2006). Focusing on the future of privacy in the digital age, Solove draws on Orwellian (big brother-like centralisation of data under a central authority) and Kafkaesque (the subjection of individuals to bureaucratic processes) metaphors to illustrate the problem of modern databases (Strandburg, 2006). There is some similarity here with the concerns of Garfinkel over what is at stake, i.e. individual autonomy. Solove suggests that distributed databases 'disempower' individuals not through the deliberate, malicious actions of an authority but: “there is a web of thoughtless decisions made by low-level bureaucrats, standardized policies, rigid routines, and a way of relating to individuals and their information that often becomes indifferent to their welfare” (Strandburg, 2006, 10). In addition, Solove raises the fear of identity theft as a consequence of growth in personal data collection and distributed storage. The risk of identify theft has increased as our digital identity is essential for so many day-to-day processes and the victims of this suffer the “pollut[ion of] their digital dossiers” (Strandburg, 2006, 11).

The discussion above provides insight into approaches to privacy and the concerns of the public. Perceived threats to privacy arise from the increasing volume of personal information that is collected in central government systems as well as the private databases that Garfinkel and Solove critique. These fears, at their heart, are concerned with the autonomy of individuals and their control over their own information, the loss of which conjures images of totalitarian, bureaucratic dystopia or a future of identity theft and cyber-crime where our digital identity is synonymous with our true self. This literature also provides different perspectives on privacy and how to conceptual the public versus private balance, Nissenbaum’s contextual integrity in particular reminds us that the contexts in which information is collected and processed are complex and subject to societal norms. These discussion provide further background context, alongside the description of legal frameworks above, that should be carried forward into the proceeding sections and will be explicitly revisited in section 2.6.

2.3 Ulrich Beck and the Risk Society – Statistical Disclosure as Meta-Risk

Having discussed approaches to privacy and some of the fears over threats to privacy and the rights of individuals, the concept of risk in social theory is now considered as the significant first step from the contemporary, pragmatic, world into the theoretical and philosophical. It would be remiss to not first look for a contemporary framework of risk that can be adapted for the purposes. Ulrich Beck’s work on the risk society, developed during the 1980s with a backdrop of growing concern over environmental risks (and subsequently the Chernobyl disaster), provides a theoretical framework for a reflexive modernity that is in a constant state of flux and uncertainty, assessing, mitigating and re-assessing perceived risks and making value judgements about them. This framework is set out below starting with his definition of risk, and then it is applied to the context of statistical disclosure. The application of this framework sets statistical disclosure in the position of a meta-risk, or risk born of the risk assessment (reflexive modernity) process itself. This allows for the problem of disclosure control to be discussed in terms of Beck’s sociology of risk, which at its core sets the evaluation of these risks as a interdisciplinary challenge that must navigate both scientific and social rationality.

Beck in *Risk Society* defines risk as “a systematic way of dealing with hazards and insecurities induced and introduced by modernization itself” (Beck, 1992, 21). Underpinning this definition is Beck’s approach to modernisation, which is described as:

> “[...] surges of technological rationalization and changes in work and organization, but beyond that includes much more: the change in societal characteristics and normal biographies, changes in lifestyle and forms of love, change in the structures of power and influence, in the forms of political repression and participation, in views of reality and in the norms of knowledge. In social science’s understanding of modernity, the plough, the steam locomotive and the microchip are visible indicators of a much deeper process, which comprises and reshapes the entire social structure” (Beck, 1992, 50).

One could include the current expansion of data as one such visible indicator which is reshaping the social structure. As we have seen in the preceding section, new technological innovation in the collection and storage of data have questioned and subsequently reshaped societal attitudes (reframing of existing boundaries – section 2.5) and our relationship with governments (data protection legislation – section 2.1), other institutions (data sharing and user agreements – section 2.6) as well as each other (the rise of social media for example). Also from Beck’s concept of modernisation, it is important to note that the modern era is distinguished by the rise of ‘manufactured’ (consequences of deliberate human agency) rather than ‘external’ (natural disasters for example) risks (see (Beck, 1992, 183), but also (Giddens, 1999)).

Having laid some of the groundwork the focus moves back to Beck’s definition of risk, which can be broken down into two parts. First, ‘dealing with hazards and insecurities’ i.e. the risks that societies face. At this stage in the definition this could apply to past epochs as no distinction is made about the type of risk. Second, “these hazards are induced by modernisation itself”, seizes on the point made above, these hazards are manufactured through the very process of modernisation. Further, as a constant reflexive process, modernisation does not just induce risks but through this modernity it also mitigates against these risks.

A note of caution should be sounded before this discussion of risks proceeds, as Beck notes, risk is not the fulfilment of the potential, it is not the occurrence of a hazard but the potential for such a consequence: “The concept of risk thus characterizes a peculiar, intermediate state between security and destruction, where the perception of threatening risks determines thought and action” (Adam et al., 2000, 213). Similarly, Armstrong articulates this in the health context: “Risk has no fixed nor necessary relationship with future illness, it simply opens up a space of possibility. Moreover, the risk factor exists in a mobile relationship with other risks, appearing and disappearing, aggregating and disaggregating, creating spaces within and without the corporal body" Armstrong (1995, 401).

What position does data, and specifically big data, occupy in this modernity? As Kerr and Earle suggests it is a product of the risk society which offers predictive powers and "opportunities like never before to anticipate future needs and concerns, plan strategically, avoid loss, and manage risk" (Kerr and Earle, 2013, 66). Reflexive modernity has been characterised by the production and processing of ever-increasing volumes of data to capture, assess and mitigate perceived risks. To return to Beck, he describes the sociology of this risk as "a science of potentialities and judgements about probabilities" (Adam et al., 2000, 213). Data facilitates these 'judgements' and arguably forms part of the science of potentialities (see the alignment with Armstrong's medical risk factors) but Beck also introduces the idea of subjectivity through the use of the term judgement, who makes these judgements?

Beck recognises the importance of information in the risk society, and the power that those that have the information wield. However, he suggests that this reflexive modernity, in which society focusing in on itself "throws all of the [...] basic principles into flux" (Beck et al., 2003, 2) - such as the dominance of the nation state, reliable welfare systems and the nuclear family – is also leading to a democratisation of risk. Another tension can be identified here between scientific and social rationality. Beck argues that "in the definition of risks the sciences' monopoly on rationality is broken." The perception of risk by the public is a challenge to scientific authority, whose ability to claim expert status is questioned: "Where and how does one draw the line between still acceptable and no longer acceptable exposures?" The problem for science in assessing risk, Beck argues is that "one must assume an ethical point of view in order to discuss risk meaningfully at all" (Beck, 1992, 29).

The subject of this research, disclosure control, can be positioned within these concepts. If the collection, processing and analysis of data is an intrinsic element in society's reflexive process of risk identification, monitoring and mitigation, statistical disclosure – is a risk produced by this process of reflexivity itself. Webster (2002) describe such risks as 'meta-risks' or the 'riskiness of risk assessment'. Therefore, disclosure control forms a response to this meta-risk of disclosure. Establishing disclosure as a meta-risk allows us to apply the sociology of risk and the balance of scientific and social rationality, within the contemporary social power structures. This tension can be seen in how data are viewed as personal, or sensitive, and where relative anonymity is pursued through the manipulation of data to mitigate perceived risks to privacy.

To take Beck's line of argument, the ethical dimension is important to consider in any handling of data or analysis of disclosure risk. It also justifies to some extent why attempts as 'experts' to fully automate assessment and management of risk are ill advised (see Solove on Kafkaesque bureaucracy) or fall into technological determinism (see Heidegger on the technological way of thinking in section 2.4). Instead, the process should be democratised by providing tools that allow the sciences, the public and government to reach an agreement on what is 'acceptable exposure' in the pursuit of a specific 'public good' — whether that be identifying and analysing patterns of social inequality, or searching for new medical advances.

It is also important to note, before leaving the reader with the notion that science is redundant in the sociology of risk, that Beck's work does not propose an uninformed dictatorship of public perception. It recognises that risk is a quasi-scientific interplay of mathematical probability and ethics, a combination of factual and value claims. Therefore the statistical models devised to estimate risk cannot be separated from questions about what society considers to be acceptable or unacceptable. For research into disclosure control, as an exploration of meta-risk, Beck sets an interdisciplinary challenge: "[Risk statements] can be deciphered only in an interdisciplinary (competitive) relationship, because they assume in equal measure insight into technical know-how and familiarity with cultural perceptions and norms" (Adam et al., 2000, p215). This challenge is seen in the statistical disclosure control literature, which with few exceptions, notably Duncan et al. (2011)), approach disclosure control from a purely technical or statistical direction. This has resulted in a series of advancements in anonymisation techniques and how the risk of re-identification is assessed (see Elliot et al. (2005a), Reiter and Mitra (2008) and Dankar et al. (2012)). However, these advancements have occurred in relative isolation from the real-world workflows that combine the ethical and technical.

As has been described above, Beck provides a set of useful definitions of risk and modernisation that can be used to describe the big data risk society. By exploring the reflexive process statistical disclosure can be positioned as meta-risk born of the process itself, and therefore the estimation of risk must be undertaken in the interdisciplinary relationship Beck notes above. This framework supports the interdisciplinary approach set out in this thesis through the interplay of the methods set out in Chapter 3. However, Beck's analysis does not capture the 'why', as in why at a philosophical level statistical disclosure is considered a risk at all. Some of this is hinted at in the above section on approaches to privacy. However, in order to go deeper and inspect the substrata beneath these foundations, a more philosophical understanding of society's relationship with technology and its influence on how people relate to each other is needed. For this another German writer's work is considered; the narrative of Martin Heidegger will be applied to the contemporary context.

2.4 Martin Heidegger, Enframing, Aletheia and Gelassenheit

Martin Heidegger, the 20th Century German philosopher, is considered to be one of the most important philosophical thinkers of that century, despite his extremely controversial support for National Socialism. Being and Time is Heidegger's best known work and is said to have influenced a number of leading thinkers that followed (Heidegger, 1962). Normally associated with the schools of phenomenology, hermeneutics and existentialism, it might seem odd to call upon Heidegger to provide insight into the basis for statistical disclosure control and the balance of the public and private spheres. However, despite Heidegger's deeply philosophical writings on the nature of being, he also engaged in writing about subjective experience and its connection to contemporary thought. He also fills a gap in the earlier consideration of Beck; as will be shown, Heidegger captures something of the fear and insecurity that surrounds and permeates the 'big data' revolution.

In addressing this revolution as the deluge of digital data that is harnessed to allow individuals, organisations and societies to make sense of their world through actions like mass surveillance, the deployment of 'smart' sensor networks or the electronic tracking of transactions; Heidegger and ontological philosophy offer concepts to aid in the description of the world in which disclosure control operates. 'Erschlossenheit' or world disclosure is Heidegger's concept for how human comprehension of objects is developed through interaction – the day-to-day use of an object and interactions with other people gives it meaning (Heidegger, 1962). It is also important to note that this ontological description of an object and its attributes is the theoretical basis for the idea of ontologies in the field of information science (for example see Gruber (1993)) and is used extensively in big data technologies that use linked data strategies or semantic web approaches.

Given Heidegger's own subjective experience as witness to the Second World War, it is not surprising that he would write about the effect of technology on the relations between human beings, and it is this work on the philosophy of technology that will be drawn upon. Born in 1889, throughout his lifetime he witnessed the rapid advancement of technology toward a state of total war. In an essay written in 1950, *The Question Concerning Technology*, he expresses his disdain for the way technology has converted all things into resources on stand-by, waiting to be manipulated, with a particularly disturbing tone:

> "Agriculture is now a motorized food industry, essentially the same as the manufacture of corpses in gas chambers and extermination camps, the same as the blockade and starvation of countries, the same as the manufacture of hydrogen bombs" (Heidegger, 1978).

Polt (2003) argues for looking beyond the superficial shock that this quote generates, and for readers to recognise the challenge to ask whether or not these processes are all 'essentially' the same. This requires reflection on the technological approach to the world that dominates the modern era. Of course these actions are not 'essentially' the same, but this description sets the scene for Heidegger's conceptualisation of technological thinking. Polt also associates this view of technology with Huxleyan (as opposed to Orwellian) visions of the future; a future where all people and things are heavily conditioned resources existing in an artificial, safe and happy society that spurns the individual and any attempts to subvert this controlled existence. Heidegger uses two labels for this managed resource view of the world, Machenschaft (machination) and Ge-stell (enframing); enframing will be used from this point onward (Polt, 2003, 142 & 172).

Echoes of 'enframing' can be seen in the public's apprehensive approach to the use of their personal data. When asked, people raise questions about what data will be used, who will have access, and will it be secure. They discuss their data as if it were a commodity and a resource to be managed. They also express a view on authorisation and ownership. Consent, to the use of their data, is seen "as being a means of respecting individuals" (Aitken, 2012, 2) and is often considered essential, despite the practical problem of collecting such consent for the use of data collected, stored and analysed in different physical locations, at different times and for different purposes.

Heidegger also posits that the technological way of thinking is driven for its own sake. Enframing transforms everything into resources, a power base for further development of more technology and so on (Zimmerman, 1990, 248). This 'because we can' reasoning for the accumulation of resources also resonates with public fears over the use of their data. Contemporary debates over the draft Communications Bill in the UK, which would see an expansion of the state's power to intercept, collect and store personal communications data, has been criticised from a position of 'just because we can, does not mean we should'. For example, Liberty, the civil liberties campaign group in the UK, responded to the draft bill by concluding that the bill "not only exacerbates human rights concerns but also makes clear that this proposal is about extending rather than maintaining the ability of the State to monitor communications" (Liberty, 2012). Liberty in this response questions the enframing, technological thinking and concludes that just because we can collect more data does not mean we should. In this statement's shadow, Beck can be seen separating the perceived risk of the Communications Bill into factual statements, 'we can collect it', and value statements 'we shouldn't collect it'.

For Heidegger, this enframing of personal data would likely have been another step toward his fatalistic view of technology; society's unquestioning use of technological solutions to perceived problems. However, this chapter is concerned with tensions and balances and Heidegger's enframing is, in the personal data context, set against his concept of Mitsein

(Being-with) – being in the world together with one another. Olafson (1998) succinctly describes Mitsein:

> “The plain fact is that, at every step we take in whatever endeavour the lives may be devoted to, we are constantly supplementing the own observations and the own recollections from a common fund to which we call contribute. So profound is the dependency that the idea that we could draw a line demarcating what we ourselves have supplied out of the own resources from what we have drawn from the common stock is more than a little problematic” (Olafson, 1998, 26).

So between enframing and Mitsein there is a tension. Heidegger criticises the enframing of existence because it causes the relationship as human beings to be dominated by a technological way of thinking, which in turn reduces everything to resources to be operationalised, manipulated and optimised. However, he acknowledges through Mitsein that individuals make sense of the world as social beings, through shared subjective experiences and a common pool of knowledge. As an attempt to describe the use of personal data in relation to this tension, the public versus private dichotomy set out in the beginning of this chapter re-emerges. Mitsein or ‘being-with’ others can be defined as the pooling of collective experience – and in this context, this collective pooling takes the form of data stored for processing. This data allows individuals to ‘discover the world’ together as Heidegger might have described it, through the identification of patterns or trends in the data. This interpretation of Mitsein provides the public weight in the balance between public versus private good; data are questioned, interrogated and analysed in order to help make sense of the world. The personal weight, scales with the enframing threat the data represents – how much of oneself is given up as resource (or perhaps how detailed are the data and who has access)?

Heidegger also realises that this tension cannot be resolved simply, it is not Mitsein or enframing; to use technology or not. Instead he uses the term Gelassenheit or ‘letting-be’ a term in this context described by Polt as an alternative to a crude choice between action or passivity (Polt, 2003, 172). Heidegger suggests technology can be used without being dominated by it, or by succumbing to enframing and the technological way of thinking. His central concern is that the use of technology should not subsume an individual’s ‘being’ or their ‘being-with’ others, it should not affect their ‘core’. He expresses this in Discourse on Thinking:

> “It would be foolish to attack technology blindly. It would be short-sighted to condemn it as the work of the devil. We depend on technical devices; they even challenge us to ever greater advances. But suddenly and unaware we find ourselves so firmly shackled to these technical devices that we fall into bondage

> to them. [...] We can affirm the unavoidable use of technical devices, and also deny them the right to dominate us, and so to warp, confuse, and lay waste to the nature" (Heidegger, 1969, 53,54).

To return to the contemporary world again, the interplay of Heidegger's Mitsein versus enframing tension has been described, and it has been shown how this relates to the public versus private balance when personal data is used for research. Therefore, to extend the argument beyond these crude dualities, Heidegger's concept of 'letting-be' must be incorporated. This is constituted as a cautious acceptance of technological advancement (in this case greater analytical, linkage and storage capacity of data). It is cautious engagement that introduces the opportunity to calibrate the public versus personal balance to prevent this technology from affecting one's core as Heidegger suggests. Or to really bring together the argument as a whole; what can be done to mitigate the risks to one's core being is the application of statistical disclosure control methods.

Having positioned disclosure control in the space created by gelassenheit and the acceptance of technology, we can also describe disclosure control in Heideggerian terms. 'Aletheia', taken from the Greek for truth, unconcealment or disclosure is adopted by Heidegger to discuss when a 'world' is disclosed (see the earlier discussion of Erschlossenheit) (Bartky, 1979). We can see how the act of statistical disclosure, which disclosure control is set against, is therefore a revelation, a disclosure of an entity. Heidegger also offers an example of how an object relating to a being can disclose some aspect of that being's character. In a description of a painting by Van Gogh, Heidegger describes how a pair of shoes reveal something of the character of the peasant women who owns them:

> "From the dark opening of the shoes the toilsome tread of the worker stands forth. In the stiffly solid heaviness of the shoes there is the accumulated tenacity of her slow trudge through the far-spreading and ever-uniform furrows of the field, swept by a raw wind. On the leather there lies the dampness and saturation of the soil. Under the soles there slides the loneliness of the field-path as the evening declines. In the shoes there vibrates the silent call of the earth, its quiet gift of the ripening corn and its enigmatic self-refusal in the fallow desolation of the wintry field." cited in Bartky (1979, 214).

One could draw a similar, if less poetic, comparison with the details of a census, survey or patient record that reveals something of the character of the data subject not its true nature. For example their position in an occupational classification does not reveal the intimate detail of a persons day-to-day life but it does establish the broad sphere in which one would expect that they operate in.

It is also important to note here that Heidegger in later writings, and more contemporary scholars (Kompridis, 2011), have sought to separate aletheia from absolute truth: “To raise the question of aletheia, of disclosure as such, is not the same as raising the question of truth. For this reason, it was inadequate and misleading to call aletheia in the sense of opening, truth” cited in Kompridis (2011, 188). This is important because in our context, we know from Solove and Garfinkel (see 2.1) that conflating data that discloses information about an individual and the absolute truth pertaining to their identity is dangerous for the negative effects it can have on individual autonomy and integrity. This revision of aletheia adds a subtlety that makes it an appropriate concept for disclosure in the current context.

Lastly on Heideggerian disclosure and disclosure control, the beginnings of a link between the concerns over privacy articulated by Solove and Garfinkel and statistical disclosure as a form of aletheia was framed above. Within the concept of disclosure itself Heidegger makes no mention of the agency of beings or control over the process of disclosure. As Garfinkel and Solove suggest privacy is a problematic term but encapsulates autonomy and self-possession, here then one would want to assert an individuals agency in what disclosures are made through the exploration of objects pertaining to themselves. IN the sotrage of data and its subsequent processing we as a society cede our individual agency to organisations, governed by societal norms (read Nissenbaum’s tests for appropriateness and distribution), laws or user-agreements (see section 2.6). Statistical disclosure control is then a device for mitigating the risk of unauthorised aletheia and the associated risks to an individuals privacy.

The above might seem like a convenient philosophical judgement to make, however there is evidence of it in how the public and researchers’ view storage and analysis of personal data. To return to the public engagement work of the Scottish Health Informatics Programme, participants expressed a strong preference for their data to be anonymised to protect their confidentiality, so as to protect themselves from any injury resulting in their re-identification and misuse of their data. Participants also noted something that resonates with the philosophical point about enframing in this context:

> “In one discussion group a researcher described how he used data and that he did not tend to think of data subjects as being ‘real people’, this was described as being reassuring by another participant. Thus, anonymisation - or depersonalisation - of data was perceived as an important reassurance in relation to uses of personal medical data” (Aitken, 2012, 2).

This last quote raises another challenge that should be answered here. The researcher’s view that data subjects in a dataset are not ‘real people’ is something that at first glance might cause Heidegger to label such thought as enframing. The researcher sees a data subject’s data as a resource ready at hand for his manipulation. However, the anonymisation encompassed in statistical disclosure control facilitates the authorised disclosure of information that provides some characterisation of their identity. That resource which can now be pooled with that of others in order for the process of 'being-with' to proceed. This also chimes with the legal frameworks discussed earlier, the DPA's exemption for research includes the need to avoid being "processed in such a way that substantial damage or substantial distress is, or is likely to be, caused to any data subject" and stipulates that "the results of the research or any resulting statistics are not made available in a form which identifies data subjects" (UK Parliament (1998)), in effect these rules call for data controllers and data users to satisfy the tests that Nissenbaum set out for appropriateness and information flow.

To further combine the theoretical framework set out above with the main subject of this research, two further areas are addressed. In light of recent developments at the European Parliament it seems prudent to discuss the impact of proposed reforms to the Data Protection Directive and how these might fit with this research. It would also be valuable to compare and contrast public and private sector approaches to this topic, the collection and use of personal data is discussed in the government and social media context.

2.5 Proposed Revisions to the EU Data Protection Directive

Earlier in this chapter, the legislative framework, including the Data Protection Act (DPA), was set out to provide a contextual backdrop for this research. To further enhance the case for considering statistical confidentiality in an interdisciplinary context, it is useful to consider proposed changes to the DPA. These changes are being discussed at the European level and would subsequently reform the UK's legislative framework. It will be shown that the legislative framework is not static and therefore a deeper understanding of statistical confidentiality is required to adequately address the problem at the centre of this research.

The proposal set out in European Commission (2011) seeks to replace the 1995 Data Protection Directive and reinforce individuals' rights to privacy. In the Commission's report they provide a justification for reforming the 1995 Directive: "The current framework remains sound as far as its objectives and principles are concerned, but it has not prevented fragmentation in the way personal data protection is implemented across the Union, legal uncertainty and a widespread public perception that there are significant risks associated notably with online activity" European Commission (2011, 3).

The EU's call for individuals' data privacy rights to be strengthened has been welcomed by data privacy advocates and campaign groups. Much of this support has culminated around high profile clashes between large technology firms, such as Google, and the European Union, for example see Big Brother Watch (2014). While public opinion welcomes these moves by legislative bodies to enhance legal protections for data privacy, the research community has been quick to point out potentially serious threats to future research if the proposed changes go ahead without consideration of their impact on issues of statistical confidentiality and consent.

In an editorial for the British Medical Journal, Ploem et al. (2013) highlights how these new proposals threaten the research exemptions and justifications that have been discussed earlier in this chapter. Focusing on amendments put forward by the European Parliament's Committee on Civil Liberties, Justice and Home Affairs, Ploem draws on the Committee's suggestion that "[p]rocessing of sensitive data for historical, statistical and scientific research purposes is not as urgent or compelling as public health or social protection. Consequently, there is no need to introduce an exception which would put them on the same level as the other listed justifications." In effect what the Committee proposes is the removal of the research exemption that underpins much of the research carried out using personal data at the present time. In addition to this, the Committee recommends limiting the ability of individual member states from providing exceptions to proposed rules by setting the bar high with an 'exceptionally high public interest' test for research to proceed without the direct consent of data subjects (EU Committee on Civil Liberties, Justice and Home Affairs, 2012).

Mascalzoni et al. (2013) probes deeper into the proposed changes and considers the motivations of the Civil Liberties, Justice and Home Affairs Committee. In their editorial for Nature, Mascalizoni et al note that the real target for the Committee's amendments are private companies who might seek to use the research exemption to use personal data for its own ends without consent from individuals. Furthermore, they highlight the potential for private companies to link personal data for more detailed processing outside it's intended use. This provides the European Parliament with a difficult navigation of public opinion, in regards to the use of their data by private companies for example , and the needs of the research community. However, as Mascalzoni also suggests, it is not only the EU that finds itself in a difficult position. A number of public bodies have incorporated data sharing requirements in their rules for awarding research funding.

The UK Research Councils (RCUK), for example, insist that research data should be passed to data repositories for future use by other researchers working on potentially very different research projects. The Economic and Social Research Council have spent significant resources promoting the secondary reuse of data generated from publicly funded projects. This has included the Secondary Data Analysis Initiative (SDAI) (ESRC, 2012) offering grants of £200,000 to analyse existing data sources. Not only have this type of initiative increased the emphasis on data sharing, but it has in part also changed the economic model for research funding by reducing the focus on data collection.

Recognition of the benefits of data sharing in research communities has not only been driven from the top by funders. The more grass-roots ‘open access’ movement has also supported efforts for greater replicability of research findings through the public archiving of research data and supporting methodological work. The Public Library of Science (PLoS) which publishes a number of open access science journals, recently introduced a new policy on open research data and the need for authors to share data for published articles (Silva, 2014). Given the degree to which the research community is moving toward greater data sharing, the threat posed by new EU legislation is serious on a number levels as highlighted in the discussion above.

To round out the evidence in this vein, Andersen and Storm (2013) seeks to simulate the cost to research if a new EU Directive curtails the research exemption. Their case study focuses on the impact to public health if the use of cancer registries are effected by new legislation. The potential costs include: loss of research which cannot feasibly be carried out using controlled trials; increased error in statistical results as potential data sources shrink in size; and the loss of unbiased population level data which is then substituted by anecdotal, small trial or data based on stakeholder interests.

In this section, the shifts in the legislative framework that yield potential threats to data sharing and secondary analysis of research data have been established. As such, it also cements further the principle of statistical confidentiality as an interdisciplinary subject that must lie between legal, ethical and technical constituencies in order to fully encapsulate the needs of both the individuals whose data are being made available with the data users and the communal structures that exist to promote the public good. This section also sits at the sharp edge of the theoretical grounding on risk from Beck’s work set out earlier.

2.6 Public and Private Sector Approaches

So far the narrative of this chapter would indicate that individuals jealously guard their personal data, that they are sceptical of data sharing and are wary of the influence of technology on their lives. However, from anecdotal experience, it would seem obvious that the real narrative is not so linear. Individuals’ attitudes to their private data are highly context-specific, as captured in Nissenbaum’s concept of contextual integrity (see the end of section 2.1). In this section it will be shown that different contexts can provide a fundamental juxtaposition of the public’s concern for data privacy and their willingness to distribute and publish their data. This will be done using evidence of the public’s approach to data sharing within the governmental context and the context of social media, such as Facebook. A tension exists between the lengthy deliberations on ethics and the use of personal data in the public sector and the terms of service consent model used by private companies. This tension stands in stark contrast to the public's attitude which often sees the first as problematic and the latter much less so. These tensions will also be discussed with reference to the theoretical framework set out above and the literature on privacy.

To take the public sector context, if the contemporary issues of surveillance and government 'snooping' are left to one side, the focus can shift to public sector use of personal data in the delivery of public services and research. Through its routine business, government collects a vast quantity of data on its citizens. This data collected for administrative purposes also forms the basis for research - so much so that the Office for National Statistics (ONS) has recommended supplementing and in the future replacing the UK Census with data drawn from administrative sources (Office for National Statistics, 2014). To set out a brief case study for comparison with the private sector, consider the use of secondary care health data collected in hospitals in Scotland. This data is used in research such as McAllister et al. (2013), in which academics in collaboration with NHS Scotland analysed ten years worth of hospital admissions data to test for a link between socio-economic deprivation and the risk of hospitalisation during the winter months. In order to access data of this kind, the researchers would need to have sought ethical approval from their institutions ethics committee as well as seeking ethical approval from the NHS' Privacy Advisory Committee (PAC), a process which in the case of this thesis research took over a year (for information on PAC see NHS National Services Scotland (2013b)). This is complemented by studies of public opinion to data sharing such as Wellcome Trust (2013) and those specifically in the health field Scottish Government (2011a) which explore both participants' reluctance and support for their data to be shared. What emerges from both studies is that individuals expressed a degree of cynicism when discussing governments collection and storage of personal data. This cynicism was often supported by anecdotal personal experiences or by citing newspaper stories of data loss (BBC (2009) for example). This cynicism also speaks to the general concern over governments centralising of data that Garfinkel discusses (Garfinkel, 2000).

As for attitudes toward data sharing in the health field, the research suggests that the cynicism expressed in the general case was tempered by support for a 'public good'. The idea, common among participants, was that if data sharing resulted in improved public health then the collective good should take precedence over individual rights to privacy, with the caveat that data should be treated as sensitive and handled with particular care. As shown in the Scottish Government report, this attitude became more pronounced if participants accessed more information about how data sharing is carried out and what safe guards are in place Scottish Government (2011a, 3.32). This contextual approach to the use of data for a public good, is similarly described by Nissenbaum as satisfying the tests of contextual integrity, it is both deemed an appropriate level of data to reveal to legitimate health researchers and with the various measures, including data security, it conforms to the norms of information distribution that society would expect from this context. To illustrate this further, when describing societal norms of information flow in the healthcare context Nissenbaum suggests ways in which the physician-patient relationship change with the context, i.e. “where it poses a public health risk, and where it is of commercial interest to drug companies” (Nissenbaum, 2004, 124).

These discussions of public attitudes to data sharing in the public sector resonate more strongly with the more pessimistic outlook of Heidegger and the enframing described above. Having set out one side of the apparent juxtaposition, the second side can be established to complete the contrast. The situation regarding data sharing through social media platforms or other web based services is much less well defined than the earlier case study. The processes for data collection, analysis and dissemination are effectively closed within the confines of private companies, although they do provide statements of intent, for example Facebook (2013) or Google (2014). However, there has been some research into public attitudes toward data sharing within these services.

Although the specific and seemingly rational fears identified by Nissenbaum, and to a lesser extent Garfinkel, seem to fit with the above discussion of the use of data in the public sector. They seem less applicable when discussing the sharing of data with private companies. In this context the scepticism discussed in Solove (2008) seems more relevant, he discusses the public’s concern regarding privacy as abstract, and citing Goldman notes that “stated privacy concerns diverge from what [they] do” (Solove, 2008, 5). This dismissal of the public conern for there on data is somewhat simplistic and the transactional nature of this engagement is explored below. Triangulated with Nissenbaum and Garfinkel this contradiction actually lends weight to parts of their argument. Garfinkel’s reluctance to allow private organisation, in his context credit agencies and here social media companies, to build myriad secret data warehouses is because of the ambiguity this introduces and the lack of public scrutiny over the use of personal data. Equally, if these two contexts are viewed through the prism of contextual integrity, there has been a great deal of public scrutiny over the appropriateness and distribution of data in the business of government and research. However, the same level of scrutiny has not been applied in the social media context, in fact the norm thus far is for individuals to provide data to these companies who then have full control of those data (see Solove’s criticism of Amazon.com’s ‘unfettered ability’ to use personal information (Strandburg, 2006)), if with some degree of self-regulation through user agreements.

To illustrate first the model used by these services, a new user is asked to agree to the terms of service which often carry wide ranging clauses on how their personal data will be processed. For example, Facebook’s terms of service[2] state the following:

> [clause 2.1] For content that is covered by intellectual property rights, like photos and videos (IP content), you specifically give us the following permis-

[2] The terms of service can be found at https://www.facebook.com/legal/terms

> sion, subject to your privacy and application settings: you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook (IP License). This IP License ends when you delete your IP content or your account unless your content has been shared with others, and they have not deleted it. ... [clause 10.1] You give us permission to use your name, profile picture, content, and information in connection with commercial, sponsored, or related content (such as a brand you like) served or enhanced by us. This means, for example, that you permit a business or other entity to pay us to display your name and/or profile picture with your content or information, without any compensation to you. If you have selected a specific audience for your content or information, we will respect your choice when we use it.

These two examples demonstrate the degree to which users provide Facebook with the ability to process their personal data. This interaction between Facebook and individuals at a personal level also allows for the processing of large volumes of data at an aggregate level. This aspect of Facebook's model has often been referred to an iceberg as illustrated in figure 2.1. Having highlighted this relationship between Facebook, users and their data, it can be shown that user attitudes do not necessarily reflect the level of access to their data ceded to social media platforms when they agree to the terms of service. Debatin et al. (2009) provides a

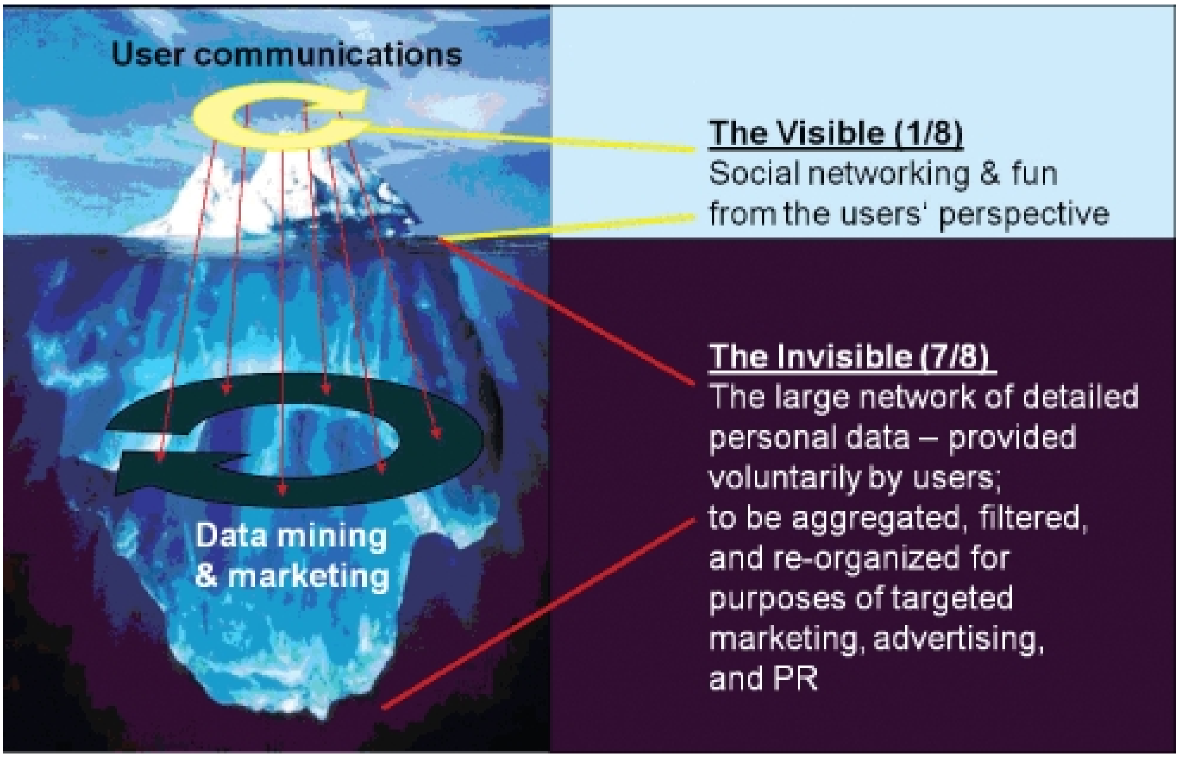


Figure 2.1: The Facebook Iceberg Model (Debatin et al., 2009)

comprehensive review of existing literature on this topic. Debatin et al draw on previous studies of Facebook users, citing Jones and Soltren (2005), for example, which revealed that over 70 percent of MIT students surveyed posted demographic data about their age, gender and location. In addition, the users largely ignored the terms of service, 91 percent agreed to it without reading it, and 89 percent acknowledged that they had never read the privacy policy. What is also striking from the literature is that the correlation between privacy concerns, trust, and greater information highlighted in the government example does not seem to apply here. Govani and Pashley (2005) suggests that even when users that express privacy concerns receive more education about privacy setting in social media platforms these remain under utilised and yet privacy concerns remain post-education.

A substantial difference between disclosing and sharing personal information with government and social media are the tangible benefits of such activity. As indicated in the literature above on public attitudes to data sharing in the public sector, individuals often support the 'public good' argument about data sharing. However, this type of collective, non-immediate outcome sits in contrast to those outcomes identified from participating in online social networks. Ellison et al. (2007) discusses the creation and maintenance of an individual's social capital through participation in online social networks. They highlight the use of these online networks to create, build and maintain relationships with local peers or the maintenance of connections if users move away from an area. These immediate benefits of social networking are also complemented by potential longer term benefits, for example, Ellison suggests "Such connections could have strong pay-offs in terms of jobs, internships, and other opportunities" Ellison et al. (2007, p1164).

To tie this brief summary of social media to the wider theoretical framework, the benefits of online social networks resonate with Heidegger at the Mitsein dimension of the enframing versus Mitsein tension identified earlier. Acknowledging that people are social beings that understand the world through shared subjective experiences and a common pool of knowledge lends itself to explaining the popularity of online social networks like Facebook. The tangible benefits to individual users serve to mitigate the privacy concerns of sharing data to an extent to which users ignore policies and settings available to protect their privacy. This hook into individuals' social understanding of the world is far less visible in the case of government or public sector data usage. This gap forms an important context for this research and lends itself to the interdisciplinary approach taken forward in the succeeding chapters.

What has been shown in this section is that individual approaches to data privacy are highly context specific, and that there is a degree of evaluation, conscious or sub-conscious, of risk and reward framed by societal norms. This balance of risk and reward, or utility, runs through this research from this theoretical framework through practical application of statistical disclosure control measures. In identifying this continual tension between approaches to online services and government data sharing and research, an avenue for further work is potentially created. Future research could consider a variation on the model used by private companies through terms of service agreements in the public sector and research sphere. This has the potential to redraw the risk and reward balance to further public knowledge, participation and trust of data sharing in this context as well as to some extent assuage concerns over informed consent.

It has been shown that there is stable conceptual ground for research into disclosure control drawing on different disciplinary perspectives. What will be set out in later chapters are further combinations of different disciplinary approaches to more applied methodological areas of statistical disclosure control. Having this conceptual grounding is important to avoid pursuing disclosure control in only the mathematical, statistical or technological sense. As Beck's sociology of risk indicates there are factual (technical and statistical) judgements as well as value (ethical and sociological) judgements to be evaluated in this process. This perhaps goes some way to explaining why the uptake and understanding of this field up to now has been limited to specialist statisticians aided by technology. As the 'big data' revolution changes how vast amounts of personal data are stored, collected and analysed, a greater understanding of the accompanying risks are needed. This understanding has to include the data subjects themselves to ensure that the public can consent, whether explicitly or tacitly, to their data being used, otherwise the cause of data analysis could suffer by tipping the balance too far in either direction as the proposed changes to EU legislation suggest. Perhaps, there are lessons for government to be learnt from the approach taken by Facebook in at least recognising the transactional nature of their interaction with their users. Heidegger' s challenge in this field remains, statistical disclosure control is a constant rebalancing of the 'enframing' of technological progress with the 'Mitsein' of the social world calibrated by a 'Gelassenheit' letting-be to protect the rights of the individual and advance the social good.

Chapter 3

Methodology

3.1 Research Design

In Chapter 1 the research problem was discussed in detail. In order to address these research questions within the evolving, practice focused fields of data management, administrative data research, and confidentiality, the research design set out below draws heavily from the pragmatic tradition. As Creswell (2013, p11) sets out that the pragmatists place the problem at the centre of the research occupying a more important position than the necessity to apply one set of methods or stick to a more rigid philosophical framework. This allows pragmatic research designs to focus on solutions and 'what works' approaches. The flexibility afforded by the pragmatic approach make it easier for researchers to mix methods from different traditions and, as Jick (1979) discusses, the products of these different methods can be engaged in triangulation to leverage data and findings that potentially address the weaknesses or biases inherent in any one method. Also as Creswell notes, audience is important when adopting a research tradition to anchor the overall design. As the audience for this research crosses both academia and data controllers working in the public sector, it was important to acknowledge the needs of both groups and provide potential solutions to current problems.

Following from the pragmatic approach discussed above, this research utilises a sequential strategy. Cameron (2009) summarises the sequential strategy as one where "One type of data provides a basis for collection of another type of data" (Cameron, 2009, p144). Creswell expands on this and describes the sequential strategy as one in which the researcher can respond to opportunities for further data collection and react to situations as they evolve. Although the methods set out below should not be considered strictly linearly sequential, they do follow some sequential patterns. For example, the interviews preceded the development of tools like NIAH, and to some extent the statistical analyses that provide evidence of data utility. In the sections below the individual methods used in the pragmatic-sequential mixed-methods adopted in this thesis are discussed. Before moving to specific methods, the e-

Science concept is discussed in greater detail because it adds further character to the chosen research approach.

3.2 e-Science as a Methodological Context

E-Science, as a concept, sits between the discussion of an overarching approach to the research (pragmatism) and specific methods that constitute the study design. As stated in 1, in the broadest sense, e-Science is defined by “the application of computer technology to the undertaking of modern scientific investigation, including the preparation, experimentation, data collection, results dissemination, and long-term storage and accessibility of all materials generated through the scientific process” (Bohle, 2013b). However, it is important to expand upon this definition and clarify why this methodological context is important for this research.

John Taylor, then Director General of Research Councils in the Office of Science and Technology introduced the term e-Science in 1999 to describe emerging collaborative, and multidisciplinary, approaches in science and engineering. At its inception e-Science was the label for large investments in computing infrastructure, especially distributed technologies like ‘Grid Computing’, to support scientific enquiry. For example, infrastructure designed to assist with the experiments planned for the, then under construction, Large Hadron Collider were at the forefront of an e-Science collaborative effort between computer scientists and particle physicists (Hey and Trefethen, 2002). From this early projects the concept solidified around the idea that modern computing techniques could be developed and applied to enable the scientist’s workflow.

With time, the concept of e-Science was applied to other disciplines, while maintaining the idea of the ‘enabled researcher’s workflow’ at its core. In 2004, the Economic and Social Research Council launched its own programme of e-Science through the National Centre for e-Social Science (NCeSS). This programme contained two strands: one focused on applications of the technological developments emerging from the e-Science sphere; and another focused on the social study of technology - questioning how these technologies were being developed and applied in the social sciences and what implications that might have (Halfpenny and Procter, 2010). Here the early focus on Grid computing was seen as less important, and instead the application of computing science techniques more broadly was given greater importance. For example, the Data Management through e-Social Science (DAMES) project was funded through this programme and was a collaboration between social scientists and computing scientists at the universities of Stirling and Glasgow to facilitate and improve data management processes such as mapping between different occupational classifications (Tan et al., 2009).

For this research, e-Science, and its social science offshoot, provides a methodological context for collaboration between the computing and social sciences which focuses on facilitating and improving the workflow of data controllers, data analysts, research coordinators and data users in their management, preparation and analysis of potentially disclosure data. This context also necessitates a multidisciplinary exploration of this workflow which provides space for the theoretical and ethical discussions of Chapter 2, as well as the development of software tools like NIAH (and the extensions of NIAH in Chapter 6), and the mixed methods of quantitative data utility analyses alongside qualitative discussions of the user experience (see Chapters 4, 5 and 7). In the remaining sections of this Chapter, the techniques and methods used are described in more detail.

3.3 Interviews

Semi-structured interviews were used across a number of purposes in this research: they provide a contextual backdrop to the exploration of the issues raised in the research questions; they offered insights into stakeholders' experiences and expectations; and they provided user requirements that helped to inform the development of analyses and tools including the NIAH toolset developed later in this thesis.

Flowing from the pragmatic mixed-methods research design, these interviews provide an opportunity to explore in-depth some of the practical issues that arise from the legal, ethical and theoretical discussions in Chapter 2. These are explored within the frame of reference of practitioners, and analysts who manage and analyse data from administrative sources, primarily in the sphere of health and social care in Scotland. Practitioners offer access to a pool of knowledge and experience that would rarely be available to an academic researcher, however negotiating access to these participants introduced limitations in the scope of the interview guide and limited the ability to publish specific extracts from individual interviews (These limitations will be discussed in detail after the range of participants has been established).

In total, five semi-structured face-to-face interviews were carried out at the participants place of work during March - June 2012. The interviews were conducted by the researcher in the presence of a Scottish Government statistician, the interviews were recorded where consent was given. The interviews had both ethical approval from the University of Glasgow, College of Science & Engineering as well as the Scottish Government. Information sheets, Consent forms, and an interview guide can be found in Appendices A, B and C. The participants were drawn from three main archetypes: data users - two local authority analysts (from different local authorities with different levels of analytical capacity) and an academic with experience of using administrative health data; data managers - a programme manager for a large project on electronic health records; and data controllers - a senior health board manager with responsibility for data confidentiality. In addition, one participant occupied both the data user and data manager space - a senior researcher from a health informatics project.

The limitations highlighted earlier were introduced for two reasons. The Scottish Government served as gate-keeper and also facilitator because these interviews were carried out during an internship with the Health Analytical Services Division. They provided access to networks of local authority and health board contacts. Further, during the interviews the researcher occupied both their own research position and an intern position and as such a compromise was reached regarding the interview guide. This compromise ensured that both the researcher and the Scottish Government obtained useful information from this data collection. The second limitation was a product of the topics sensitive nature and the positions of the participants. Participants were willing to be interviewed, but were more comfortable consenting to remain anonymous and preferred their information to be used as contextual background rather than explicitly quoted. Although these limitations are acknowledged they did not present a significant problem for the research design because they were intended to inform other aspects of the research rather than be the central research result in themselves. One benefit from the first limitation was the continued support of Scottish Government contacts which is discussed further below with reference to the case study approach.

3.4 The Case Study Approach

In order to keep this research grounded for the interdisciplinary audiences, the design makes use of the case study approach. Hammersley describes case study as a broad concept with the potential to be used in a number of different ways (Hammersley, 2004). However, there are commonalities in the uses he describes; the concentration on one or a few cases allows for greater detail and potentially different types of data to be collected. Also, the case study is a common component in research that draws on the e-Science methodological context, for example see (Goderis et al., 2006; Salayandia et al., 2006; Brown et al., 2007; Addis et al., 2003). IN fact, this use of case studies in e-Science has itself been the subject of philosophical discussion (Beaulieu et al., 2007). In e-Science case studies are often used to present tools or approaches that offer general solutions to problems (mining large volumes of text for example) but within the scope of a interdisciplinary project (applying those text-mining techniques to the parliamentary record - Hansard (Sarwar et al., 2013)) In the context of this research, the case study is used to provide practical examples of research methods (the use of the Scottish Health Survey throughout Chapter 4) and arguments and proposed tools or workflows (see Chapter 7). At its highest level of abstraction, the ‘case’ which is under scrutiny is the Scottish confidential data environment in which the Scottish Government and other public agencies operate. Within that scope lower level operations like the Scottish Health Survey and administrative data services can be probed.

In the above discussion of the use of interviews, two limitations were described. The need for compromise within the interview guide was highlighted because this compromise facilitated access to participants. In addition, this collaborative approach facilitated access to potentially rich case studies on how systems of data management, access and utilisation evolve in the governmental context.

In common with the e-Science examples cited above, the case study presented in Chapter 7 is used to propose a solution to the research problem that can be adapted and developed by future work. Lastly, to return to the idea of the audience for this research, the case study using existing administrative infrastructure provides a familiarity for practitioners in the administrative data space which potentially eases the integration or discussion of the solution proposed alongside current practices.

3.5 *k*-Anonymity as a SDC Method

K-anonymity is a measure of re-identification risk that presents an idea that can be simply understood and communicated to data controllers, researchers and the public. It was first introduced in 1998 by Samarati and Sweeney (1998a). Since then, it has been built upon (El Emam and Dankar, 2008), expanded (Stokes and Torra, 2012) and critiqued (Domingo-Ferrer and Torra, 2008). The basic definition provided by Samarati and Sweeney still holds true throughout the later incarnations. They stated that "Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least *k* individuals" (Samarati and Sweeney, 1998a, 4). Critiques of *k*-anonymity focus on its inability to protect against attribute disclosure, rather than re-identification. Domingo-Ferrer and Torra provide a clear example to illustrate this problem: "Imagine that an individual's health record is *k*-anonymized into a group of *k* patients with *k*-anonymized key attributes values Age = 30, Height = 180 cm and Weight = 80 kg. Now, if all *k* patients share the confidential attribute value Disease = AIDS, *k*-anonymization is useless, because an intruder who uses the key attributes (Age, Height, Weight) can link an external identified record (Name= John Smith, Age= 31, Height=179, Weight=81) with the above group of *k* patients and infer that John Smith suffers from AIDS (attribute disclosure)" (Domingo-Ferrer and Torra, 2008, 991). As can be seen if a set of records have the same variable value and this is considered sensitive, then *k*-anonymisation is not enough to protect those individuals from disclosure. In this research, the methodology we have used tacitly considers the threat of attribute disclosure as part of the subjective judgements made during the data sensitivity

Age	Post Code	Gender	Illness
22	G1 5FL	M	Diabetes
22	G1 5FL	M	AIDS
22	G1 5FL	M	Diabetes
36	G2 4FG	F	Heart Disease
36	G2 4FG	F	Heart Disease

Table 3.1: An Example of *p*-sensitive *k*-anonymity

step.

There are enhancements to *k*-anonymity in the literature such as *l*-diversity (Machanavajjhala et al., 2007), *p*-sensitive *k*-anonymity (Truta and Vinay, 2006), and *t*-closeness (Li et al., 2007). These enhancements introduce approaches to cope with the attribute disclosure problem highlighted in the above example. As they rely on the same logic, it is possible to provide a high-level illustration to these approaches with an example of *p*-sensitive *k*-anonymity. In order for a data set to satisfy *p*-sensitive *k*-anonymity, it must first satisfy the *k*-anonymity requirement specified in the definition (all records indistinguishable from *k*-1 other records across a set of quasi-identifiers). Within these *k*-anonymous sets of records the sensitive variables that contain the confidential information that needs to be protected (typically identified in a sensitivity analysis) must contain *p* number of different values. The value of *p* is always less than or equal to *k*. Table 3.1 provides an example data set which can be described in terms of *p*-sensitive *k*-anonymity. In this example, *p*-sensitive, *k*-anonymity is satisfied where *p*=2 and *k*=3. Therefore, an intruder cannot say with certainty that they know the illness associated with any one record should they match it to their external data. In the example, attribute disclosure would occur for the second group, here *p*=1 and *k*=2, because *p*=1 an intruder knows that a female aged 36 in that postcode must have heart disease.

The problem with implementing *p*-sensitive *k*-anonymity is that it has a devastating impact on data utility. In much the same way as, but worse than, if the data controller's assessment of the data environment resulted in a large number of quasi-identifiers. The more privacy constraints introduced, the more the raw data must be manipulated to satisfy those constraints.

3.6 Data Utility Analyses

The approach taken to data utility here draws on examples in the literature of replicating 'real world' analyses to estimate data utility, see Purdam and Elliot (2007); McCaa et al. (2013); Brickell and Shmatikov (2008) (alternatives to this approach are considered in section 4.4). In particular, Purdam and Elliot (2007) offers a case study approach to the effects of disclosure control on publicly released microdata that is particularly appealing given the pragmatic research design and our earlier outline for the case study approach favoured by this research.

The use of case studies allows for the discussion of data utility in the overarching context of SDC, and thus the practical utility of the data. The authors contacted researchers that had published work using census derived microdata samples such as the Sample of Anonymised Records (SARs) and asked them to complete a questionnaire, and for permission to replicate their analysis on different versions of the SARs with varying SDC measures applied. The results were somewhat surprising given that the authors of the work under scrutiny had been published in well-respected publications. The replication of analyses highlighted significant differences in the accuracy of some of the analytical results. One notable example detailed the inflation of minority ethnic groups within a given social class. As discussed by the authors these differences not only cause concern in terms of analysis of the specific dataset, they also raise concerns if these conclusions are used in comparison with data from other sources, smaller scale sample surveys or administrative records for example.

In Chapter 4 and 7 research questions are posed and quantitative analyses are constructed and then applied to the original and disclosure controlled data in order to demonstrate the data's practical utility to users that might carry out similar analyses. This approach was chosen because it ties neatly to the pragmatic research design, as well as providing results that can be communicated to users or data controllers with experience of quantitative analysis.

3.7 The NIAH Methodology

The NIAH toolset, which was developed during the course of this research and forms part of its overall output, is a set of command line tools that together constitute part of the SDC workflow. It is used in the analysis presented in Chapters 4, 6 and 7. Overviews of the *k*-anonymity algorithm used and the other tools in the set are provided below. This section describes the NIAH tool set in its finished form, however its development cycle closely followed the pragmatic-sequential design outlined in this Chapter. Therefore, the justification for NIAHs design and developed are not discussed here but can be found in Chapter 6.

A k-Anonymity Algorithm

At the core of this tool-set is the k-anonymity algorithm. As has been discussed earlier in this Chapter, k-anonymity was defined by Samarati and Sweeney (1998b). Their specification for k-anonymity as a privacy property is easily defined. In order to satisfy a k-anonymity requirement, all possible combinations of quasi-identifiers (also referred to as key variables) must apply to at least *k* records.

To assess whether a dataset complies with this privacy requirement, an algorithm is needed to process the data. At the abstract level, the algorithm used in the NIAH tool-set performs the following tasks:

Algorithm 3.7.1: VOID NIAH($D, nColumns, nRows, V, nV, k$)

```
 comment: Apply k-anonymity algorithm to dataset D, which has nColumns and
 nRows using quasi-identifiers V of size nV, and for value k upon return, each row in
 D that is tuple unique will be marked as such. N.B. the sort algorithm used must be
 stable
 for $i \in 1$ to $nV$
   do sort D by column V[i]
 for each subset of rows of D with the same value in column V[i]
   do if number of rows in subset <k
   then mark each row in subset as minimal unique
 comment: Now eliminate tuple combinations that do not meet required k-anonymity
 for $i \in nV$ to $1$
   do sort D by column V[i]
 for each for each subset of rows of D with the same values in columns V[1] to $V[nV]$
   do if number of rows in subset <k
   then mark each row in the subset as minimal unique
```

Each of these items is addressed in turn and a low level summary of design and implementation decisions are given. It should also be noted that the basic approach outlined above is in part based upon the first stages of the Special Unique Detection Algorithm (SUDA) set out in Elliot et al. (2005b).

NIAH takes a series of command line arguments set out in the usage statement. This includes the input CSV file. CSV was chosen as a 'lowest common denominator' so that NIAH could be integrated with various workflows, as most data analysis software will accept and export CSV format data. Integration is discussed in greater detail at the end of this chapter. Although CSV is a common format, it can be a problematic format to work with across different software tools. This is because there is no defined standard for CSV format. There have been attempts to standardise CSV, for example Shafranovich (2005). However, some inconsistencies still remain including the use of quotation marks to denote string values that may contain commas themselves. The overall approach in the development of NIAH attempted to do as little 'harm' to the input data as possible so that NIAH's output would be consistent. In keeping with this approach, instead of replacing those commas within variable values or changing the data directly, a custom split method was written for strings which ignores commas between quotation marks.

The internal data structure was conceptualised to form a table-like structure similar to a data table in a spreadsheet. Variable values are held in cells and cells are held in rows. An assumption was made that a trade-off of complexity set against availability of resources would, in this context, favour less complexity and greater allocation of system resources. This was assumed so that speed could be prioritised over memory usage - in order to provide faster performance all data are read into memory. Performance results are provided at the end of the next section.

The data are read into an ArrayList of ArrayLists in Java, as these "Resizable-array implementation[s] of the List interface" [1] allow the data structure to dynamically resize as more records are read in. Although memory usage was a secondary consideration, the decision was taken to store the data values as an Array of Bytes as opposed to Java String objects. This saves space because each element in the Byte Array only occupies one Byte as opposed to the two Bytes (at least) for each character in a String object (Strings store characters in a character array, so the space required will be dependent on the encoding). Also, within each 'cell' in the data structure, the type of the cell, string, integer or double-precision floating point number, is indicated. This is included so that type-appropriate comparison methods can be used and so that the data structure provides further functionality (this further functionality is discussed below).

The user specifies the column number(s) of the quasi-identifier variables in the command line arguments. The data is then sorted by each of these variables in turn and each value is compared with the value of the same variable in the adjacent records to produce a count for every represented value of the respective variable. If this count is less than the specified value of *k*, then the row number of the records with that value is stored in an array. To borrow the terminology from Elliot et al. (2005b), these records are referred to as minimal uniques. The logic in identifying these minimal uniques is that if variable values within one key variable do not satisfy k it is not possible for tuples that contain that value to satisfy k. This also cuts down the number of tuple comparisons needed.

Once the minimal uniques have been identified, it is important that the sort method implemented before the tuples are analysed is a stable sorting algorithm. This is necessary to maintain the relative order of the elements prior to the sort if the elements are sorted iteratively. This application of iteratively sorting the data structure by the specified quasi-identifiers results in the grouping together of tuples that share the same combination of quasi-identifiers. The data is sorted across the quasi-identifiers in reverse order. This produces the result above with one less sort operation required when compared with a sort across quasi-identifiers in their normal order.

Once the record (row) position of both minimal uniques and tuples that do not satisfy k-anonymity are recorded these records are copied into a new instance of the data structure described above and subsequently removed from the original. Both of these data sets are then written out to separate CSV files with a user specified output name.

[1] This text is taken from the Oracle website at http://docs.oracle.com/javase/6/docs/api/java/util/ArrayList.html Accessed:11/07/2014

NIAH - Further development

NIAH as a research tool was developed by incorporating the qualitative evidence from the case studies and interviews presented later in this thesis (see Chapter 6). Here, two of the development use cases are set out. These cover two common activities carried out by data controllers when considering disclosure risk. To set these use cases in their correct context it should be noted that each contain pre-conditions. These parameters drive the k-anonymity assessment, which in turn only provides useful output if there has been a rigorous process to determine suitable parameter values. These parameters are numerical representations of the ethical judgements Beck encourages risk analysts to make in Chapter 2 and the qualitative elements of disclosure risk assessment that will be discussed in Chapter 4.

1. **Initial disclosure risk analysis for microdata.**

   Actors:

   Data Custodian / Research Coordinator / Guardian

   Goal:

   Provide evidence of underlying risk in stored data.

   Preconditions:

   Actors must have selected key variables and an initial threshold.

   Summary:

   A k-anonymity assessment is performed on the data with the selected key variables and threshold.

   Related Use Cases:

   2. Statistical disclosure review of research outputs 3. Disclosure risk analysis for publication of tabular microdata.

   Steps:

Actor actions:	*System responses:*
1. Start NIAH Shell	
	2. Display shell console window.
	3. Create log file and capture log entries from now until console window is closed.
4. Invoke command 'NIAH' with necessary arguments.	
	5. Display confirmation of input arguments.
	6. Create '[output]_safe.csv' and '[output]_atrisk.csv'.
	7. Display runtime.
	8. Return to shell command line.
9. (Optional) invoke command 'Stats'.	
	10. Display Summary statistics for last run of NIAH.
	11. Return to shell command line.
12. Close console window	
	13. Save log file and exit.

Post-conditions:

The system has a record of the risk analysis carried out and the arguments specified. This can be used for replication of analysis.

2. **Statistical disclosure review of research outputs.**

   Actors:

   Data Custodian / Research Coordinator / Guardian / Analyst

   Goal:

   To provide evidence for Statistical Disclosure Control (SDC) decisions about publishable research outputs.

   Preconditions:

   The researcher has carried out some statistical analysis using their package of choice. The underlying dataset with the researchers modifications are saved in a shared location in CSV format. The data custodian has selected initial key variables and a threshold for analysis.

   Summary:

   A k-anonymity assessment of underlying data is carried out. The results of this assessment determine the actions of the data custodian: The data are deemed to be of acceptably low risk; the risk is too great and SDC algorithms are applied and the data are reassessed. Once the data are considered low risk the data can be re-analysed in the statistics package for data utility.

Steps:

Actor actions:	*System responses:*
1. Guardian: Start NIAH Shell	
	2. Display shell console window.
	3. Create log file and capture from now until console window is closed.
4. Guardian: Invoke command 'NIAH' with initial arguments and underlying dataset.	
	5. Display confirmation of input arguments.
	6. Create '[output]_safe.csv' and '[output]_atrisk.csv'.
	7. Display runtime.
	8. Return to shell command line.
9. Guardian: (Optional) invoke command 'Stats'.	
	10. Display Summary statistics for output files.
	11. Return to shell command line.
12. Guardian: (Optional) invoke 'Viewer' command.	
	13. Display safe and/or atrisk output files.
14. Guardian: invoke '[SDC-algorithm]' command.	
	15. implement [SDC-algorithm] on last NIAH input dataset.
16. Guardian: (Optional) invoke 'Undo' command.	
	17. Revert to last NIAH input dataset.
- Guardian: repeat steps 4 - 16 until satisfied with results.	
18. Guardian / Researcher: invoke 'stata' command.	
	19. open stata in a new window with current working NIAH input file.
20. Guardian / Researcher: reruns analysis for data utility (not logged)	
- if data utility is satisfactory.	
21. Guardian / Researcher: generates research outputs in stata.	
22. Guardian: Close console window	
	23. Save log file and exit.

Post-conditions:

The system has a record of the risk analysis carried out and the arguments specified, this can be used for replication of analysis. The guardian and researcher can provide evidence for decisions about published outputs.

Although the above use cases are written for microdata, they are also applicable to tabular data similar to that often published in Government statistical reports. In the tabular case, the data underlying tabular outputs are used as input for NIAH and its associated tools. This completes the overview of the research design and methods used in this thesis. Chapter 4 continues and provides a review of the literature on statistical disclosure control and positions this research design in its proper context.

Chapter 4

Review of the Statistical Disclosure Control (SDC) Literature & existing SDC Methods

4.1 A Historical Note

To provide some historical context to this chapter and also the wider research frame, some highlights in the historical development of statistical disclosure control (SDC) are drawn out and a short comparison of how different countries have approached the SDC question is set out below. The first point to make is that the origins of much of the work on SDC have been tied to the practical considerations for data collection, mostly by governments, rather than demands in public discourse or academic research. In the USA, the Census Bureau has led on much of the SDC development. Greenberg (1990) sets out that SDC or 'disclosure avoidance' became an area of research around the time that the first general public microdata release from the Census of Population and Housing 1960. Greenberg also discusses that the suppression of values (discussed later in this Chapter) was the preferred method of disclosure control for US Census outputs through the 1970s and into the 1980s.

At the same conference in 1990, Skinner et al. (1990) discusses the situation in the UK. Unlike the US, the UK were yet to make public microdata releases from the population census, but plans were laid out for a Sample of Anonymised Records (SARs) from the 1991 census. In their introduction, Skinner *et al* comment on the public's attitude to these plans: "Public fears in this regard seem to have moved away from state surveillance more to commercial intrusion and 'junk mail"'. This is interesting if only because it suggests the opposite to the contemporary position discussed in Chapter 2. Although it is also interesting to note that this development in disclosure protocols coincides with the timing of Becks work on the risk society (see section 2.3), this process of reflection on societal norms regarding privacy is an example of the process of reflexive modernity and the perceived risks of human agency, i.e. the publishing of greater volumes of data.

The UK's plans for the SARs focused on methods of disclosure control that sought to retain the 'data's integrity' (value suppression or aggregation for example) as opposed to 'contamination' methods that transform the data values themselves (e.g. adding noise). This paper also highlights a discussion about the data environment which is of particular relevance later in this chapter. The central argument in this discussion is the need for data owners to consider their release of data alongside existing data releases from their own and other external organisations. Skinner *et al* note the tension that exists here, researchers have argued that trying to review the whole data environment is an almost impossible task (McGuckin and Nguyen, 1988) while others acknowledge the difficulty, but insist that such an attempt is worthwhile (Marsh et al., 1991).

As the development of statistical disclosure control in the US and UK has been linked with the production of outputs from large scale national surveys and population censuses, it is worth drawing comparisons with other nations. The Netherlands has been responsible for a significant amount of SDC development in the European Union, which has included the development of methods and software (for example see Hundepool and Willenborg (1996)). Unlike the US and UK, the Netherlands operates a population register rather than a decennial survey based census. Therefore, one might expect their SDC development might have taken a different form. However, Nordholt (1999) discusses the rules operated by Statistics Netherlands (the national statistical agency of the Netherlands). These rules bear a strong resemblance to those of the UK and the US, despite the difference in how they collect and store data on their citizens. At the time, Statistics Netherlands implemented rules that used suppression of variables, specifically including nationality, country of birth and ethnicity. In addition they stipulate specific rules relating to geographical variables; direct regional identifiers are not included, instead aggregate descriptive values are combined and released (for example, the size class of the place of residence). The Netherlands also operate secure sites from which analysts can work on more detailed microdata that is not deemed safe for release (These secure sites are discussed further in Chapter 5). Nordholt (1999) also discusses the development of SDC processes in EU transition countries after the collapse of the Soviet Union, suggesting that many of the countries took methods and processes from established EU member states. This combined with the need for compliance with EU regulations on data privacy and security mean that a significant amount of homogenisation has occurred in SDC processes across Europe. This includes states, like the Nordic countries, that have traditionally had a different relationship with the EU. Sweden uses the well established methods of record swapping for their census outputs (Andersson et al., 2013; Jansson, 2012). In the 2011 Census, Statistics Norway took a different approach from other European states. Heldal and

Badina (2013) set out the use of rounding of small counts in tabular census outputs, which the authors argue offers sufficient protection comparable to the cell suppression used in other countries.

The Scottish Government and National Records of Scotland surveyed the methodologies used in other countries ahead of their 2011 census data releases (National Records of Scotland Census Division, 2013). This summary reinforces the homogeneity discussed above, see Table 4.1. It is important to note that despite new methods or more complex variations on existing methods (which will be highlighted in the rest of this chapter), national statistics agencies have chosen to stick with existing methods. This is important for this research because in the case studies and qualitative work that have been carried out, the majority of data owners look to their national statistics agencies for guidance on SDC for both tabular and microdata releases. Therefore, there is a gap between new statistical techniques in the literature and their implementation at the production level by data owners. A possible reason for this gap is the lack of literature that fully explores the 'real world' utility of data that has been subjected to new, more complex, statistical techniques. Cleveland et al. (2012) and Purdam and Elliot (2007) demonstrate that ensuring data utility for the widest possible set of use cases is not a trivial task. As a consequence, although new methods are highlighted in the proceeding sections the main focus is the implementation and effects of existing techniques.

Method	Countries
Pre-tabular record swapping	Scotland, England, Wales, Northern Ireland, Belgium, Austria, Israel
Post-tabular cell perturbation algorithm	Australia
Pre-tabular perturbation algorithm	Germany
Combination record swapping and post-tabular rounding	Sweden
Rounding of small counts	Norway
τ-argus post-tabulation software	Slovenia

Source: National Records of Scotland Census Division (2013)

Table 4.1: Statistical Disclosure Control Methods Used in Different Countries

4.2 Measures of Risk

The measurement of disclosure risk has to be considered as a collection of processes that collectively provide the parameters that can be input into disclosure risk estimation (for example see the discussion of 'factors which contribute to risk' in Greenberg and Zayatz (1992) and the section on 'estimating re-identification risk' in Hundepool et al. (2012)). These estimates might be metrics or more qualitative evaluations. These processes are labelled in different ways across the literature but their definitions are very similar; for examples of different labelling see El Emam and Dankar (2008) and Duncan et al. (2011). For the purposes of this research, these processes are discussed under three headings; the data environment, the sensitivity of the data, and the characteristics of the data.

The data environment is a term most associated with researchers from the University of Manchester (Elliot et al., 2011; Mackey and Elliot, 2013). The data environment is an acknowledgement that data do not exist in isolation from each other. This concept has existed in the literature for some time (see Paass (1988) and Lambert (1993)). This is of significance for statistical disclosure control because the type of intruder attack upon which this research is focused is one in which an intruder attempts to match data they already have to anonymised sensitive data. This matching is done in order to expose information about individuals. When a data controller makes data or statistical outputs from that data available they, in effect, release that information into a data environment. On the face of it, attempting to analyse the entire environment would seem impractical because of its sheer size and the number of unknowns. Despite this, there have been attempts to produce methods for doing so, such as the automated Data Environment Analysis Project (ADEA) which is under development at the Cathie Marsh Institute for Social Research at Manchester University (Elliot et al., 2011).

The more traditional way of interpreting the data environment is for an organisation to have a well-documented collection of previously released data that could contain overlaps with the proposed data or output for release. In addition, some effort is required to simulate how an intruder might cause disclosure. This simulation hinges on what knowledge the intruder has available to them, and is often done with varying degrees of sophistication. At a basic level, data controllers can consider information commonly made available to the public, such as the electoral register. The register contains demographic details; your full name, address, nationality and age. From these basic details, data controllers can then build scenarios of intrusion. This ‘quasi-criminology’ approach is described in more detail in Elliot and Dale (1999) and also to some extent El Emam and Dankar (2008). This qualitative approach to capturing the data environment is favoured by this research, and it is demonstrated in case studies presented in this Chapter and Chapter 6.

It is also important that data controllers document how statistical disclosure control has been applied in previous releases to ensure an appropriate and consistent approach. This body of evidence also assists in the accurate assessment of the data environment and ensures that data perform consistently between different output reviews. In the qualitative interview evidence collected for this research, discussions with users and data controllers raised this as an area for concern by both groups. The main concern expressed was that due to the *ad hoc* nature of disclosure control assessment and implementation, many of the subjective judgements made by data controllers are not recorded. This can introduce confidence issues in the relationship between data users and data controllers which resonates strongly with the research problem. In the field, NHS Scotland has a documentation procedure within its disclosure control protocol (NHS Scotland ISD, 2012a).

Definitions of statistical disclosure do not distinguish between the level of detail in the data or its relative sensitivity, see for example Eurostat (1996). If an intruder has been able to re-identify an individual and learn some piece of confidential information, a disclosure has occurred. This definition provides data controllers with a substantial challenge in balancing the disclosure risk with data utility. If all information contained in data sets is treated as having the same impact upon disclosure, only the most conservative risk assessment and disclosure control can be applied. Instead, the literature suggests data controllers should make an informed judgement about the impact of disclosure in relation to the variables a data set holds. This will differ between data sets, variables, and their respective levels of detail. As the health sphere is the backdrop for this research, it is worth noting what the NHS considers to be sensitive data (NHS Scotland ISD, 2012a), see Figure 4.1.

Sensitive Topics

- sexually transmitted infections
- abortions
- suicides, self harm
- pregnancies under 16 years of age
- alcohol or drugs misuse
- mental health diagnoses and treatments
- prescriptions for contraceptives, mental health or any ‘sensitive’ condition
- crime related statistics e.g. gunshot injuries, assault, stabbings
- other sensitive diagnoses or treatments

Figure 4.1: NHS Scotland (ISD) Statistical Disclosure Control Protocol list of sensitive topics

As is also noted in the protocol, this cannot be a definitive list (NHS Scotland ISD, 2012a). To mitigate this lack of definition, data controllers should carry out some form of sensitivity analysis. The most robust method would be achieved by discussing the possible impact with the type of individuals represented in the data. When the number of records is in the millions, it is not feasible to do this with every individual but strategies such as the use of focus groups and surveys can help inform data controllers. For the social care, health and housing data linkage project, the Scottish Government commissioned a piece of research that explored the public perception of the project (Scottish Government, 2011b). In a similar vein, the Scottish Health Informatics Programme research group has carried out public engagement work that can shed some light in this area (Aitken et al., 2011). Their work highlights the public's concern about the use of patient data, especially when linked to other sources. Trust appears as a central concept for individuals; trust in the organisation holding the data; and trust in the organisations seeking to access the data. As Aitken et al conclude, data controllers and researchers must engage the public in these discussions of data security if that trust relationship is to be cemented and secured for the future. Drawing on the conceptualised balance of disclosure risk and data utility set out in Chapter 2, it is acknowledged that the balance is biased toward mitigating risk rather than data utility. This is in order to properly preserve the privacy of individuals, however data sensitivity analysis allows a data controller to configure and fine tune that balance more accurately and decide just how much bias is necessary.

Having discussed the need to analyse the environment into which data are being released and the sensitivity of the variables the data set contains, data controllers can focus on the data themselves. This might appear to be an obvious statement to make; however, measuring the risk presented by the variable values and combinations of values is a non-trivial task. Removal of direct identifiers is a given for all the data sets discussed in this research, and the disclosure risk presented by variables labelled as 'quasi-identifiers' such as the demographic variables discussed as part of the data environment is difficult to measure. Historically the focus of disclosure control had been on aggregate data, usually tables of counts representing underlying individual records (see the work of the group behind the τ-argus software in Hundepool and Willenborg (1996)). For this type of data, the data controller referring to the policy of their organisation, applied a threshold to counts in the aggregate table. This threshold was used to limit the possibility of unique counts, and therefore some form of attribute disclosure, or low counts that pose an unacceptable level of risk. Low cell counts in published tables can risk attribute disclosure through differencing with other data tables (this type of attribute disclosure is considered to be distinct from the re-identification problem in contemporary microdata). A commonly used example is a table that shows age and marital status, a sixteen year old widow would likely appear as a low count. The equivalent in the case of microdata would be to look for low instances of a variable value and low instances of a combination of values. Even for small data sets of c.100,000 records (The Scottish Homecare Census, for example) this is no trivial task if attempted under human analytical power alone.

This consideration that low cell counts represent the identification of rare cases in the population, is the basis for the *k*-anonymity method used in this research and discussed in Chapter 3. To extend that discussion beyond *k*-anonymity and its enhancements, it is also worth mentioning here that techniques which use probabilistic rather than observed methods exist in this field (Skinner and Shlomo, 2008; Carlson, 2002). However, these focus on the risk associated with sample survey microdata because of the ambiguity introduced by using a sample subset. In these cases a data controller needs to be able to estimate the risk that an intruder finds a match in the sample data that is also a correct match in the context of the whole population. A series of processes that, when combined, provide data controllers with a methodological framework for assessing the risk of disclosure, have been set out above. Where this type of framework has been used by organisations and embedded in their own processes will now be considered.

Skinner and Elliot (2002) provides us with a succinct summary of different approaches employed to measure disclosure risk. In order to discuss these approaches, some basic context for the data should be described. What is being considered here is a set of records that constitute a micro-level data set, which is itself drawn from a finite population. An example of this data context would be the individual level Sample of Anonymised Records (SARs); this is a 3% sample of the national Census and is maintained by the Cathie Marsh Institute for Social Research. One such method, the population unique method, revolves around the uniqueness of a record in the population; whether a record can be considered unique to a particular population if it is found to be unique in the sample. This presents a rather basic measurement of risk. For example, given a situation where an unauthorised person seeks to access the data for malicious purposes, and has knowledge that can identify a record in the population, they can also identify that record in the sample (if it is present) and disclose additional information about that record.

As discussed in Greenberg and Voshell (1990), this method has been used on large data sets, such as the US Census. Their primary focus in this paper was the effect of geographic region on the percentage of population uniques, and this provides a useful example of this method in practice. For a given geography, there might be 100,000 records of which 10% are unique to that geography, this 10% present a high risk of disclosure. Therefore, a data provider could use the above measure of risk to consider expanding this geography to a larger area in order to reduce the percentage of population unique records. In their conclusions, Greenberg & Voshell note that different combinations of identifying variables and geographies result in different degrees of disclosure risk; from this they stress the need to consider each data set individually; this is a subject that will be returned to in the later section on the practical implementation of SDC.

The sample unique method flows from the same logic as the population unique method. Instead of measuring the percentage of population uniques, the percentage of sample uniques contained in a given microdata release can be considered, as only these can potentially be population uniques. The risk in this method is expressed as the probability that a sample unique is also a population unique and that a successful match can be achieved by an attacker.

The third measure, as discussed in Skinner and Elliot (2002), is an attempt to construct a more nuanced approach to the measurement of risk. The problem the authors discuss is that thus far no inference method currently exists that can adequately predict the relevant values needed for the methods previously discussed without the use of a series of modelling assumptions. As a high level overview it is important to note that differing conclusions on the measurement of disclosure risk exist in the theoretical literature.

A brief discussion of the probability of a sample unique also being a population unique as an adequate measure is useful, despite Skinner & Elliot's description of this method as 'optimistic'. Optimistic in this sense means that the estimate of disclosure risk generated by this model should be seen as an under-estimate of the disclosure risk. To justify the inclusion of this 'optimistic' measure in this review of the literature, the empirical evidence from Skinner and Holmes (1998) is considered. In this study, the authors present an implementation of sample uniqueness as a measure of risk. This is demonstrated through an analysis of a 10% sample of data from the 1991 UK Census, that once sampled and analysed using a log-linear model, present accurate estimates for the number of sample uniques that prove to also be population unique. The actual results of this example find that approximately 0.2% of the 45,000 records sampled present a high-risk of disclosure, i.e. they present a large risk of being population unique as well as sample unique.

This section has provided an overview of the literature on measures of risk, where risk is framed as a collection of processes that encompass three distinct areas of activity: the data environment; the data sensitivity; and the characteristics of the data themselves. It has been shown that data controllers should acknowledge that their data does not exist in isolation and that attempts to capture the data environment allow for more sophisticated disclosure risk assessments. In addition, not all data are equal in terms of the impact a disclosure might have and therefore data controllers should conduct a sensitivity analysis at the variable level, which can be achieved through qualitative engagement with the data's subjects. Lastly, it has been established that data controllers should interrogate the data directly through quantitative testing. This testing can take the form of policy-driven thresholds, or probability driven metrics such as measures of sample-to-population unique estimates for example. Having now established some basis in the measurement of disclosure risk, a case study using the Scottish Health Survey is set out in the following subsection.

4.2.1 A Risk Analysis Example: The Scottish Health Survey (SHeS) 2003

To illustrate the disclosure risk assessment methods discussed in the previous section, a risk assessment example case study using publicly available data from the Scottish Health Survey

Quasi-identifiers
Sex
Age
Health Board
Marital Status
Ethnicity
Standard Occupational Classification

Table 4.2: Quasi-identifying Variables from the Scottish Health Survey (SHeS) 2003

2003 (SHeS) has been set out. The SHeS is a sample survey and participants are drawn from the Postcode Address File. It is not an annual survey and the collection periods have varied. In 2008 the SHeS changed to a continuous survey with a report published each year. The latest sample (2012) is not used as there is a particular interest in the treatment of Standard Occupational Classifications (SOC2000 in this case). Therefore the 2003 sample which is the most recent to carry the SOC codes was used. The survey provides a detailed picture of the health of the Scottish population and includes variables on physical activity, diet, and socio-economic factors. As a sample survey, were this a real disclosure risk assessment the risk is greatly reduced as the dataset does not represent the whole population. Therefore unique individuals in the sample might not be unique in the population. Also as a public release file, the survey content has almost certainly had some statistical disclosure analysis applied to it, so this analysis should only be viewed as a demonstration of the methods and concepts previously discussed. Due to the survey design the dataset is lopsided (in terms of the number of questions preceded by filter questions) this only affects the sensitive variables, i.e. the variables to protect from disclosure, therefore this should not impede the risk assessment.

The main concern throughout this analysis is the possibility of re-identification; could an intruder use the anonymised data and their own knowledge to identify individuals in the dataset and learn something new about them. As a first step, in this risk assessment the variables in the SHeS that might form quasi-identifiers, variables that an intruder can use for re-identification, have been considered. These are usually demographic details that could be easily obtained or known *a priori* by an intruder. The variables selected for this purpose are contained in Table 4.2. As discussed in the previous section, the data environment that the data is released into forms the basis of our reasoning for choosing quasi-identifiers. In this case, the decision was taken to use quasi-identifiers that mostly represent physical or spatial attributes that could be observed by an intruder or that they could know *a priori* from publicly available data. In this scenario the intruder could fit the journalist archetype fishing for information on an individual. In order to construct these quasi-identifiers they could use a public social media profile, knowledge from a neighbour or data from the electoral register.

This set of quasi-identifiers is quite limited in terms of scope for a realistic SDC analysis.

This is mainly because of the lack of geographical detail. However, there are enough variables here for an exploration of their respective influence on attempts to re-identify individual records. For a more realistic assessment of disclosure risk see the case study of education data in Chapter 5.

Using the prepared 2003 Scottish Health Survey, k-anonymity, where k=3, was implemented. For this implementation the set of key variables were; age, sex, marital status, SOC2000 code, and ethnicity. The NIAH suite of tools were used to partition the dataset into those records that meet the k=3 requirement and those that do not. This means that records are only deemed safe if there are at least three instances of their combination of key variables values. This partitioning provides two datasets which can be analysed to highlight where the disclosure risks are concentrated. The first step in this analysis is a comparison of summary statistics for the raw data, the 'safe' records and those deemed 'at risk'.

An initial observation is that for k=3, a relatively loose privacy requirement in k-anonymity terms, the number of records deemed 'at risk' is very high. Of the $7,897$ original records[1], $7,854$ (99 per cent) are deemed 'at risk'. This is likely to be due to the high level of detail provided by the SOC2000 codes. However, it is important to clearly understand the k-anonymity output in this case. Those records marked as 'at risk' do not satisfy the k=3 criteria, but at this stage in the analysis there is insufficient information to state that these records present a credible threat to privacy. To illustrate this point further the unknowns can be stated: it is not known which variable or combination of variables has the strongest influence on a record's position of being safe or at risk; the distribution of key variable values for those at risk records is unknown; and the potential interaction effects of the key variables are not known. It is therefore clear that in order for a disclosure control decisions to be made, the risk assessment must probe the k-anonymity results at a deeper level. In this vein consider the following summary statistics. Table 4.3 shows us the comparison of averages for the three versions of the data with missing and non-applicable values ignored.

Given that there are only 1 per cent of safe records to compare against it is not surprising that the at risk records look like the original data. However, it is interesting to note the composition of the remaining 43 safe records. The homogeneity present in the original data is further exaggerated here, the records are 100 per cent white and most likely married. Also it is worth noting that despite the small number of records the mean age does not shift away from the original data significantly. This again is unsurprising as it is expected that the only records to satisfy the k-anonymity requirement will be clustered around the average value on each of its key variables. To follow on from the hypothesis that the level of detail in the SOC2000 codes could be the biggest factor in the k=3 results, the same measures as in Table 4.3 are used in Table 4.4 with SOC2000 removed from the key variables. From this second

[1] Children have been removed from the data for the following analysis, as earlier experiments showed that they skewed the data for the age variable significantly

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=43)	NIAH At Risk Output (N=7854)
Age			
Mean($\pm1\sigma$)	(33.3) 50.6 (67.9)	(31.4) 50.6 (69.7)	(33.3) 50.6 (67.9)
Sex			
Mode (% of N)	Female (56%)	Female (93%)	Female (55%)
Marital Status			
Mode (% of N)	Married (57%)	Married (65%)	Married (57%)
Ethnicity			
Mode (% of N)	Scottish (84%)	Scottish (100%)	Scottish (84%)
SOC2000			
Mode (% of N)	7111.Sales and retail assistants (5%)	9233.Cleaners, domestics (44%)	7111.Sales and retail assistants (5%)

Key Variables: Age, Sex, Marital Status, Ethnicity, SOC2000, Health Board.

Table 4.3: SHeS: Summary of NIAH Output for all Key Variables and Health Board, *k*=3

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=3603)	NIAH At Risk Output (N=4294)
Age			
Mean($\pm1\sigma$)	(33.3) 50.6 (67.9)	(34.7) 50.3 (65.8)	(32.3) 50.9 (69.9)
Sex			
Mode (% of N)	Female (56%)	Female (57%)	Female (54%)
Marital Status			
Mode (% of N)	Married (57%)	Married (79%)	Married (37%)
Ethnicity			
Mode (% of N)	Scottish (84%)	Scottish (98%)	Scottish (73%)

Key Variables: Age, Sex, Marital Status, Ethnicity, Health Board.

Table 4.4: SHeS: Summary of NIAH Output by Health Board without SOC2000, *k*=3

partitioning for k=3 shifts in the marital status and ethnicity variables can be seen. Also it should be noted that age has remained largely static across the partitions. This is of interest because the hypothesis regarding SOC2000 was based on the number of possible categories. There are 353 SOC unit codes in SHeS, and 81 possible ages in the data set. If the number of categories alone was a decisive factor it would be reasonable that age might present a greater challenge. However, another factor in this hypothesis should be the relatively sparsity of the values. SOC2000 has a number of sparsely populated categories among the 353 possible unit codes.

Also in both of these iterations of k=3 no details of the geographic variable health board, have been included. Figure 4.2 provides the frequency distribution for Health Board for the at risk records for both iterations. Despite the big change in N for the two at risk partitions the distribution of health board has been largely unaffected.

From the above summary statistics the SOC2000 codes appear to be providing the majority of the disclosure risk for k=3, followed by some influence from marital status and ethnicity. Although age has stayed relatively static, the analysis was repeated dropping age from the key variables to see what effect that had. The results of this are presented in Table 4.5. In Table 4.5 the level of detail in the SOC2000 codes was reduced to represent just the two digit sub-major encoding. The descriptive summary indicates that age has little effect when removed from the k-anonymity assessment. There is a small increase in the size of the safe partition, however this could also be due to the recoding of SOC2000. To scrutinise the relationship between age and the partitions further a logistic regression model was fitted for the effects of age on a binary 'at risk' flag which denotes which partition the record was in after the k-anonymity analysis. This results in a regression coefficient of 0.0002691 and an R^2 value of 0. This indicates that the effect has no statistical significance which would support the descriptive results above. This type of scrutiny was also applied to the SOC2000 codes. Table 4.6 shows the effect of the SOC major groups on whether or not a record is considered at risk or not (our reference category here is the lack of a recorded SOC group). All of the SOC groups return a statistically significant result, however the R^2 only indicates that the SOC groups alone account for 1% of the variation in the at risk variable. This low R^2 is perhaps unsurprising when it is already known that the at risk variable is a product of a complex cross classification of all of the key variables.

To reinforce the conclusion that the SOC variable is the major influence on the partitioning of the data into safe and at risk records, the SOC variable has been removed completely from the key variables. Figure 4.3 presents three scatter plots showing age plotted against health board for the original data, the safe partition when the SOC sub major groups are included and the safe partition when no SOC codes are included. Although the scatter plots do not give a particularly instructive description of the health board profile, the contrast of the three scatter patterns clearly highlights the effect of the SOC variable in reducing the availability

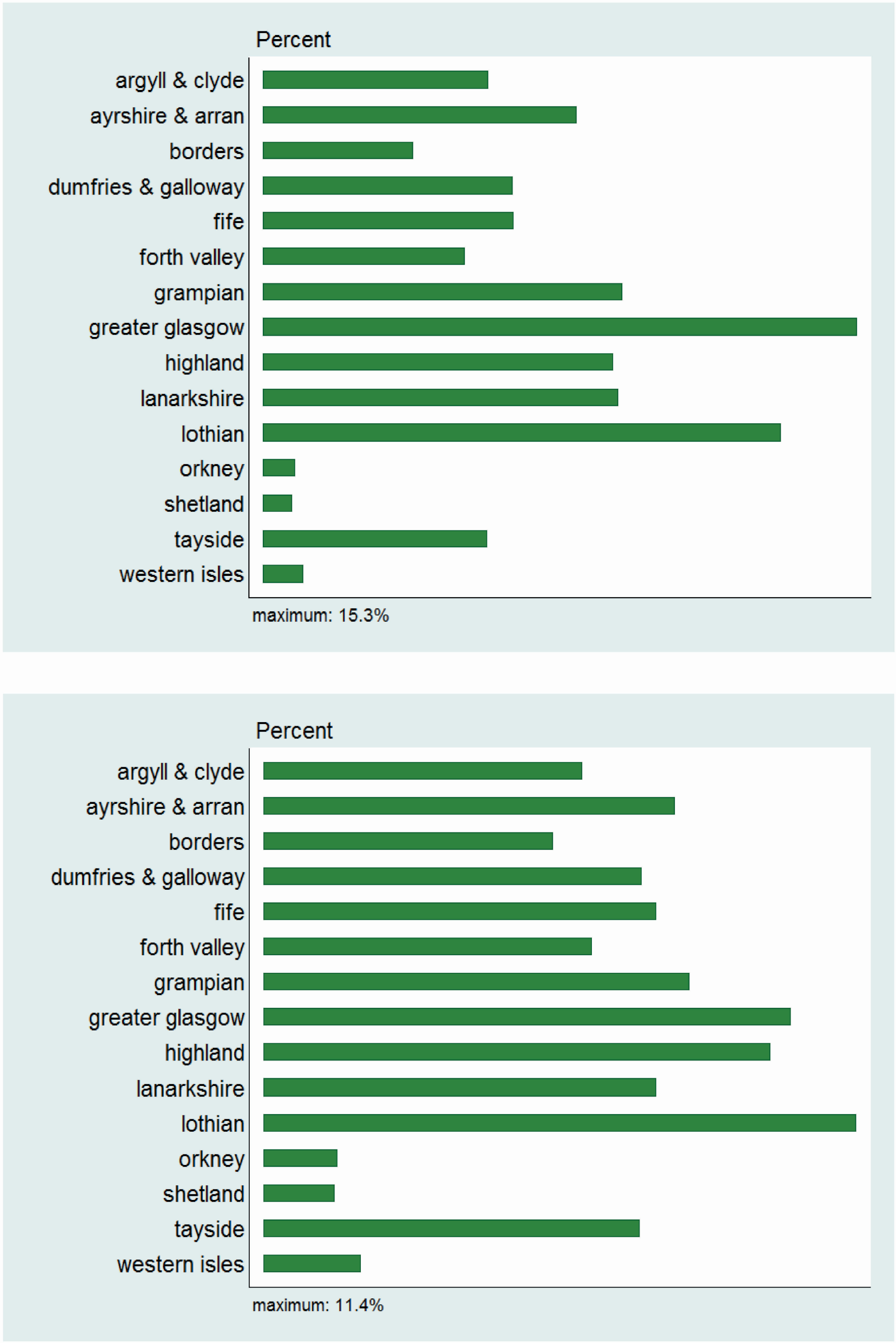


Figure 4.2: SHeS: Frequency Distribution of Health Board for 'at risk' Records, all Key Variables (N=7854) and Frequency Distribution of Health Board for 'at risk' Records, SOC2000 removed from Key Variables (N=4294)

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=383)	NIAH At Risk Output (N=7514)
Sex			
Mode (% of N)	Female (56%)	Female (60%)	Female (56%)
Marital Status			
Mode (% of N)	Married (57%)	Married (80%)	Married (56%)
Ethnicity			
Mode (% of N)	Scottish (84%)	Scottish (100%)	Scottish (84%)
SOC2000 Sub-Major Groups			
Mode (% of N)	92. Elementary administration & service occupations (12%)	92. Elementary administration & service occupations (27%)	92. Elementary administration & service occupations (11%)

Key Variables: Sex, Marital Status, Ethnicity, SOC Sub-Major Groups, Health Board.

Table 4.5: SHeS: Summary of NIAH Output by Health Board, SOC2000 Recoded and Age Removed from Key Variables, *k*=3

At Risk	Coef.	Std. Err.	t	P>$\|\|t\|\|$	95% CI.	
Administrative and secretarial occupations	.0299	.0050	5.95	0.000	.0200	.0397
Associate professional and technical occupations	.0330	.0050	6.51	0.000	.0230	.0428
Elementary occupation	.0182	.0049	3.72	0.000	.0086	.0277
Managers and senior officials	.0330	.0051	6.46	0.000	.0230	.0430
Personal service occupations	.0330	.0053	6.22	0.000	.0226	.0434
Process, plant and machine operatives	.0330	.0052	6.39	0.000	.0228	.0430
Professional occupations	.0292	.0051	5.70	0.000	.0192	.0393
Sales and customer service occupations	.0173	.0054	3.21	0.000	.0067	.0278
Skilled trades occupations	.0330	.0051	6.49	0.000	.0230	.0429
constant	.9670	.0044	218.28	0.000	.9583	.9757

Table 4.6: SHeS: Logistic Regression showing the effects of SOC2000 Major Groups on the Records marked as ‘at risk’, (N=7897)

safe data. Having established a preliminary survey of the disclosure risk for the 2003 Scottish Health Survey data, these results were then used to inform the the SHeS case study which continues to explore disclosure control methods in an applied context, see section 4.3.1.

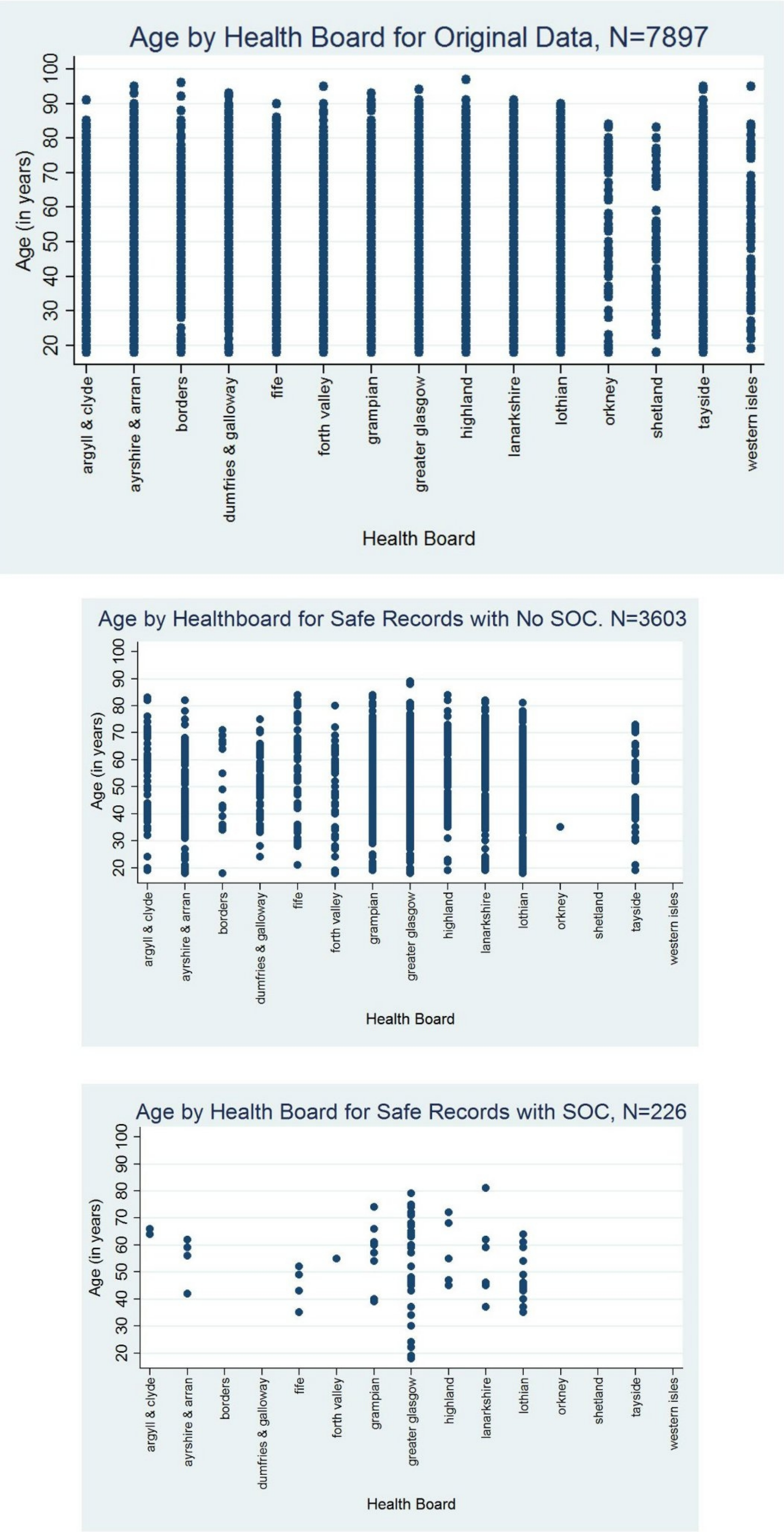


Figure 4.3: SHeS: Scatter Graphs of Age by Health Board for Original, ‘safe’ and ‘at risk’ Data

4.3 Disclosure Control Methods

Having discussed approaches to estimating disclosure risk, this section reviews existing methods devised to prevent statistical disclosure. These can be divided into three distinct groups; recoding and suppression; the addition of noise to disguise the data; and the less common approach of generating synthetic data. Fienberg and Willenborg (1998) uses a similar classification of SDC strategies setting out three approaches: "Reporting only a subset of the data, by selection of cases and/or variables; modifying the data in some form; Not reporting the observed data at all, but only 'pseudo data"' (Fienberg and Willenborg, 1998, 338). Each group will be discussed and the positive and negative aspects of their approaches considered before summarising within the context of this research and the types of application these methods could be used for. However, it is important to note, as Fienberg does, that these approaches are not mutually exclusive and data controllers can use combinations of methods from all three.

Recoding and suppression are techniques common to data analysis and provides perhaps the most intuitive methods for statistical disclosure control (SDC). Recoding is also frequently used by researchers to prepare their data for analysis, and this forms a vital part of the data management process. For a useful discussion of what is meant here by data management, with a worked through example, see Lambert and Gayle (2008). As intuitive methods that feature in common data workflows, recoding and suppression for SDC have existed in the literature since SDCs inception, for example see the discussion of data suppression and 'rolling-up' (recoding) in (US Federal Committee on Statistical Methodology, 1978) and a more contemporary example in Willenborg and De Waal (2001). Ultimately, if data are too sensitive to release they are suppressed by data controllers. Suppression as a method needs little further scrutiny apart from to note that it operates as a binary switch between disclosure risk and data utility for those values that are suppressed. However, the consequences of choosing to suppress some data and not others does provide a more complex challenge to the data's utility. For example, if a variable by case matrix were constructed representing 100 per cent data utility and then a series of counts for occurrences of a particular variable value (or combination of values) were produced. These counts might indicate that only a small number of cases display this combination and deem those cases as a potential disclosure risk (this represents a simplistic version of the disclosure risk assessments illustrated in Section [3.a]). At this stage a data controller could simply suppress these records or blank out their values for variables deemed to be at risk. As Fienberg et al note, this particular strategy has a significantly negative impact on the data's utility: "The principal problem we have with cell suppression as a method is that it intentionally 'distorts' the information in the table by purposely selecting cells to suppress. As a consequence, users can be led into misleading and, in particular, biased inferences on the basis of the cell values that are reported" (Fienberg and Willenborg, 1998, 488). For this reason the rather blunt instrument of record or cell suppression is often only seen as a 'last-resort' in the SDC literature. Despite this it does feature in many data controller statistical disclosure control protocols and is often used in tabular data releases, for example see Scottish Government (2007).

Recoding is another possible technique that data providers can deploy and it covers a number of different sub-types. Winkler (2004) offers a succinct summary of some of these methods. Macrolevel recoding, referred to by Winkler as 'global recoding', addresses the level of detail presented in a variable. This involves the aggregation of possible values into higher level categories. As indicated earlier, this is often done by researchers in the preparation for their analysis. For example, a measure of occupations such as the International Standard Classification of Occupations (ISCO) can be collapsed or expanded to fit the requirements of the researcher in furthering their analysis and providing a clearer narrative. In the SDC case, data controllers restrict access to the lower levels of the hierarchy. In practice, data that contains age as an individual integer frequently could represent a disclosure risk. If the highest age values are relatively rare in the distribution, suppressing this value could drastically alter the statistical properties of the data. Therefore as an alternative, ages could be grouped by recoding the individual integers into bands of an appropriate length. Another common example of this technique is employed when data carries geographic identifiers: if the number of cases within a particular geography is considered to be too small, the geographic variable can be recoded to a higher level of abstraction, for example from post code to parliamentary constituency in the UK context (Witkowski, 2008).

In the geographical case, this type of global recoding can prove problematic because different geographical hierarchies do not match up exactly, or additional geographical detail from other variables could increase the risk of disclosure rather than reduce it (Steel and Sperling, 2001). There is also some debate over the assumption that disclosure risk scales with geographical detail. As suggested by Greenberg and Voshell (1990), intuitively the number of rare records should decrease as the size of the geographical area, and the underlying population, grows. However, as Elliot et al. (1998) suggests this relationship can be non-monotonic. Although this approach is easily implemented in most instances, it does create problems for the researcher trying to utilise the data for analysis. These problems include the aforementioned loss of information at the top or bottom end of a distribution and also that the aggregation of values prevents finer-grained analysis. As an alternative to macro level recoding, microlevel aggregation performs a similar coarsening of detail but this is achieved as a function of the data values themselves. As Domingo-Ferrer and Mateo-Sanz (2002) sets out, records can be grouped based on similarity and then average values can be calculated for each group and these average values replace the original record values. To the user of the data, records will seem more homogeneous than they are in reality, and the disadvantage over global recoding is that the user cannot easily attach this homogeneity to a variable hierarchy that they recognise. Domingo-Ferrer and Mateo-Sanz (2002) primarily deals with the numerical case, however the same type of simple aggregation has been applied to categorical data as well, see (Mares and Torra, 2012).

Data-swapping can be considered a sub-type of recoding. This type of method has been used by national statistics agencies including the US Census Bureau and the Office for National Statistics (Zayatz et al., 1999; Office for National Statistics, 2011). In order to protect the data, variable values are swapped between a pair of records. This can be done with reference to some other variable (often referred to as a control (Dalenius and Reiss, 1982)). For example, if income is considered sensitive then a pair of cases selected from different geographical areas could have their values for income swapped. These swaps are often done systematically and therefore it is important that the data provider keeps the details of their swap system confidential so that they cannot be undone. This technique targets and distorts the data on the basis of a few cases and variables. This distortion has to be managed, for example, if given a control such as geography, the values of a sensitive variable would ordinarily follow some identifiable distribution, then the swapped-in values would seem anomalous and could indicate to attackers that a swap has occurred. This distortion would also effect the results of statistical analysis, or create impossible situations; records for prisoners could be swapped to an area that has no prison. There is a difference in the way that data swapping protects the data from disclosure when compared with global recoding. As opposed to suppression or recoding where values are removed or coarsened, data swapping creates uncertainty. Census outputs are usually samples so an intruder has to grapple with the sample unique versus population unique problem, and by swapping a percentage of the records intruders cannot be certain that their potential match has not been swapped.

New methodologies that seek to address the problems of record swapping include the 'data-shuffling' methodology proposed by Muralidhar and Sarathy (2006); Muralidhar et al. (2006) which was evaluated using Irish Census data in McCaa et al. (2013). Shuffling reassigns the values of individual records to different records with reference to the rank order correlation of the data set as a whole. Muralidhar and Sarathy (2006) suggests that this ensures that all monotonic relationships between variables are maintained which provides a higher level of data utility compared to swapping. On evaluation this claim does stand up to scrutiny at least for the example statistical analyses carried out by McCaa et al. (2013).

At this juncture, the majority of disclosure control methods applied in practice has been covered. Many data controllers rely heavily on suppression and recoding, some argue that these methods provide the only robust solution to the privacy versus utility problem that have a predictable impact on the data utility (Cleveland et al., 2012). However, the literature continues to supply older methods, like noise addition, and newer methods such as synthetic data generation.

The addition of noise as a technique for disclosure control has been established for some time (Kim, 1986). For continuous numerical data this method is reasonably simple to implement. Consider a variable by case matrix that contains values for age and income, to obscure the original data one could increase each value by 10% (or any appropriate percentage) so it is no longer possible for an attacker to identify the individuals in the sample using exact matching. As with data-swapping this addition of noise is deployed systematically, and therefore the data provider must keep that system secret. In reality applications of this method are more complex and involve adding noise relative to the covariance of the original data, so that covariances and means remain accessible (Yancey et al., 2002). Proponents of this method provide evidence that noise added in this way yields strong analytical validity, however as noted by Winkler (2004) specialist software is required to analyse data that has had this method applied. As such, it is an unpopular choice because it forces the user to operate outside their normal software environment.

Noise addition can also be implemented for categorical data, although a categorical variable cannot be increased by 10% because the values are discrete and may contain no order or hierarchy. Gouweleeuw et al. (1998) proposed an implementation of the Post Randomisation Method (PRAM) in order to apply the concept of noise addition to categorical data. Noise in this case is regarded as the deliberate misclassification of the value(s) for a given record in a given variable. Gouweleeuw *et al* provide a neat example to illustrate this, which is summarised here. In this example the variable in question is gender with two possible scores, male and female, which are assigned the value 1 and 2 respectively. PRAM is applied as a probability mechanism for the number of possible values, in this case it is generated as 0.9. Consider a dataset that holds 100 males and 100 females. The new disclosure controlled dataset would still have 100 males and 100 females, however 10% (the inverse of the probability mechanism value) would be misclassified. Therefore in the controlled data 10 females are actually male and 10 males are actually female. As the above example illustrates this method lowers the probability that an attacker can make an exact match using their external knowledge because the misclassification creates uncertainty. This method has potentially severe implications for variables that have a strong correlation, for example if another variable in the data contains values that are gender specific, such as number of pregnancies, an independent variable misclassification would result in a number of males with a number of pregnancies greater than zero.

In addition, PRAM is very dependent on the selection of the probability mechanism, an issue highlighted in Domingo-Ferrer and Mateo-Sanz (2002) and De Wolf et al. (1999). De Wolf et al also reflect on another disadvantage of PRAM; In order to achieve realistic results from statistical analysis of a dataset that has had PRAM applied, standard statistical methods have to be altered to mitigate the PRAM's effect on results. The data provider could issue the controlled release data with details of the Markov matrix or its inverse so that the researcher can estimate the variance from the truth created by PRAM. This has a significant risk attached to it as a user with the Markov matrix could use this to undo the effects of PRAM at the individual level. As a partial solution it is suggested the data controller could provide multiple controlled datasets so that a researcher could conduct analysis across the set and take an average of the results. Again this carries a high level of risk. An intruder could analyse the multiple data files and given that the value for a given record is unlikely (due to the randomisation element) to be misclassified across all the data files, one could simply calculate the value that occurs most frequently and take this as a strong candidate for the real value. This method of using multiple datasets also draws strong parallels with the literature of multiple imputation used in the treatment of missing data, see (Rubin, 1987; Schafer, 1997; Schafer and Olsen, 1998).

So far the above methods attempt to obscure the real data by altering the values to avoid record matching by an attacker with some *a priori* knowledge. These methods are employed by national statistics agencies and other data controllers, and to date the number of disclosures is very small when compared with the sheer amount of data that has been collected. This is confirmed in Committee for the Coordination of Statistical Activities (2014, 2), in which the Committee for the Coordination of Statistical Activities acknowledged that “[w]hile there has been no major incidents reported so far, this risk cannot be ignored. In response to these risks many data producers and depositors have adopted a conservative approach by severely limiting or excluding access to their microdata”. However, it should be noted that the primary focus of the methods in this section were static public release files where SDC methods are applied once before data are released. The advances in data availability and data-linkage discussed in the introduction to this research mean that data controllers need to be able to perform SDC on outputs generated in secure-settings rather than one-off public release files. This also coincides with the increase in computational power of the would-be-intruder. As noted in a number of the papers cited above, the scenario that has been the main focus of work in this area is one in which the attacker employs exact matching techniques for re-identification. In modern data management and analysis, exact matching is now combined with probabilistic data-linkage methods to provide far greater accuracy. Probabilistic matching (sometimes referred to as fuzzy matching (Bell and Sethi, 2001)) matches records that, although not identical, have a high probability of being a correct match (Fellegi and Sunter, 1969). Jaro (1995) provides an early implementation of this method in the linkage of large public health micro-level datasets. With the development of probabilistic linkage, the data masked by SDC can potentially be unmasked if an attacker treats recoding, data-swapping, and noise injection as simply misclassification or data collection error. Although this may not provide an attacker with exact matches to their a priori knowledge it may provide a much smaller subset of the population that could be correct matches. Further methods could then be utilised, for example triangulation through the analysis of other external data to reduce the subset of possible matches. This also has implications for what is considered a disclosure breach, if an attacker can obtain a subset of individuals that are potentially correct matches; could that be considered a disclosure by a data controller. This issue is discussed further in the context of PRAM in Shlomo and Skinner (2010).

The above discussion sets the scene for the third SDC strategy. The creation of synthetic datasets for public release is still a relatively new process especially outside of the United States. This method was born out of a developments in data management and analysis intended for the treatment of missing data. For a comprehensive review of dealing with missing data see Schafer and Graham (2002). Synthetic data is data generated in an attempt to maintain the relationships, and therefore the validity of inferences, without being 'real' in the sense that it does not represent the observed values of any one individual in a population. Reiter (2002) discusses the creation of synthetic datasets:

> "In this approach, the agency selects units from the sampling frame and imputes their data using models fit with the original survey data. The approach has three potential benefits. First, it can preserve confidentiality, since identification of units and their sensitive data can be difficult when the data for some or all of the variables in the data set are not actual, collected values. Second, with appropriate estimation methods based on the concepts of multiple imputation [...], the approach can allow data users to make valid inferences for a variety of estimands without placing undue burdens on these users. Third, synthetic data sets can be sampled by schemes other than the typically complex design used to collect the original data, so that users of synthetic data can ignore the design for inferences" Reiter (2002, 2).

Having described synthetic data it is necessary to briefly describe methods for its generation. There are two proposed methods in the literature. The first is drawn from the multiple imputation literature on missing data. Drechsler et al. (2008) used the German Institute for Employment Research (IAB) Establishment Panel. The data controller first takes a sample of the population data and treats the rest of the data in the population as missing. The values for this 'missing data' are then imputed. For a given missing value, a number of imputations are generated, drawn from a probability distribution for that variable based on the other variables and values present in the sample. These imputations are then combined using a set of rules to allow analysts to carry out their analysis as they would normally. A number of the imputed datasets are created and then these are released to the public. The accuracy of statistical analyses is preserved by the combined effect of the probability distributions, dependent on the number of datasets released. The results in this example looked at standard statistical analyses (regression analyses and summary statistics). The results from the synthetic data were highly accurate, the majority of analyses returned significance of the same magnitude, and those that did not were only slightly attenuated,for example with significance values at the 5% level rather than 1% level. This level of accuracy is encouraging, especially when the disclosure risk is very low, because the released data does not contain any real representations of individual records.

Drechsler's example involved the imputation of all variables in the release dataset. The partial synthesis of data has also been proposed. Partial in this context is described as a selection of variables considered to be of high disclosure risk (Reiter, 2003). The intruder does not know which variables have been synthesised and which are original. The most significant disadvantage of this approach is that the accuracy of the synthetic values, in terms of relationships preserved, is reliant on the model which generates the new values. Also statistical models feature parametric assumptions, which may or may not actually be true for the target data. In terms of efficiency, this limits the usefulness of this approach if a data provider has a significant amount of data to which SDC must be applied. Recognising that this problem exists, Reiter and other authors began working on non-parametric implementations for the generation of synthetic data. This forms the second generation method, the non-parametric model considered here is the Classification and Regression Tree (CART) model used by Reiter (2005). The CART model is an approach borrowed from machine learning, which breaks the prediction of values down using a decision tree. The model can handle categorical and numerical data, by using the classification and regression aspects of the model respectively. In the context of SDC, fitting CART models to a sample of the population allows for the prediction of the values for a given set of identifying, or sensitive, variables. The key to the model's use in SDC is the pruning of the tree structure; this is done to avoid over-fitting the model which might return the real values of the population, which would defeat the models purpose. Reiter provides some advice on this topic, however the pruning tends to be governed by some external disclosure threshold or similar measure. Once the necessary tree has been constructed and pruned appropriately, it is used to generate estimates for the values needed. From this pool of estimates, the values to be imputed are selected at random. At this point, the analysis of this synthetic data can be done using standard statistical methods that a researcher would use on any data.

As before, the data controller might issue a number of these datasets so that a researcher can combine the results of analyses and increase the accuracy. There are various advantages to this method; it is not beset by parametric assumptions; the generation of data is semi-automatic as these trees can be generated by computer without the need to fit complex statistical models; the accuracy of the resulting analyses has proven to be good in the literature. However, it also carries disadvantages; it is possible that CART will not capture all the relationships in the population (this relies of the selection of the sample). If the relationship is not present when the tree is grown, it will not be captured. Another perhaps less technical disadvantage of synthetic data in general is its 'unnatural' feel. This second disadvantage is said to have the bigger impact (Reiter et al., 2009). If the simulated evidence presented in the methodology literature is correct, then this problem would seem to be one of perception rather than substance. However, until more studies involving synthetically generated data are published, this problem will likely persist.

In this section, the existing literature on robust disclosure control methods has been reviewed. It has been shown is that the simplest methods, such as suppression and recoding, often have the greatest impact on data utility in order to protect confidentiality, yet are still considered the methods of choice for data controllers. The more recent developments in terms of synthetic data might provide a better balance of risk and utility given the contemporary context. However, this approach has yet to prove itself popular, especially with researchers. Data shuffling might also prove to be fruitful, however the lack of clear documentation for its implementation prohibits its use here. As this research seeks to provide a set of algorithms that allow data controllers to review outputs, rather than create static public (or restricted) release files, it was decided that the robust methods currently used by data controllers were the best candidate for development in response to the research problem. These are well documented and tested and allow for comparison between existing approaches and the approach set out in this research. As will be shown in the next sub-section, these traditional methods are applied to the Scottish Health Survey as an example case study, continuing from the estimation of risk in 4.2.

4.3.1 A Disclosure Control Example: Scottish Health Survey continued

In this section the result of the risk analysis in section 4.2.1 are combined with some of the techniques reviewed in the previous section to generate potential disclosure control scenarios for the Scottish Health Survey (SHeS) data. To briefly summarise the results from section 4.2.1, it was shown that the SOC2000 (Standard Occupational Classification) codes are the primary focus of disclosure risk for a k-anonymity assessment with k=3 and the key variables: age, gender, health board, marital status, ethnicity and SOC2000. In addition to SOC, age was also considered as a target for disclosure control because of the large number of age categories when age is presented in single years. These will be the targets for statistical disclosure control methods. Using the NIAH tool-set (detailed in Chapter 6) banding, recoding, noise addition and random misclassification of categorical variables will be implemented. K-anonymity assessments will also be used to illustrate the effects of these disclosure control methods, however a more detailed analysis of the resulting data's utility will be discussed in section 4.4.1.

The first method implemented was banding; this consists of grouping variable values into higher level categories. The age variable was coarsened from single year age to age in 5 year bands. This type of coarsening is often carried out by analysts to structure research data in a particular way to help with the clarity of their analytical narrative (for example see the use of age groups in Townsend et al. (1994)). For this reason, it is the least radical of the disclosure control methods, which is an important consideration for data controllers, because it provides common ground on which to discuss disclosure control with data users. Age is often the target of this method of disclosure control because integer age values can be easily collapsed into categorical age groups (see Willenborg and De Waal (1996)). As a first iteration, age was grouped into 5year categories, with the exception of the first and last category; the data excludes children and is bottom-coded from 21 and younger, and top-coded for ages 97 and above.

The SOC2000 values are recoded from four digit unit codes to two digit sub-major groups. The structure of the SOC2000 variable makes this transformation simple to implement as the 2 digit sub-major group can be extracted from the 1st two digits of the unit code. This hierarchical structure lends itself well to coarsening the detail in this way. However, as will be discussed later in this section, it can cause complications for other disclosure control methods that fail to capture this hierarchy and make illogical changes to data. For example, the data values might be changed to those outside a hierarchical sub-grouping. In addition, different codes at the same level of the hierarchy are not symmetrical. For example, some four digit codes could cover broad enough categories to make collapsing them unnecessary, while some two digit codes might still provide too much specificity when tested for statistical disclosure.

Table 4.7 provides the results of a k-anonymity assessment based on the data reconstituted with 5 year age categories and SOC sub-major groups. Again, a high proportion of the records do not meet the k=3 criteria. Despite the recoding of SOC2000 codes, it still appears to be the variable that carries the greatest disclosure risk when combined with the other key variables.

To illustrate this further, table 4.8 shows the same descriptive analysis for the recoded dataset without the SOC codes included in the key variables. The level of records that do not satisfy k=3 are reduced significantly. In addition the partitions also provide enough depth to highlight the shifts in modal values between safe and at risk. For example, the table seems to indicate that a safe record is more likely to be married and ‘white: Scottish’. The geographical variable, health board, that is not shown in the table, is shown in Figure 4.4, which provides a comparison of the health board distribution across all records and the safe records. From this comparison, it is clear that the smaller health boards such as the Western Isles disappear almost completely and those records that remain are concentrated into larger health boards like Greater Glasgow.

Table 4.9 pursues a further compromise of the age variable detail in an attempt to retain SOC at the sub-major group level. Ages has been grouped into 10 year categories and top

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=1631)	NIAH At Risk Output (N=6266)
Age			
Mode (% of N)	37-41(10%)	52-56(14%)	37-41(10%)
Sex			
Mode (% of N)	Female (56%)	Female (63%)	Female (54%)
Marital Status			
Mode (% of N)	Married (57%)	Married (82%)	Married (51%)
Ethnic Group			
Mode (% of N)	White: Scottish (84%)	White: Scottish (99%)	White: Scottish (80%)
SOC2000 (Sub-Major Groups)			
Mode (% of N)	92. Elementary administration and service occupations (12%)	92. Elementary administration and service occupations (21%)	92. Elementary administration and service occupations (10%)

Key Variables:Age(5yrs), Sex, Marital Status, Ethnic Group, SOC Sub-Major Groups, Health Board.

Table 4.7: SHeS: Summary of NIAH Output using Recoded Age and SOC2000 Sub-Major Groups, *k*=3

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=6156)	NIAH At Risk Output (N=1741)
Age			
Mode (% of N)	37-41(10%)	37-41(10%)	37-41 & 52-56(10%)
Sex			
Mode (% of N)	Female (56%)	Female (57%)	Female (52%)
Marital Status			
Mode (% of N)	Married (57%)	Married (65%)	Married (28%)
Ethnic Group			
Mode (% of N)	White: Scottish (84%)	White: Scottish (92%)	White: Scottish (55%)
SOC2000 (Sub-Major Groups)			
Mode (% of N)	92. Elementary administration and service occupations (12%)	92. Elementary administration and service occupations (12%)	92. Elementary administration and service occupations (11%)

Key Variables:Age(5yrs), Sex, Marital Status, Ethnic Group, Health Board.

Table 4.8: SHeS: Summary of NIAH Output using Recoded Age and SOC2000 has been Removed from Key Variables, *k*=3

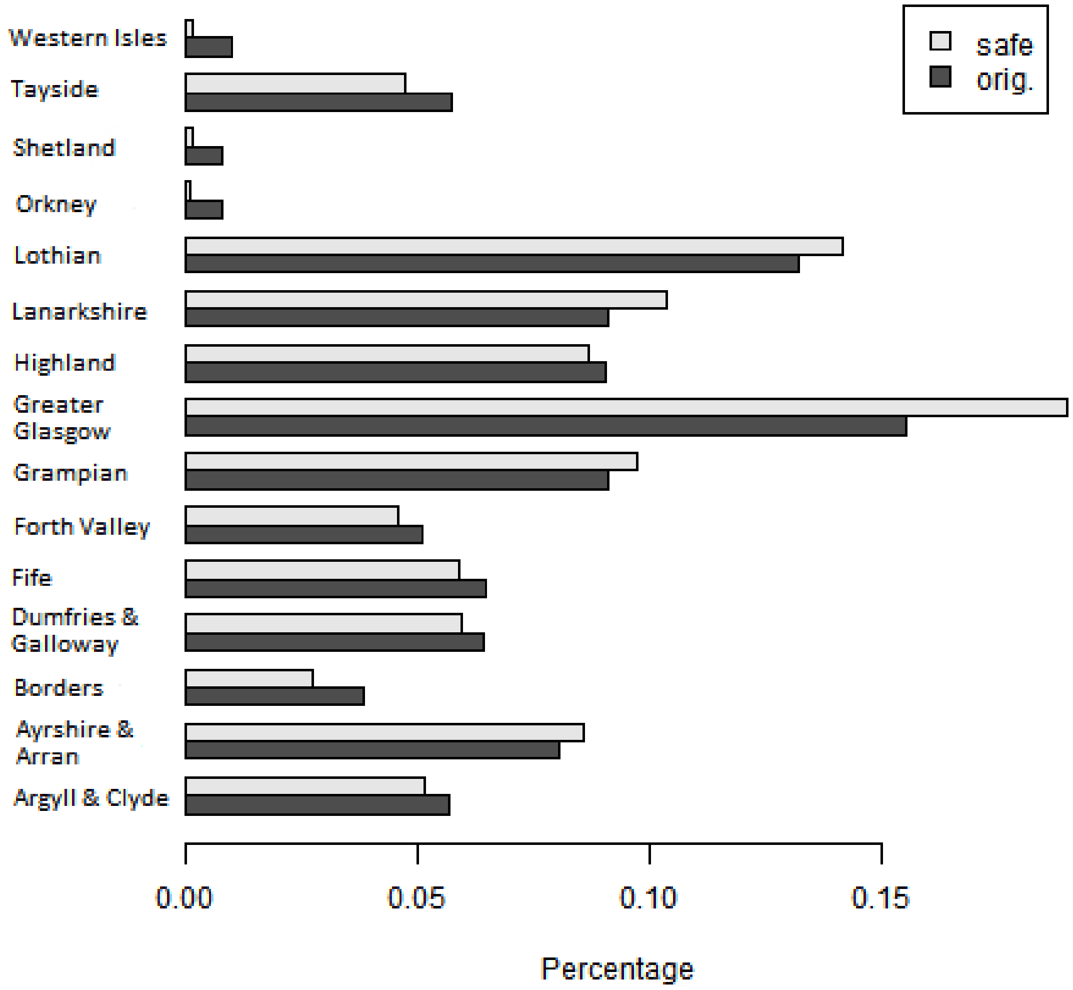


Figure 4.4: SHeS: Frequency Distribution of Health Board for All Records (N=7897) and ‘safe’ Records (N=3603)

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=3016)	NIAH At Risk Output (N=4881)
Age			
Mode (% of N)	38-47(20%)	68+(22%)	38-47(20%)
Sex			
Mode (% of N)	Female (56%)	Female (60%)	Female (54%)
Marital Status			
Mode (% of N)	Married (57%)	Married (77%)	Married (44%)
Ethnic Group			
Mode (% of N)	White: Scottish (84%)	White: Scottish (98%)	White: Scottish (76%)
SOC2000 (Sub-Major Groups)			
Mode (% of N)	92. Elementary administration and service occupations (12%)	92. Elementary administration and service occupations (17%)	92. Elementary administration and service occupations (9%)

Key Variables:Age(10yrs), Sex, Marital Status, Ethnic Group, SOC Sub-Major Groups, Health Board.

Table 4.9: SHeS: Summary of NIAH Output for Age Recoded into 10yr categories and SOC2000 Sub-Major Groups, *k*=3

coded from 68 years and upward. The disclosure risk situation, in terms of k-anonymity, has improved with 38% of records deemed safe at k=3. For comparison, Table 4.10 repeats the k-anonymity assessment removing the SOC variable. As is expected the size of the at risk partition reduces still further from that presented in Table 4.9. It can also be verified that the health board distribution has been largely unaffected by these changes.

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=6959)	NIAH At Risk Output (N=938)
Age			
Mode (% of N)	38-47(20%)	38-47(20%)	48-57(20%)
Sex			
Mode (% of N)	Female (56%)	Female (57%)	Male (50%)
Marital Status			
Mode (% of N)	Married (57%)	Married (61%)	Married (26%)
Ethnic Group			
Mode (% of N)	White: Scottish (84%)	White: Scottish (91%)	White: Scottish (39%)
SOC2000 (Sub-Major Groups)			
Mode (% of N)	92. Elementary administration and service occupations (12%)	92. Elementary administration and service occupations (12%)	92. Elementary administration and service occupations (12%)

Table 4.10: SHeS: Summary of NIAH Output for Age Recoded into 10yr categories without SOC2000 Sub-Major Groups, *k*=3

This method of coarsening data into categories has reduced the potential for disclosure by reducing the number of outliers in the data set as a whole and therefore limiting an intruder's ability to target specific records with certainty. However, this reduction comes at a cost of excluding more than 70% of records as they still do not meet the k-anonymity requirement unless the SOC variable is suppressed completely or there is a further reduction in the detail of the age variable. Even if SOC is suppressed, the resulting safe dataset over reports the number of married, and white records. This should be considered when assessing the data's resulting utility, which will be covered in section 4.4.1.

Having explored banding and the effects of further categorisation, noise addition will be considered. As set out in section 4.3 noise addition is used in this context to obfuscate variable values and introduce uncertainty for a would be intruder attempting to match *a priori* knowledge with a data set. In this scenario the age variable has had random noise applied to

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=288)	NIAH At Risk Output (N=7728)
Age			
Mean($\pm 1\sigma$)	(32.4) 50.2 (68.0)	(21.7) 40.7 (59.7)	(32.5) 50.1 (67.7)
Sex			
Mode (% of N)	Female (56%)	Female (66%)	Female (55%)
Marital Status			
Mode (% of N)	Married (56%)	Married (58%)	Married (56%)
Ethnic Group			
Mode (% of N)	White: Scottish (85%)	White: Scottish (100%)	White: Scottish (84%)
SOC2000 (Sub-Major Groups)			
Mode (% of N)	92. Elementary administration and service occupations (11%)	92. Elementary administration and service occupations (22%)	92. Elementary administration and service occupations (11%)

Key Variables: Age(10yrs), Sex, Marital Status, Ethnic Group, Health Board.

Table 4.11: SHeS: Summary of NIAH Output for Age with Added Noise, *k*=3

it. Every record has either been increased or decreased by a random integer between 0 and 5. Table 4.11 shows the familiar summary of NIAH output. The first observation to make is the change in N, which has increased to 8016 from 7897. This is because the number of adults has increased. This is because some records younger than 18 have had their ages increased and this group was larger than the number of records over 18 that had their ages reduced to below 18. The overall picture is less than satisfactory with 4% of records deemed safe. However, this is a suitable juncture to acknowledge the limitations of k-anonymity in this respect: k-anonymity cannot capture disclosure risk when variable values are altered beyond their 'true' value. That is to say it cannot capture the uncertainty introduced by the added noise. Also as discussed in section 4.3, a different approach is needed to capture the disclosure risk in these cases. One form, briefly presented here, is an exploration of the false positives an intruder might find in an attempted attack.

In this example an intruder believes they have partial knowledge of a data record. The intruder has data for a 45 year old female postal worker. With access to the raw data the intruder would find a unique match for their *a priori* data. As the SHeS dataset is survey data there is a degree of uncertainty inherent in the dataset. This is because the intruder does not know for sure that the unique match they have found is unique in the population and therefore the record they specifically mean to target. A conservative estimate might suggest that if a sample survey that is a 10% sample of the population has been used then if a unique match is found it is estimated that there are likely to be 10 people in the population with those characteristics. If the intruder were given access to the dataset where the age variable has had noise applied to its values they would find no record that directly matches their knowledge. At this stage an intruder may assume that as there is no match for the record they wish to target, it is not in the survey. However, perhaps they persevere and attempt some fuzzy matching technique to find likely candidates for a match. If they broaden their search to include all records with an age between 40 and 50 inclusive (which actually has the intruder implementing a banding scheme of their own), their search returns 3 results (this includes the unique match from the raw data whose age was altered to 42 from 45). To add further uncertainty, the SOC values are released at the sub-major group level. In this scenario an intruders search query would return 109 records of females aged between 40 and 50 with a SOC sub-major group of 92, which includes postal workers.

Noise addition in the above limited example appears to offer a degree of protection to the records in the dataset. However, as will be discussed in section 4.4.1, the implications of such wide spread randomised noise addition on the utility of the dataset are difficult to predict and are potentially serious for data analysis.

To complete this exploration of applied statistical disclosure methods another form of noise addition is considered. Random misclassification of categorical variable values has a similar effect to that of noise addition for numerical variables. Discussed in the earlier section is the

Post-Randomisation Method (PRAM). A simpler method for illustration purposes is used here. For this example, 10% of the records were selected at random and their geographical variables are randomly misclassified using the whole range of possible values. Given this is a simple example, the algorithm implemented to select the records to be misclassified was a Fisher-Yates shuffle of the record indexes (similar to randomly selecting records out of a hat). Then the first 10% of the shuffled records were misclassified. To miss-classify the record, one of the possible variable values was drawn at random from a uniform distribution and assigned to the record in place of its original value. Table 4.12 presents the summary of NIAH output. As with the noise addition scenario, the overall summary is problematic, k-anonymity cannot capture the effect of changing variable values and 98% of records do not meet k=3. This summary also does not include the health board variable. To verify the position post-misclassification consider figure 4.5. This shows the distribution of Health Board for the original and misclassified data. The distributions are very similar; the largest effect has been on the smaller Health Boards, but the relative distribution is the same. Therefore, the 10% misclassification does not appear to have affected the data significantly. As shown earlier with numerical noise addition, a possible intruder scenario to demonstrate this statistical disclosure control's effects can be explored. Consider an intruder, with *a priori* knowledge that their target is a 33 year old female from the Forth Valley health board area. If the intruder had access to the original data direct matching would return 3 records. If the same direct matching were carried out 4 records would be returned from the misclassified data. So in the first instance the intruder might be able to distinguish between the 3 original records by obtaining more knowledge of their target, and therefore suffer only the uncertainty that their target is not population unique. In the second instance the intruder could attempt to do the same, however their uncertainty is also compounded by the uncertainty that their target was one of the 10% of records misclassified. The act of misclassification in this case has already added 1 further potential match for their *a priori* knowledge. It should also be noted that, as is the case for all SDC methods that directly alter variable values, the intruder is very unlikely to know the SDC scheme that has been implemented. This adds further disclosure protection to the records.

It has been shown that different methods can be applied to microdata to prevent disclosure. In the following section, data utility is discussed and the Scottish Health Survey example is continued with a demonstration of capturing some measure of data utility post-SDC. To provide a more robust dataset for the data utility experiments, ethnicity was also recoded into white and non-white categories, an approach that is often used when reporting administrative statistics, see (Scottish Government, 2011c). This further decreases the size of the at risk partition when a k-anonymity analysis is carried out. To show how this changes the NIAH output partitions consider table 4.13. The results for recoded age and SOC are used as the basis for this further SDC application. Although the recoding of ethnicity does have

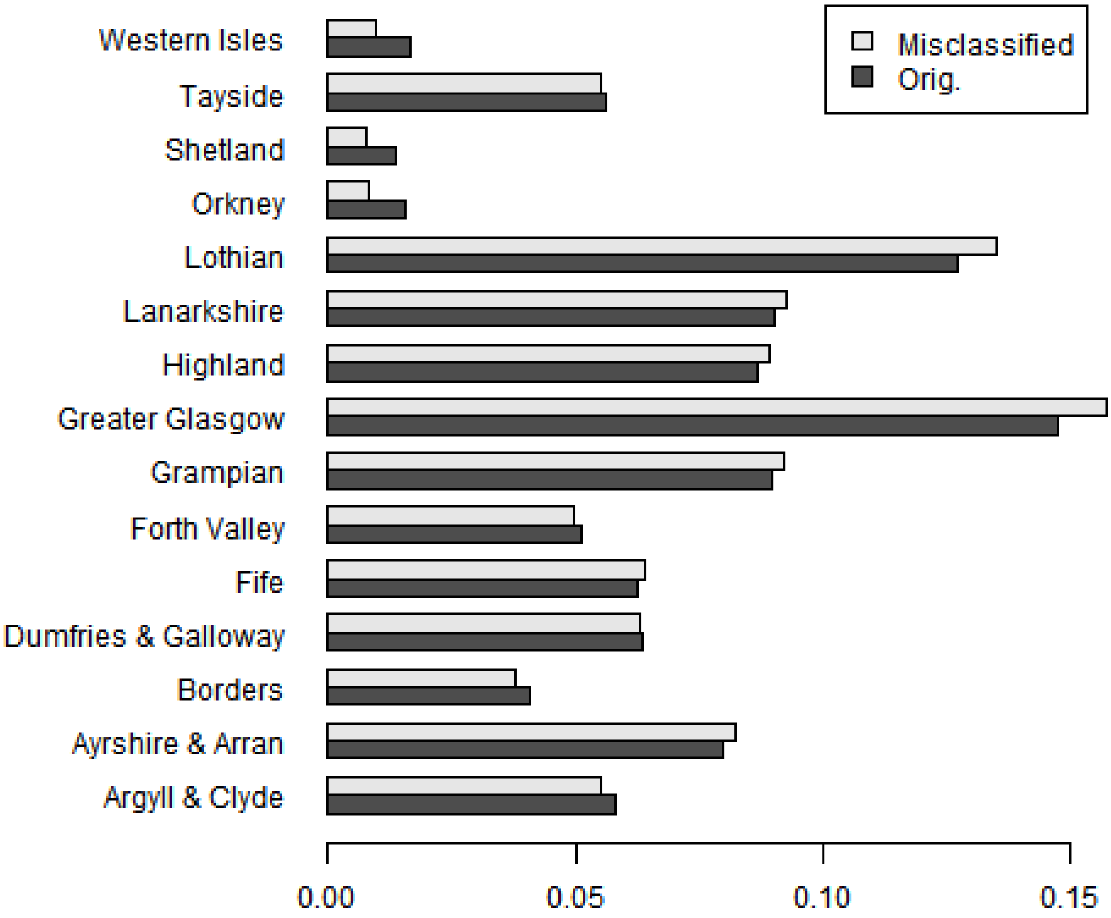


Figure 4.5: SHeS: Distribution of the Health Board Variable for the Original Data and 10% Misclassified Data

a positive effect on the size of the safe partition, the SOC codes still provide a significant barrier to records being deemed safe at k=3. Therefore, the final dataset used in section 4.4.1 will suppress the SOC codes. This can be carried out in this case because the example research question does not require SOC values. Suppressing SOC reduces the at risk partition to 505 records representing 6% of the total data. This remaining 6%were suppressed and the effects of this will be discussed in the analysis of the data's utility.

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=199)	NIAH At Risk Output (N=7698)
Age			
Mean($\pm 1\sigma$)	(33.3) 50.6 (67.9)	(37.4) 52.8 (68.2)	(33.3) 50.6 (67.9)
Sex			
Mode (% of N)	Female (56%)	Female (66%)	Female (56%)
Marital Status			
Mode (% of N)	Married (57%)	Married (77%)	Married (57%)
Ethnic Group			
Mode (% of N)	White: Scottish (84%)	White: Scottish (100%)	White: Scottish (84%)
SOC2000 (Sub-Major Groups)			
Mode (% of N)	92. Elementary administration and service occupations (12%)	92. Elementary administration and service occupations (28%)	92. Elementary administration and service occupations (12%)

Table 4.12: SHeS: Summary of NIAH Output for Health Board with 10% of Records Randomly Misclassified, *k*=3

Variable (Measure)	Original Data (N=7897)	NIAH Safe Output (N=3661)	NIAH At Risk Output (N=4236)
Age			
Mode (% of N)	38-47(20%)	48-57(20%)	38-47(21%)
Sex			
Mode (% of N)	Female (56%)	Female (59%)	Female (53%)
Marital Status			
Mode (% of N)	Married (57%)	Married (77%)	Married (40%)
Ethnic Group			
Mode (% of N)	White (98%)	White (99%)	White (96%)
SOC2000 (Sub-Major Groups)			
Mode (% of N)	92. Elementary administration and service occupations (12%)	92. Elementary administration and service occupations (16%)	92. Elementary administration and service occupations (8%)

Key Variables:Age(10yrs), Sex, Marital Status, Ethnic Group (Binary), SOC Sub-Major Groups, Health Board.

Table 4.13: SHeS: Summary of NIAH Output for Binary Ethnicity with SOC Sub-Major Groups, *k*=3

4.4 Data Utility

Data utility is at the heart of the tension between data controllers and researchers. The potential for disclosure is a serious problem for data controllers, not least due to the aggressive data protection framework within which they have to work. In the health field, a good example of this are the Caldicott principles used in the NHS (Department of Health, 1997) or the European Union's Data Protection Directive (discussed in Chapter 2). As with the measurement of risk, a number of different approaches to data utility have emerged. It is useful to set out some definitions for data quality. Put simply, for researchers, data quality is defined as the comparison of the released data to the original data in terms of its operationalisation when conducting statistical analysis. By extension, statistical outputs reviewed by data controllers should provide the same results after review.

Approaches to data utility can be divided into two broad types; replication of 'real world' analyses (Purdam and Elliot, 2007; McCaa et al., 2013; Brickell and Shmatikov, 2008), or targeted utility measures independent from analyses (Rastogi et al., 2007; Li and Li, 2009; Askari et al., 2012). Both have their disadvantages; data owners that use research questions to test the utility of their data cannot guarantee that it will perform as well in all possible analytical iterations that potential users might select. While on the other hand, those that use techniques specifically designed to assess data utility struggle to articulate how data might perform for a given user. Brickell and Shmatikov (2008) provides a succinct summary of the difficulty inherent in measuring utility: "[The U]tility of any dataset, whether sanitized or not, is innately tied to the computations that one may perform on it. For example, a census dataset may support an extremely accurate classification of income based on education, but not enable clustering based on household size. Without a work-load context, it is meaningless to say whether a dataset is 'useful' or 'not useful,' let alone to quantify its utility" (Brickell and Shmatikov, 2008, 5).

This juxtaposition, has led to a call for verification services to be provided by data owners, post-statistical disclosure control and analysis. In this vein, Reiter et al. (2009) recommends data owners to provide verification servers that could carry out a users analysis on the original data and the data to which SDC methods have been applied. This server could then report a measure of 'fidelity' which users could include when making claims about statistical inferences. A measure of 'fidelity' also assuages a demand by users to know what the data owners have done to the original data; data owner's do not publish specific details regarding their statistical disclosure control regimes for fear this would aid an attack on the data.

Away from specific techniques, Zaslavsky and Horton (1998) offers a conceptual framework which sets out the problem of whether to disclose data as a decision problem. They discuss the application of decision analysis, and present the case for calculating the impact of decisions to disclose data in terms of loss of information that could have been publicly useful.

Using this framework they plot data utility against potential disclosure risk to provide a sensitivity analysis for different non-disclosure decisions. However, as acknowledged in their conclusions, this approach does not provide an implementation that data owner's can readily use.

As has been shown above, different approaches to data utility exist in the literature. Given the pragmatic research design outlined in Chapter 3, and the conceptual ideas of trust between data owners and users, discussed in Chapters 2, 5 and 6, the approach pursued in this research takes the real world, case study, model as a template. This has the advantage of presenting data utility in a language familiar to data users, which makes it easier for data owners to articulate. In the next section, assessing data utility is demonstrated using the Scottish Health Survey and the SDC work from the preceding sections.

4.4.1 Estimating Data Utility: An analysis using the Scottish Health Survey

The literature and associated methods for assessing data utility are captured in the preceding section 4.4 In this section the Scottish Health Survey (SHeS) is used to demonstrate the application of some of the methods discussed. In section 4.3.1 a 'safe' dataset was constructed by applying a series of statistical disclosure control measures (SDC) to the SHeS dataset. This safe data is used for the analysis throughout this section. Also the two supplementary datasets constructed using forms of noise addition will be considered. First, some descriptive statistics are provided to show an overview of what the dataset looks like post-SDC. From this position of re-familiarisation, a demonstrative research question is examined to ascertain the effects of the SDC methods employed when the 'safe' data are compared with the original SHeS data.

To provide a simple overview of the aggregate differences in the original and disclosure controlled data, demographic variables are compared. Figure 4.6 shows the comparison of ages in both datasets. Here, the detailed age variable in the original has been recoded to fit the age categories in the SDC dataset. This allows for greater comparability and the extent of data utility lost here will be picked up further in the analysis. As is shown, the overall distribution of age has been unaffected by the suppression of the 6% of records deemed at risk.

In addition to the changes to the data set out above, the ethnicity variable has again been recoded to a common scheme so that a direct comparison can be made (this binary representation is referred to in section 4.3.1). This shows that the homogeneity increases further from 98% white to 99.5% white for the safe data. The loss in detail will be examined further in due course. However, from the research question and regression modelling, it will be difficult to

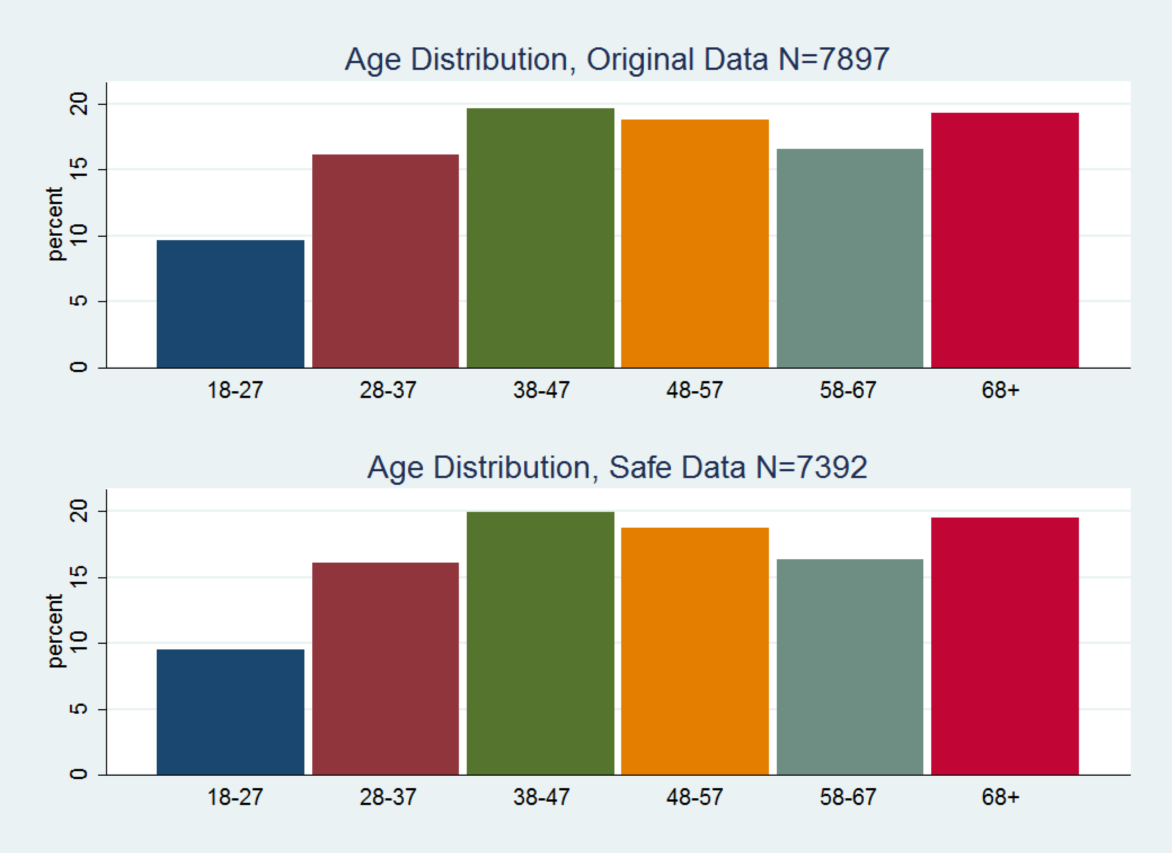


Figure 4.6: SHeS: Age Distribution for the ‘safe’ and Original Data

show the impact of this data loss as even in the original data the total number of non-white ethnicities is potentially too small to yield significant results. What can be highlighted at this stage is that the safe data potentially renders any research with a focus on ethnicity particularly difficult and perhaps not viable at all. This effect also has wider ranging implications outside the ethnicity example provided here. Although the focus is more likely to be on the effect of SDC on summary statistics, the elimination of rare cases in datasets could be critical across any combination of variables relevant to a given research question.

Figure 4.7 compares the distribution of marital status across both datasets, revealing no significant changes between original and safe data. Also a comparison of gender indicates a small increase of 1% in the number of women in the safe dataset when compared with the original data.

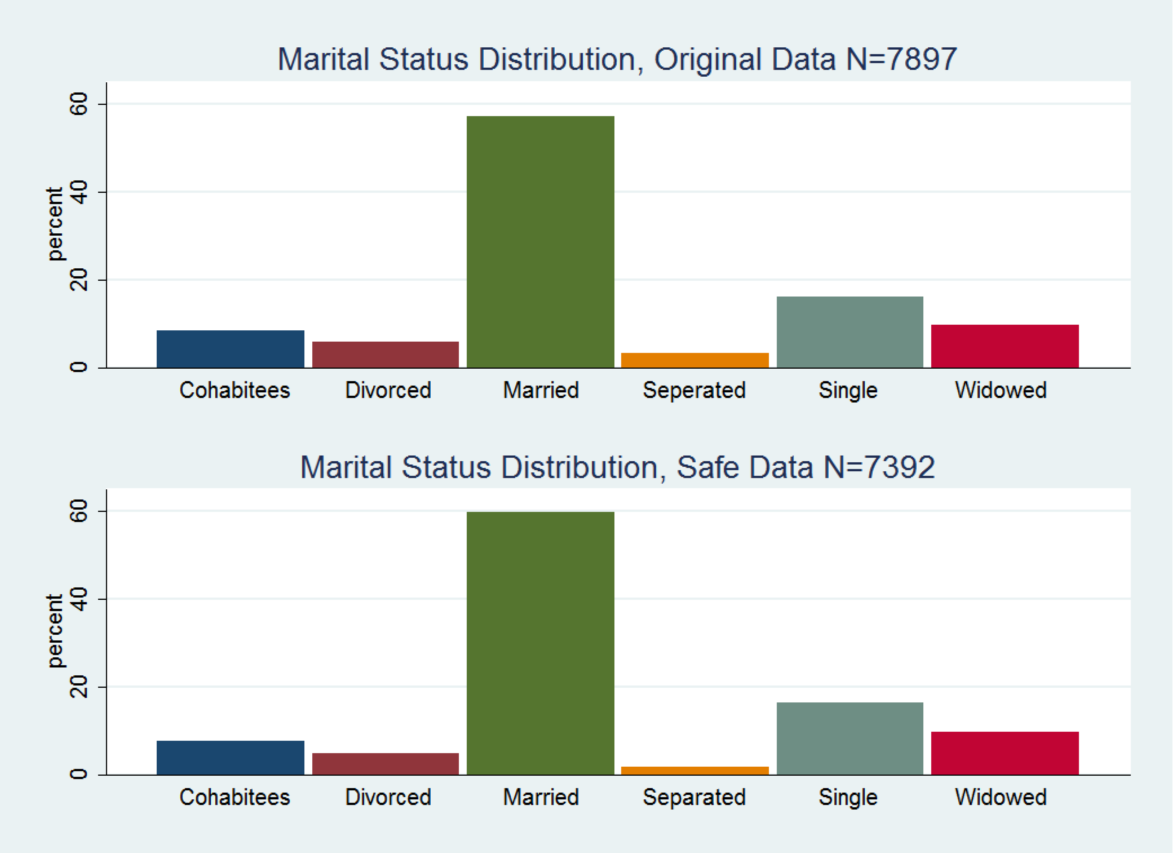


Figure 4.7: SHeS: Marital Status Distribution for the ‘safe’ and Original Data

The results above are to be expected given that the data values themselves were not changed but aggregated to some more coarse level of detail. In addition, the low level of records suppressed (6%) to enforce the k=3 k-anonymity requirement has made little difference to the overall distributions of demographic variables. The additional two datasets constructed in section 4.3.1 that carried some form of noise or misclassification were also analysed to ascertain what effect the SDC has had on high level distributions. Figures 4.8 and 4.9 provide distribution comparisons for the two additional disclosure controlled data sets. Figure 4.8 shows the distribution for age for the original data and the data that has had noise applied to the age variable.

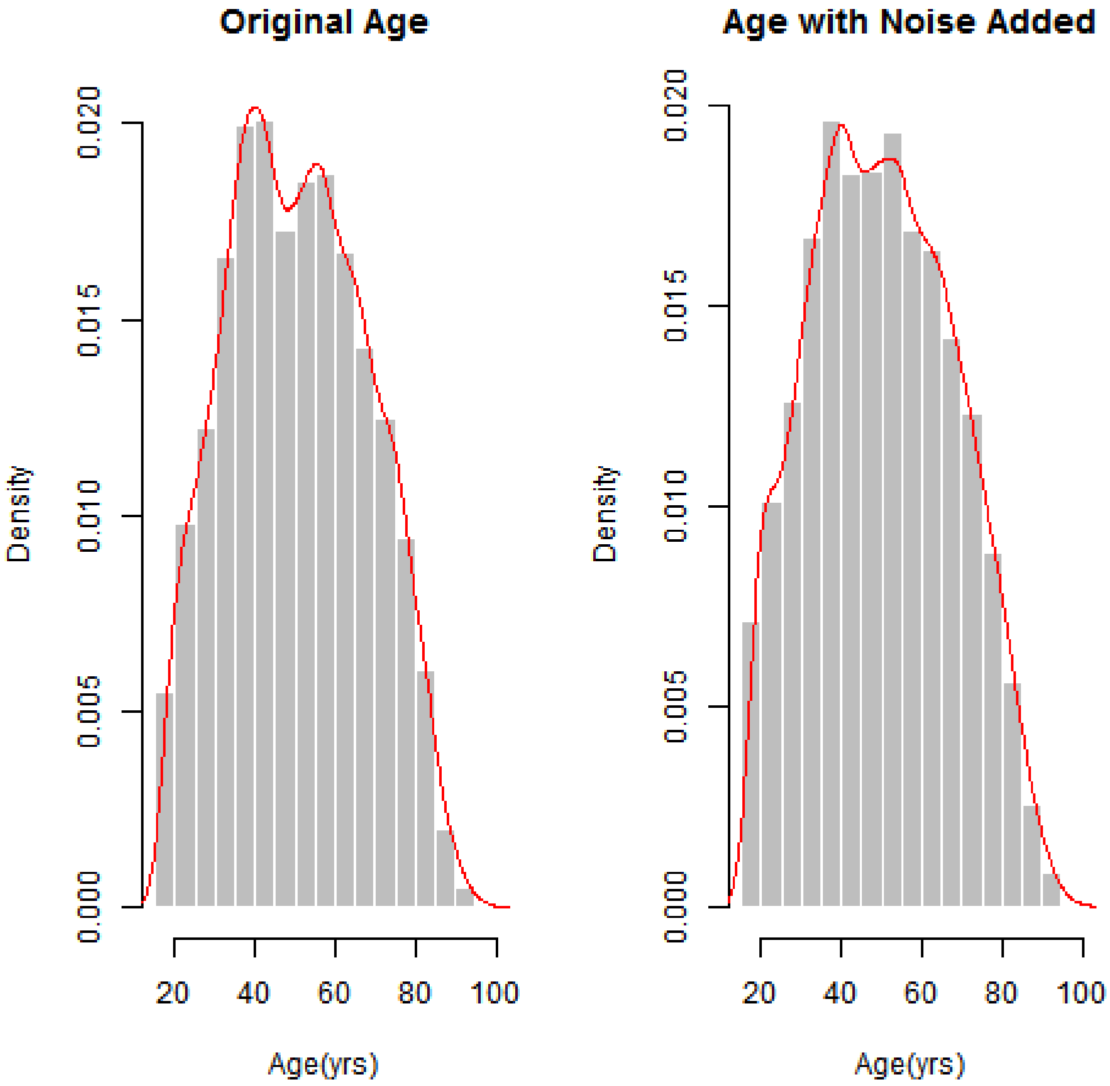


Figure 4.8: SHeS: Distribution Plots for the Original Age and Age with Noise Added Variables

Figure 4.9 presents a picture similar to that of the data with noise added. The distributions of the health board variable are broadly the same for both datasets. The geographical location of 10% of the records has been randomly misclassified. From this data, it is observed that the less prominent health boards have been inflated by $\tilde{1}$% but the overall distribution remains relatively static. Given these results, a degree of confidence could be expressed when producing univariate statistics. Therefore, an analyst could generate univariate statistics from the data that has had disclosure control applied with confidence that the statistics were accurate for the underlying original data. However, the data that has had either numerical noise or categorical misclassification applied has directly altered the total counts for the respective variables. In terms of the age variable with random noise applied, the number of adults in the dataset has increased from 7897 to 8016, an increase of 1.5%. This is perhaps an insignificant difference however, a number of statistical publications, especially those released by governmental organisations, rely heavily on the accurate use of aggregate tables of counts and therefore the disclosure control used here could be prohibitive.

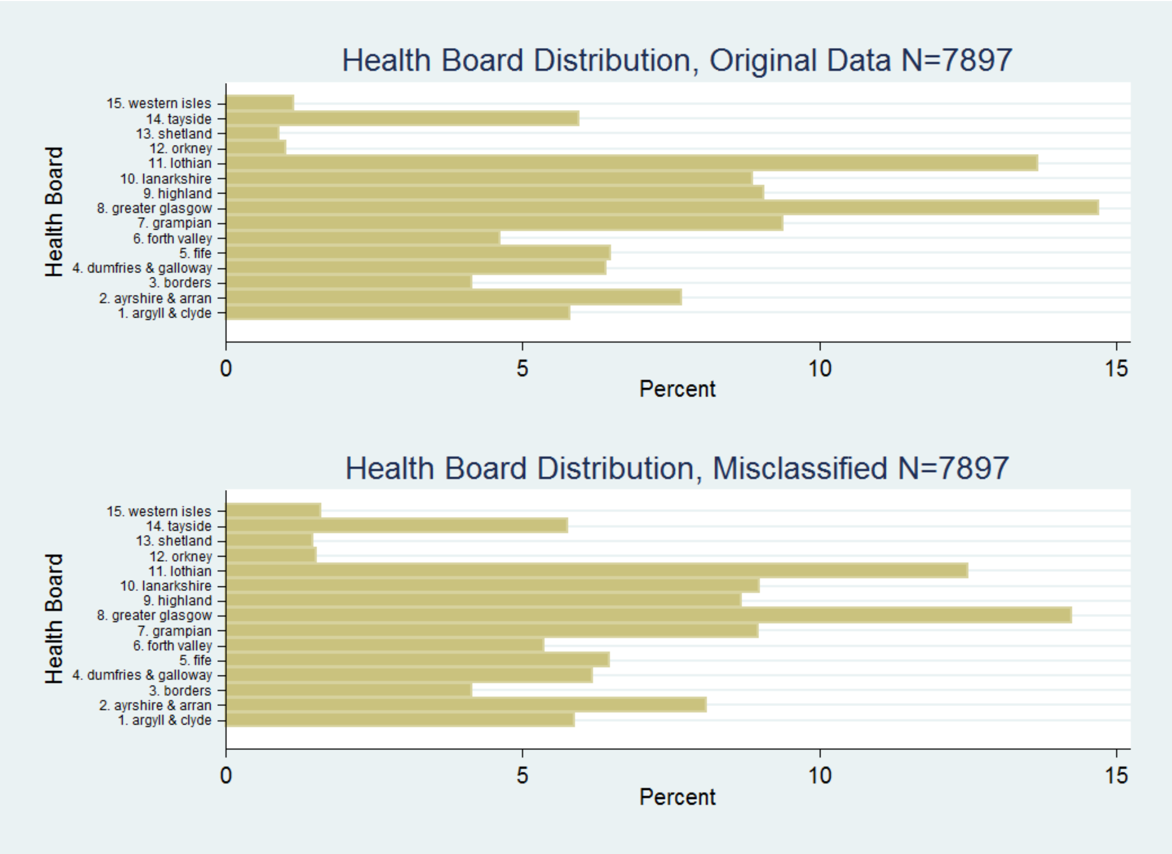


Figure 4.9: SHeS: Distribution of Health Board for the Original Data and 10% Misclassified Data

Having considered the validity of univariate statistics, it would useful to scale up the level of complexity and analyse the effects of the disclosure control measures applied in section 4.3.1 to some bivariate statistical analyses. The relationship between age and marital status was explored using the data with disclosure control applied and the original raw data for age in years. An ANOVA test for age and marital status was conducted and no significant change in the correlation was found (original age(F=758.88***) and age with noise added(F=777.15***)). From the earlier comparison of distributions, the overall distribution of age compared with noise added age seemed to indicate no significant shift and from that it could be assumed that maintaining a single year of age measure could yield greater data utility than the 10 year age category recoded age variable, and these results support this assumption.

Scaling up the level of complexity further, the following example research question is now considered: what is the effect of social class on healthy eating habits? In order to examine this, an indicator based on the Government's guidance for individuals to consume at least five portions of fruit and vegetables a day has been constructed from variables regarding diet in the health survey. Having given attention to a number of the variables that will be included in the analysis, a description of the indicator for the 5-a-day guidance is also included. Figure 4.10 illustrates the prevalence of this indicator for the original data. This shows that only just under a quarter of the sampled adult population actually meet the Government's guidance. There are numerous studies that have included some analysis of the '5-a-day' government message, for example see Ashfield-Watt et al. (2004), Naska et al. (2000) and Pomerleau et al. (2005). Social class is represented in the health survey using the National Statistics Socio-Economic Classification (NS-SEC). For this analysis, the 8 category NS-SEC has been used. Figure 4.11 shows the distribution of the NS-SEC categories across the surveyed population.

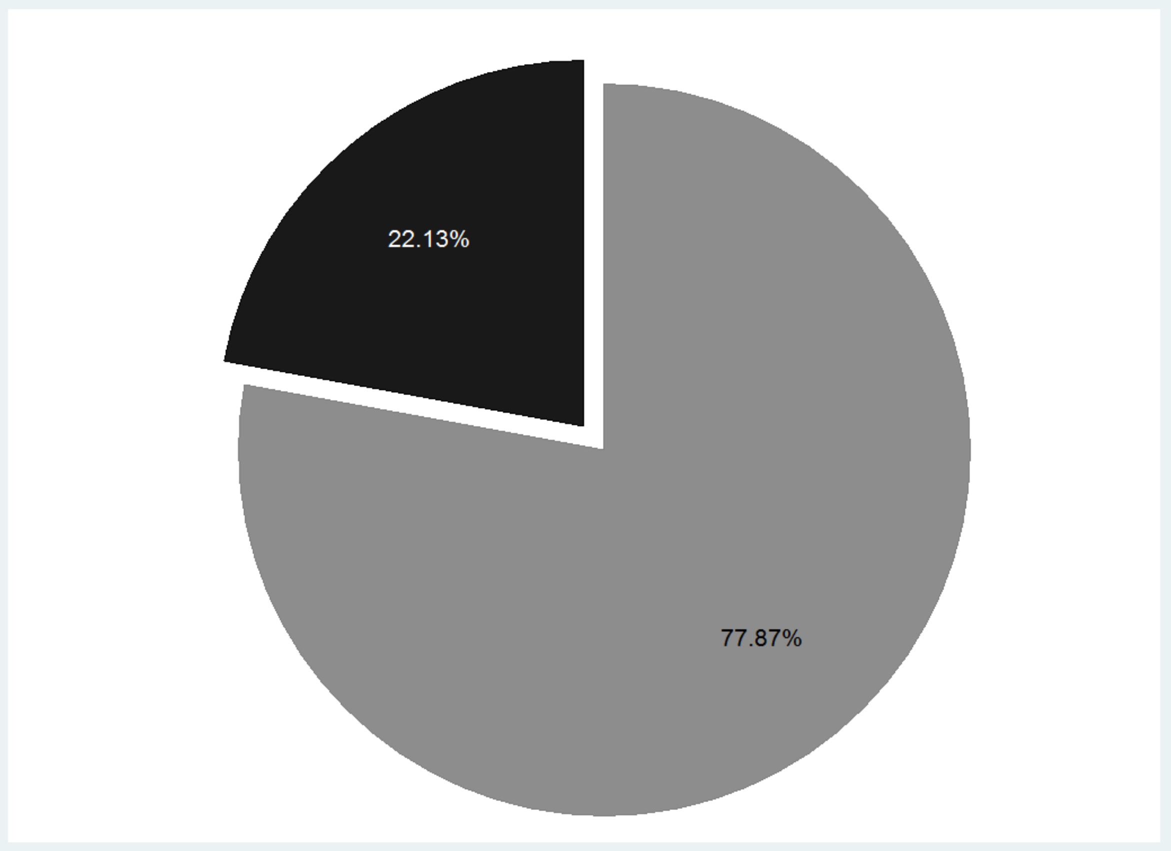


Figure 4.10: SHeS: Pie Chart of the '5-a-day' Fruit & Veg Indicator using the Original Data (N=7897)

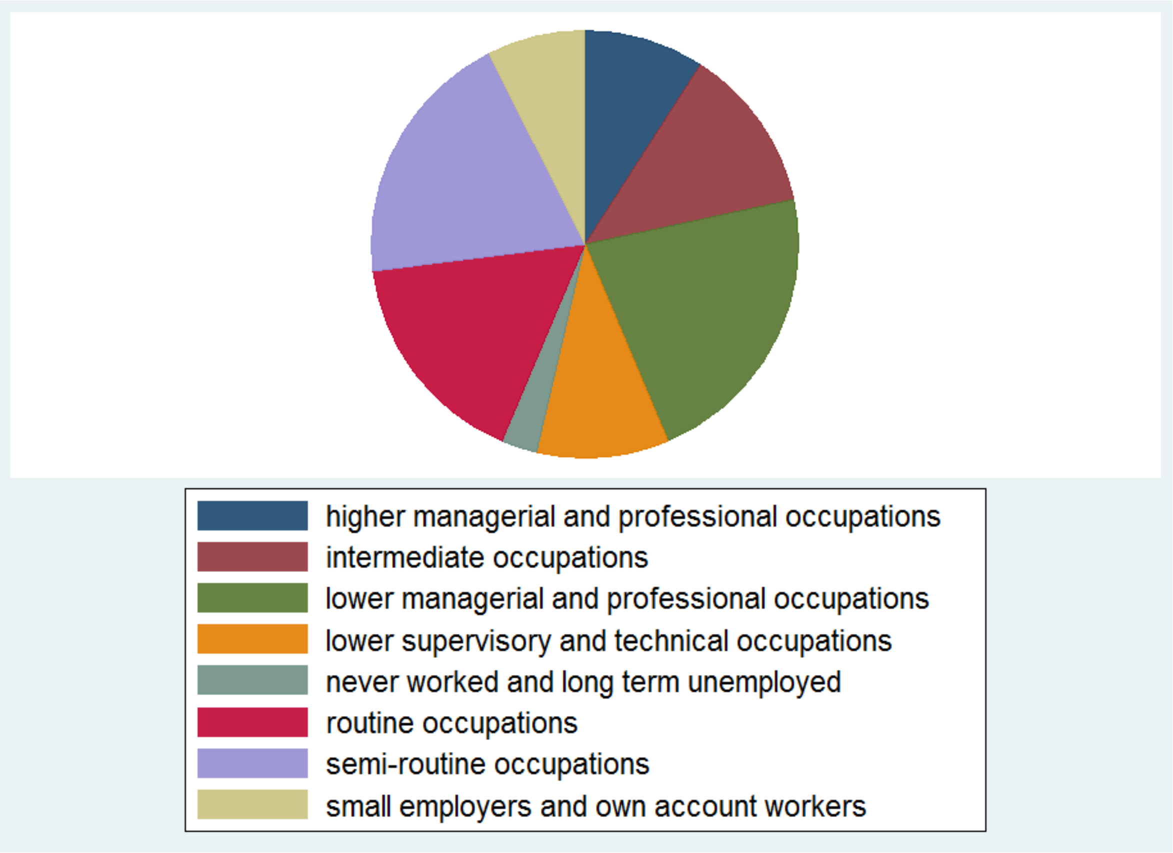


Figure 4.11: SHeS: Distribution of NS-SEC using the Original Data (N=7897)

To determine the effect of social class on whether a respondent is likely to meet the Government's guidance, a series of binomial regression models have been constructed. Table 4.14 shows the results of these models for the original data and the safe dataset from section 4.3.1. Model 1 and 1a present a base-line model using age, gender and marital status. Marital status has been simplified into a binary indicator of married or not-married. This was simplified to capture some effect from marital status without constructing dummy indicators for other marital status values, some of which have low counts. Also, the obvious difference between these two models is the way age of respondents is presented. Our original data has single year of age and the safe data has age grouped into categories that span 10yrs. These base line models explain very little of the variation in the data with R^2's of 0.003 and 0.008. Being female has a significant positive effect on whether a respondent consumes 5 portions of fruit and vegetables a day. Similarly, being married also has a significant positive effect. Should further analysis of this data be explored, there is potential for an interaction effect to test for being female and married. Studies in this area have also suggested that marital status has a pronounced affect on men's diets, for example see Eng et al. (2005). This could also be explored further looking at an interaction effect of being male and married. Age presents potentially counter-intuitive results suggesting that the safe data does a better job then the original, however these results can be unpicked. Continuous single year of age does not have a significant effect on the 5-a-day indicator. A quadratic relationship was also tested by using age squared values, but this also did not provide a significant result. Whereas the grouping of age into 10yr bands does produce significant results for the 48-57 and 58-67 age brackets. Both of these groups show a significant positive effect when tested against the reference category 18-27. What in essence has been shown in these two models is that there is not a linear relationship between age and the consumption of 5 portions of fruit and vegetables. Model 1a appears to give a better analytical narrative, however using the raw original data a number of different categorisations of age could be constructed to pursue the relationship between the two variables. If the safe data are used the only option is to reduce the level of detail still further.

Models 2 and 2a introduce a set of indicator variables to explore the geographical effects on 5-a-day consumption. The reference category here is Greater Glasgow, therefore all results should be interpreted as being in a given health board rather than being in the Greater Glasgow health board. With the introduction of geography, there is a modest increase in the R^2 values. These indicate that $\tilde{1}$% of the variation in the datasets are explained by the models. For both models, gender and being married as still positive significant influences. Some of the Health Boards show statistically significant results. Respondents from Fife and Highland are more likely to consume 5-a-day than those from Greater Glasgow. Those respondents from Lanarkshire and the Western Isles are less likely to consume 5-a-day than those from Greater Glasgow. Single year of age continues to be not significant. The introduction of geography has also reduced the significance of being in the 58-67 age category in model 2a.

Models 3 and 3a extend the analysis further to incorporate the NS-SEC. The reference category for the NS-SEC indicator variables is NS-SEC 1: Higher managerial and professional occupations. The demographic variables follow the same narrative as the previous two groups of models. Being female and being married provide a significant positive effect on 5-a-day consumption. The introduction of NS-SEC has diminished the significance level of being married from the 99% to the 95% level. Single year of age continues to be not significant and the two older age groups of 48-57 and 58-67 are still presenting a significant positive effect. The geographical effects follow closely those of model 2 and 2a. The NS-SEC classifications all produce significant effects on the consumption of 5-a-day when compared with the reference category. There is an identifiable trend across the hierarchy of the NS-SEC classification. The further away from 1: Higher managerial and professional occupations a respondent is the higher the negative effect on whether they will consume 5-a-day. The only exception to this linear trend is an overlap between intermediate occupations and small employers and own account workers.

From these three pairs of models, it has been shown that there are subtle variations in how each model performs when compared across the original dataset and safe data. As has been highlighted, the serious consequence for users using the safe data is lack of extra-exploratory depth beyond models similar to those above. The lack of detail in the age variable makes unpicking the relationship between age and 5-a-day consumption difficult. Also having suppressed the SOC occupation codes from the dataset (as was done by the Scottish Government in the years after 2003) it is not possible to explore the effects of different types of occupations on the 5-a-day indicator. It is necessary to make the caveat that NS-SEC is derived in part from SOC so using NS-SEC as an occupation based measure will to some degree explain the relationship between occupation and 5-a-day consumption but this is not easy to show coherently. For comparison, Table 4.15 shows a model that replaces NS-SEC with the major SOC groups. SOC major group 1: Managers and Senior Officials is excluded from the model as a reference category. This model shows that the major groups have a significant effect on 5-a-day consumption. The only exception is the associate professional & technical group which is arguably sufficiently close to the the reference category so as not to provide a significant difference. The R^2 for this model is marginally higher than Model 3, it is not possible to draw out a direct difference between the effectiveness of these two models. However, it supports the loss of data utility narrative by demonstrating further that the safe data is hampered by a lack of analytical opportunities.

To round out this exploration of data utility, the two additional datasets constructed in section 4.3.1 are considered in comparison with the original data. Model 5 and 5a compare the original data with the data where the age values have had a random amount of noise added (+/- a random integer between 0-5). Although continuous single year age has not been significant throughout the proceeding analyses, the difference here is that the co-efficient for age is even closer to zero. Earlier it was shown that adding noise of this type to age did not significantly affect the correlation of age and marital status and it has had no significant effect here either. Models 6 and 6a provide a similar test of the data where noise has been applied; here the random misclassification of the health board variable is examined. In this case the scope of the noise is reduced as only 10% of records were misclassified. However, the results still show significant disruption to the statistical relationships contained in the original data. Only one consistent result remains across the two models (Highland still has a significant positive effect). Also the R^2 suggests that a model with these variables explains less of the variation than is actually the case. Should geography be a primary concern to a user's research question, using the misclassified data would likely lead to false conclusions. One possible explanation for the significant changes here is linked to the change in distribution seen in 4.9. As Greater Glasgow is our reference category and it's share of the distribution has increased relative to the other health boards, it is possible that strong correlations between Greater Glasgow and other health boards have been muddied by the misclassified records assigned to Greater Glasgow, and those removed and assigned elsewhere.

What has been shown in this section is that data controllers must proceed carefully when considering the statistical disclosure control that they wish to apply. Although reducing the level of detail might appear simplistic, in the research example shown above it performs better than noise addition and misclassification. Noise addition and misclassification in this case severely disrupt the data's inherent statistical relationships even if the general variable distributions are preserved. The potential for erroneous conclusions has severe consequences for the data user experience, and the level of trust between users and data controllers. As will be discussed elsewhere (Chapter 4) the ability for users to obtain some results, if limited in scope, that are correct when compared with the original data is an important consideration. In this vein, ultimately, users can build on early results and possibly obtain greater data access to areas of interest to their research.

Table 4.14: SHeS: Regression Models for the Raw Data (1,2,3) and Safe Data (1a,2a,3a), that explore the factors that influence the ‘5-a-day’ Fruit and Vegetables Indicator

Variable	Model 1	Model 1a	Model 2	Model 2a	Model 3	Model 3a
Constant	-1.397	-1.600	-1.412	-1.615	-0.768	-0.855
(Standard Error)	(0.093)	(0.107)	(0.112)	(0.122)	(0.113)	(0.146)
Age(yrs)	-0.002		-0.002		-8.99e-06	
	(0.002)		(0.002)		(0.000)	
Female	0.181***	0.180***	0.188***	0.188***	0.285***	0.283***
	(0.055)	(0.058)	(0.055)	(0.058)	(0.060)	(0.063)
Married	0.233***	0.172***	0.233***	0.180***	0.133**	0.094
	(0.056)	(0.063)	(0.057)	(0.064)	(0.057)	(0.066)
Age (48-57)		0.392***		0.365***		0.315**
		(0.122)		(0.123)		(0.128)
Age (58-67)		0.260**		0.241*		0.287**
		(0.127)		(0.127)		(0.132)
Health Boards						
Fife			0.262**	0.253**	0.262**	0.250*
			(0.122)	(0.127)	(0.124)	(0.129)
Highland			0.411***	0.410***	0.419***	0.419***
			(0.108)	(0.111)	(0.110)	(0.113)
Lanarkshire			-0.433***	-0.404***	-0.450***	-0.423**
			(0.125)	(0.126)	(0.126)	(0.128)
Shetland			0.539*	0.137	0.552*	0.188
			(0.288)	(0.443)	(0.292)	(0.450)
Tayside			-0.183	-0.253*	-0.205	-0.280*
			(0.139)	(0.148)	(0.141)	(0.150)
Western Isles			-0.947**	-0.852**	-0.979**	-0.854**
			(0.379)	(0.408)	(0.382)	(0.413)
NS-SEC Categories						
Intermediate Occupations					-0.718***	-0.708***
					(0.383)	(0.117)
Lower Managerial & Professional					-0.372***	-0.392***
					(0.095)	(0.100)

Lower Supervisory & Technical					-0.818***	-0.828***
					(0.119)	(0.125)
Never Worked					-1.449***	-1.408***
					(0.227)	(0.242)
Routine Occupations					-1.366***	-1.39***
					(0.115)	(0.121)
Semi-Routine Occupations					-1.009***	-1.056***
					(0.105)	(0.111)
Small Employers & Own Account Workers					-0.579***	-0.696***
					(0.125)	(0.134)
Pseudo R^2	0.003	0.008	0.012	0.016	0.038	0.042
No. Obs.	7897	7392	7897	7392	7897	7392

*Only statistically significant values are reported above. This model included controls for: all age categories above 18-27 and all healthboards except Greater Glasgow

Variable	Model 4
Constant	-1.103
(Standard Error)	(0.139)
SOC Major Groups	
Administrative & Secretarial	-0.415***
	(0.114)
Elementary Occupations	-0.844***
	(0.113)
Personal Service Occupations	-0.253**
	(0.125)
Process, Plant & Machine Operatives	-0.976***
	(0.135)
Professional Occupations	0.431***
	(0.107)
Sales & Customer Service	-0.608***
	(0.136)
Skilled Trades	-0.502***
	(0.119)
Pseudo R^2	0.039
No. Obs.	7624

Only statistically significant values are reported above. This model included controls for: age, gender, marital status, health board, and all SOC Major Groups except Managers and Senior Officials

Table 4.15: SHeS: Regression Model for the Raw Data, exploring the effect of SOC2000 Major Groups on the ‘5-a-Day’ Fruit and Vegetables Indicator

Table 4.16: SHeS: Regression Models for Raw Data (5,6) and Data with Noise Added (5a,6a), that explore the factors that influence the ‘5-a-day’ Fruit and Vegetables Indicator

Variable	Model 5	Model 5a	Model 6	Model 6a
Constant	-1.397	-1.548	-1.412	-1.473
(Standard Error)	(0.938)	(0.107)	(0.112)	(0.117)
Age(yrs)	-0.002	-0.000	-0.002	-0.002
	(0.002)	(0.002)	(0.000)	(0.002)
Female	0.181***	0.175***	0.188***	0.182***
	(0.055)	(0.055)	(0.055)	(0.055)
Married	0.233***	0.300***	0.233***	0.223***
	(0.056)	(0.063)	(0.057)	(0.056)
Health Boards				
Ayrshire & Arran			-0.155	0.283**
			(0.122)	(0.126)
Dumfries & Galloway			-0.007	0.209*
			(0.129)	(0.151)
Fife			0.262**	-0.055
			(0.122)	(0.132)
Forth Valley			-0.024	-0.585***
			(0.139)	(0.174)
Highland			0.411***	0.329***
			(0.108)	(0.111)
Lanarkshire			-0.433***	-0.052
			(0.125)	(0.126)
Shetland			0.539*	0.738***
			(0.288)	(0.260)
Tayside			-0.183	0.293**
			(0.139)	(0.128)
Western Isles			-0.947**	-0.123
			(0.379)	(0.278)
Pseudo R^2	0.003	0.004	0.012	0.0093
No. Obs.	7897	8016	7897	7897

*Only statistically significant values are reported above. This model included controls for: all healthboards except Greater Glasgow

Chapter 5

Barriers to Efficient Analysis: The case for researcher experience

In the preceding chapters, the attention has been focused on the theoretical, legal, and ethical underpinnings of statistical disclosure control, as well as the existing literature on risk analysis, disclosure methods and data utility. The focus now shifts to expand the existing scope of the literature by considering the inclusion of the data user's experience as a factor in statistical disclosure control workflows. This means supplementing the risk and utility assessments seen in Chapter 4 with the addition of some assessment of the experience of data users. This experience is considered in respect of user's access to data, including any pre-conditions, their analysis of the data and the generation and treatment of outputs. This Chapter's argument is developed across three sections; models for data access, the generation of outputs and virtual research environments (VREs) before concluding with a potential model for increasing the use of VREs in the health sphere and also some thoughts on how the user experience could be assessed.

5.1 Models for Data Access

Before beginning the discussion of data access, it is worth reinforcing a point made in Chapter 2. Data access in the majority of cases is granted by the organisation that collects or owns the dataset[1], however this is itself premised on the agreement, whether tacit or explicit, of the data subjects themselves.

[1] An exception to this in the health field might be data controlled by the Privacy Advisory Committee (PAC) although part of the wider NHS Scotland PAC is an arms-length body from Information Services Division (ISD) that actually hold the data. Also, in the more general case, the UK Data Archive might exercise the role of data owner on behalf of others.

As such, increasing the awareness of how data are used and processed has an effect on data access considerations, especially when building trust between researchers and participants (see discussions of information sharing and transparency in Thomas and Walport (2008); UK Clinical Research Collaboration (2007); Council for Science and Technology (2005)).

To illustrate this point, some qualitative work was carried out as part of an Economic and Social Research Council (ESRC) Festival of Social Science 2012 event[2]. This event was organised in conjunction with the National Records of Scotland, the Scottish Government and The University of Edinburgh. It sought to draw together an audience that cut across the stakeholders in the data linkage and data privacy area. As such, attendees included members of the public, local authority data controllers, government analysts, researchers and public pressure groups (e.g. Stonewall Scotland and Liberty). During the event the audience were asked a number of questions, and were provided with electronic voting devices to record their answers. These devices were assigned at random and no links between an individual and their answers were recorded. This was done to ensure a degree of anonymity for participants. In addition not all members of the audience chose to participate during every question. For this reason these results are only offered as a backdrop and a potential contribution to scoping out areas for future work.

The question most relevant to the current point on increasing the flow of information about how data are recorded and processed was question 2: “How do feel about your personal data being used for research, especially when it’s linked with other data?” This question was asked at the beginning of the event and then repeated at the end. Figure 5.1 shows the results before and after. It should be noted that the N for each is different as not all participants answered this question in both instances. However, the general trend toward the left of the graph (the more positive sentiment) does resonate with the more rigorous findings of similar studies, such as Aitken et al. (2011); Aitken (2012); Scottish Government (2011b). It is the position of this research that this approach of greater sharing of information could also be adopted in the relationship between data owner and data user and this theme will run throughout this chapter.

Data Access

The problem of giving researchers access to potentially sensitive data for analysis is not new; below are some examples of data access models from this century and a description of their constituent parts. At an abstract level these methods differ in only one respect - should the data be brought to the researcher or should the researcher go to the data. Traditional approaches to data access have focused on providing researchers with data directly through some physical medium (e.g. CDs or disks). Projects including the Cross National Equivalence File, which provides contains equivalently defined variables from 1970 through 2009

[2] As noted here: http://www.esrc.ac.uk/news-and-events/events/festival/events-archive/2012-specific/perceptions-privacy.aspx

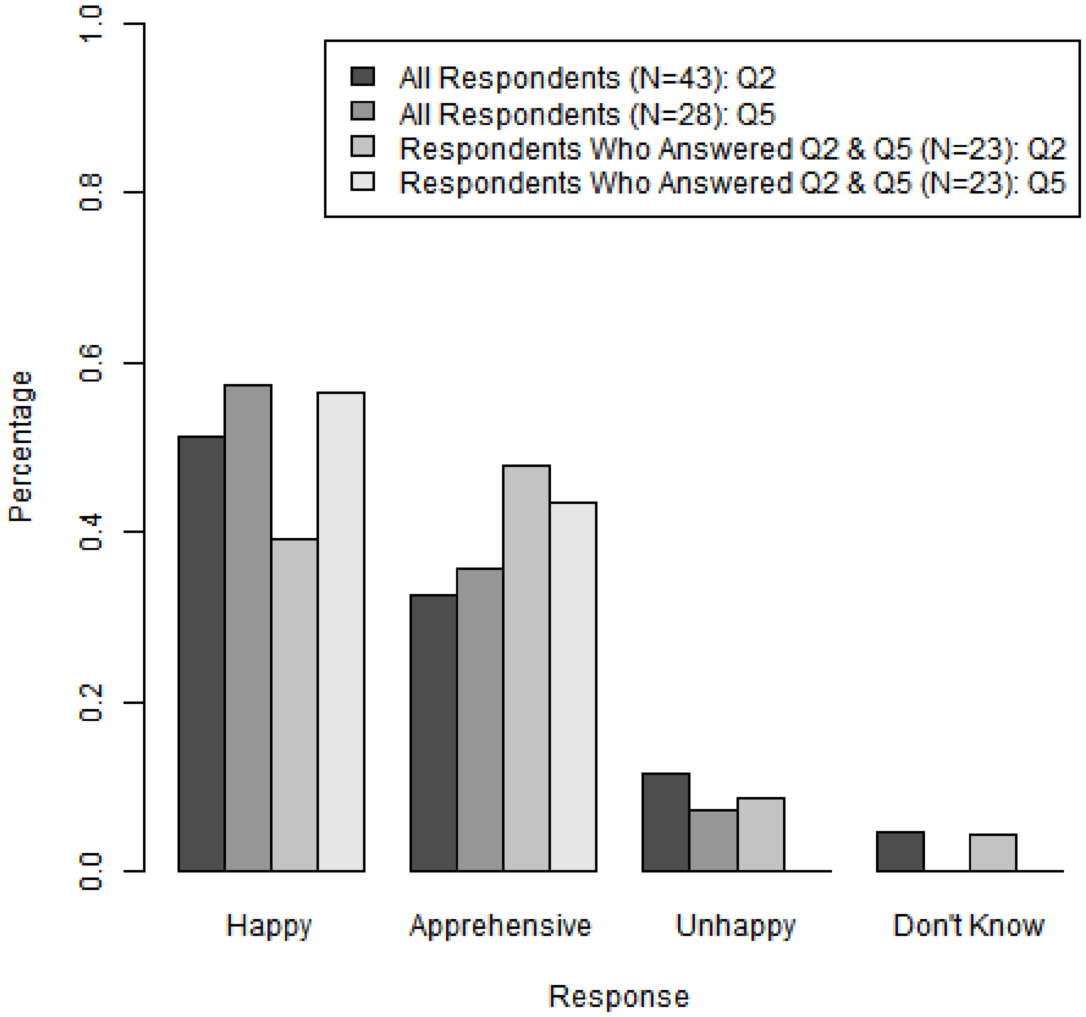


Figure 5.1: ESRC Festival Event Q2 & Q5: How do feel about your personal data being used for research, especially when it's linked with other data?

for the British Household Panel Study (BHPS) and similar studies from Australia, Korea, the USA, Russia, Switzerland, Canada and Germany, still use CDs as their primary method for data distribution (Ohio State University). Some projects also still offer this as a legacy option alongside a downloadable dataset, for example see the Surveillance, Epidemiology and End Results Program (SEER) (National Cancer Institute, 2014). These types of data access models require an agreement between the researcher and the data owner to be signed[3]. These agreements have not changed significantly since the move from physical media to downloadable data. They place a number of conditions on the researcher's use of the data and provide the potential for penalties if the agreement is broken. This model is also common across data centres where data has not been marked as in need of special restrictions (see UK Data Archive for the the UK Data Archives standard End User Licence). The penalties in these agreements can be severe and serve as a deterrent not only for the user themselves but also in some cases their institution. The UK Data Archive's (UKDA) standard End User Licence includes permanent or temporary suspension of access to the UKDA services and the threat of possible legal action.

When data are deemed too sensitive to offer for direct download or distribution on physical media, the contemporary approach is to require researchers to go to the data. Traditionally this has been in the form of physical infrastructure such as bricks-and-mortar secure settings. For example, the UK Data Service Secure Lab has a 'Safe Centre' at the University of Essex; The Scottish Longitudinal Study has a 'safe-setting' at the National Records of Scotland's Ladywell House; and within the health context, NHS Scotland Information Services Division has a 'Safe Haven' in Edinburgh. This approach is also not limited to the UK, the Minnesota Population Center[4] operates a 'Secure Data Enclave' in Minneapolis.

Some projects have attempted to allow researchers to 'virtually go to the data' by providing remote access to secure settings. These remote systems have either been based on access via virtual private networking (VPN) from the researcher's own computer or by creating space (a room) for a dedicated secure-setting within institutions. In 2013 the Administrative Data Taskforce carried out a survey of existing secure-settings and found of the 71 Universities that responded, 26 had some form of secure-setting for accessing sensitive data (Administrative Data Taskforce (Technical Group), 2013, 27). From this it is safe to assume that at present for a number of researchers, accessing sensitive data would mean going to a location outside their own institution. The VPN approach, from a user's own machine, is used by the Health Informatics Centre (HIC) at the University of Dundee; figure 5.2 shows the system used by HIC for remote access. The Centre's software allows for a secure connection where no data can be moved from the secure-setting to the local machine (Health Informatics Centre (Dundee)).

[3] SEER's sample agreement can be seen at: http://seer.cancer.gov/data/sample-dua.html

[4] Minnesota Population Center - https://www.pop.umn.edu/

For an example of the dedicated safe setting, in 2014 the ESRC announced funding for prefabricated secure settings, ‘SafePods’, structures with the dimensions 2m x 2m x 2m that can be installed within institutions (Administrative Data Liaison Service, b). These SafePods offer organisations like the ESRC an opportunity to standardise conditions nationally across secure-setting locations. In a technical report for the Administrative Data Taskforce (Administrative Data Taskforce (Technical Group), 2013), it was noted that the SafePods would operate thin-client terminals that connect the UK Data Service by secure VPN. In addition, the SafePods have controlled access, CCTV, screens to prevent anyone other than the user seeing the data, and strong visual cues to remind users of the need for, undefined, ‘safe behaviour’. This report also noted that perhaps automatic verbal messages could be played within the SafePod.

As the level of data sensitivity is higher in a secure-setting, the penalties for a breach of the usage agreement are more severe. For example, The UK Data Archive’s Secure Lab has a penalty policy that at its most extreme includes “Permanent suspension from all ESRC data services ([for the] individual); [and a] 5 year suspension from all ESRC data services ([for the]institution); [and] permanent sanction from ESRC funding ([for the] individual); [and a] 5 year sanction from ESRC funding ([for the] institution)” (UK Data Archive, 2014, 9). As such a breach would be a serious violation of data confidentiality, the policy also includes the possibility of a criminal offence under the Statistics and Registration Service Act (2007). This Act includes the possibility of a fine of £2000, a two year jail sentence, and a criminal record.

Working with data in a secure-setting

Having described a number of access arrangements, the position of the researcher is compared in three different scenarios: local physical access (including data downloads), physical secure setting access (including SafePods) and remote access to secure settings. It is useful to set this comparison on a scale from maximum user experience to maximum data security as this resonates with the conceptual balance of disclosure risk and data utility in Chapter 2. For users local physical access occupies the most familiar position on the user experience spectrum. For users with this type of access (it is assumed here that the baseline of ethics approval and access being granted have been successful - this will be assumed for all three types) will notice little disruption to their usual workflow. Data are often supplied in a user specified format (or a common format such as Comma Separated Values (CSV)). This means that a user can use their favourite software packages on their local machine to carry out analysis and generate outputs. Not only do they have access to software, but also their own previous work, code snippets, and literature, as well as an internet connection and ready access to online resources. These resources are combined with time flexibility and the physical environment the user is used to operating in, which often also includes access to colleagues for assistance with methodological queries. Potential threats to data confidentiality

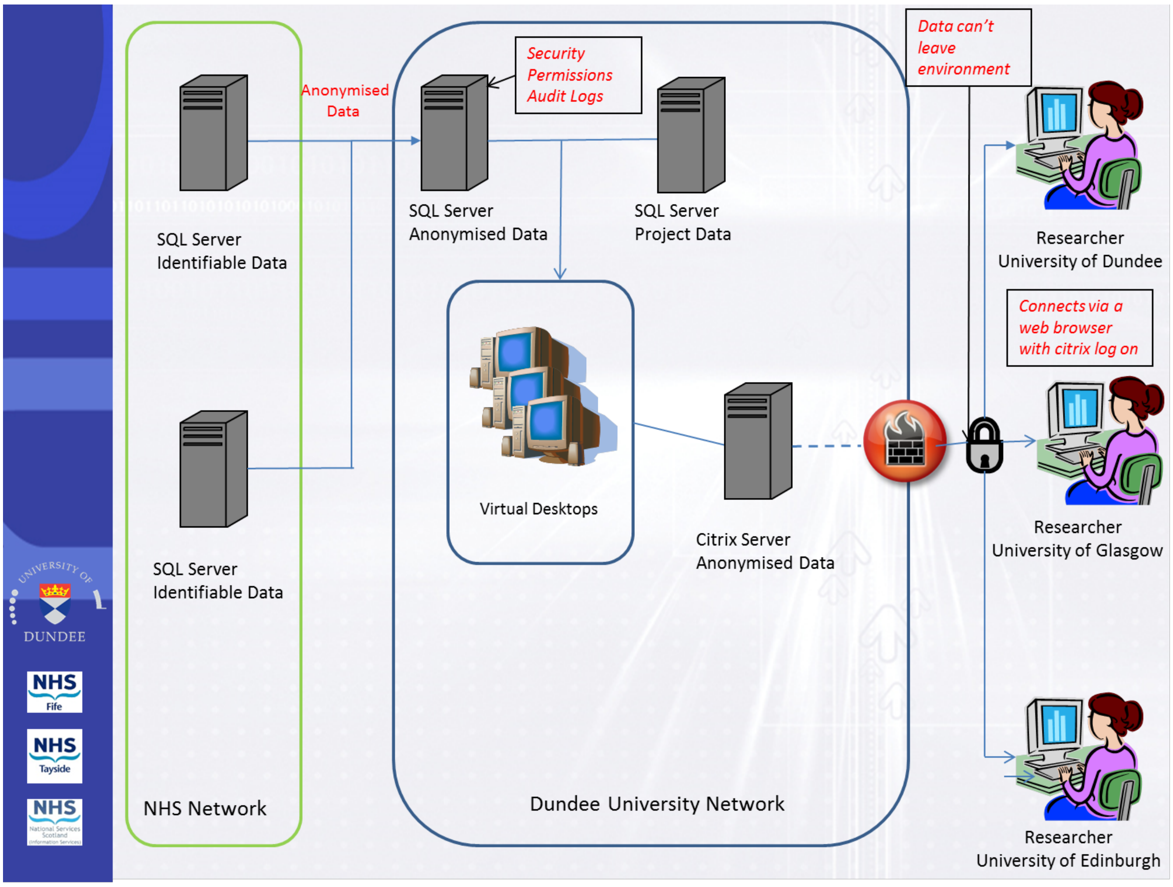


Figure 5.2: The Health Informatics Centre Services Safe Haven Environment

Source: https://medicine.dundee.ac.uk/hic-safe-haven include loss of physical media, unauthorised access to a researcher's machine, distribution to unauthorised users, lack of a user activity log, unauthorised persons viewing the data, and failure of users to check outputs for statistical disclosure. This type of access relies heavily on the trust between users and data owners mediated by a user agreement. It should be noted that despite these threats, there is little evidence that suggests these problems have been realised except in isolated cases.

At the other end of the scale is a dedicated physical secure-setting, the SafePod will be used as an example of this type of access in contemporary debates although most operate similar procedures, for example NHS Scotland's electronic Data Research and Innovation Service (eDRIS) has similar rules to those extrapolated from the Administrative Data Taskforce's technical group report (NHS National Services Scotland, 2013a). A user who has been granted access to sensitive data for their research must first attend a compulsory accredited training course on data security (for example, the Administrative Data Liaison Service's - Safe Researcher Training (Administrative Data Liaison Service, a)). Once their training is complete, they must book into a SafePod either at their own or another institution. They cannot take into the secure-setting any printed material, physical media, or equipment (e.g. a mobile phone) that they could use to record or store data from the secure-setting unless it is vetted on entry and exit. All activity in the secure-setting is monitored using CCTV and logging software on the client machine. The researcher will have access to a range of common statistical software packages, however if extra analytical packages or user written extensions are required these must be requested in advance. Their time is limited dependent on the booking system and popularity of the particular secure-setting. Unless pre-arranged and vetted, the user does not have access to their own previous coding or statistical package scripts. They have no direct access to colleagues for methodological support and can only communicate with the data owner, research coordinators or other members of the their research team who are named in the data access agreement. The workspace might also have posters and audio reminding the user of the rules for using the space or deterrent messages about 'safe behaviour'. In this data access scenario the risks to data confidentiality are minimal. The data owner retains all control over how the data is accessed and used. Importantly, almost no trust is placed in the user beyond that they do not memorise data for later malicious use.

The third scenario considered lies somewhere between the two scenarios described above. In this scenario, a user is given remote access to a secure setting from their own machine through the use of secure virtual private networking software. Data are stored at the server side and there is a firewall between the local client and the server. The user has access to a common range of statistical packages on the server (again any additions to this need to be agreed in advance). Analysis can be carried out at a time convenient to the user and over any duration. They have access to their own physical environment including literature and previous work. No files can be transferred from the client to the server, so previously written scripts and code cannot be used directly but are readily available as a reference. They can have access to the internet (via a separate machine) for methodological support, and access to colleagues. All activity conducted while working within the secure-setting environment is logged and can be reviewed by data owners. The threats also lie between the two scenarios above; users could write down or photograph on-screen data, they could communicate data to unauthorised persons (intentionally or not), if the remote system security does not contain some form of personal verification or biometrics then users could provide their login details to unauthorised persons. In this scenario there is a balance of trust in users with providing a level of practical data utility to facilitate analysis.

All three scenarios share the need for a data access agreement, which require the users to agree to common principles of data security, regardless of the mode of access, for example compare NHS National Services Scotland (2013a); National Cancer Institute (2014); UK Data Archive. Therefore, the significant difference is the trust between users and data owners, and the amount of control data controllers are willing to cede to users. This distinction provides another hook into the theoretical framework of Chapter 2. This balance of trust and control versus data utility, in the form of best facilitating a user's workflow, is an example of Beck's fusion of value and ethical judgements with technical and scientific statements. It is technically possible to implement a range of access controls, but does it fit with one's values to implement a particular regime? If data owners want to facilitate increased and efficient usage of their data, this is a judgement that they need to make. As a practical example of type of judgement, members of staff from the Health Informatics Centre at Dundee were included in a series of semi-structured interviews. The full details of these are considered in greater depth in Chapter 6 however some respondent observations are relevant to this discussion:

The introduction of safe havens had been met with some scepticism by research respondents. Academics discussed the negative impact of accessing data for a prescribed amount of time in a location away from their own work environment. However they acknowledged that the changing information governance landscape has meant that changes to traditional data access methods were necessary. They also discussed the positive aspects of working within a safe haven environment. These largely centred on the capacity for collaborative working supported by the virtual research infrastructures (These are discussed later in this Chapter) that could accompany safe havens (e.g. websites, forums, shared drive spaces).

Based on these observations, the remote-access secure setting operated by organisations like HIC can enforce a high level of data security while also being sensitive to the needs of users. The proposed SafePods, that at least in their design, appear to provide little compensation to user needs, could limit the volume of research carried out on the sensitive data to which they were supposed to facilitate access. In the next section, generating outputs becomes the focus, and the process of output checking is considered in the context of these access arrangements.

5.2 Generating Outputs

In this Chapter the models of access and the process of analysis have been discussed. Three models for data access were scrutinised for their effect on the researcher experience. In this section those three models are evaluated in terms of their impact on how research outputs are generated and how they are processed by data owners. The treatment of outputs can have a significant effect on the researcher experience as ultimately it is the research outputs that are published or disseminated to further the body of knowledge. This is also where the research into statistical disclosure control described in this dissertation is felt most keenly as data owners seek to protect the confidentiality of their data perhaps to the detriment of the researcher's analytical narrative.

Let's return to the examples of the traditional 'data goes to the researcher' model. The SEER data sets are covered by a user agreement. This user agreement passes the responsibility for the reviewing of outputs to the researcher. The project's sample agreement contains the common principle that researchers will not: "present or publish data in which an individual patient can be identified [...] will not publish any information on an individual patient, including any information generated on an individual case by the case listing session of SEER*Stat [...] will avoid publication of statistics for very small groups" (National Cancer Institute). A breach of this agreement is subject to the penalties described in the preceding section. Similarly, the UK Data Archive passes the responsibility of output checking (for standard licensed datasets) to the user through clause 8 of its End User Agreement: "To preserve at all times the confidentiality of information pertaining to individuals and/or households in the data collections where the information is not in the public domain. Not to use the data to attempt to obtain or derive information relating specifically to an identifiable individual or household, nor to claim to have obtained or derived such information. In addition, to preserve the confidentiality of information about, or supplied by, organisations recorded in the data collections. This includes the use or attempt to use the data collections to compromise or otherwise infringe the confidentiality of individuals, households or organisations" (UK Data Archive).

Using the conceptual scale of security versus user experience, this transference of control to users allows them the greatest freedom to use their own judgement in the creation and publishing of research outputs while the data controllers retain the right to retrospectively assess whether any breach of the user agreement has been made. This freedom afforded the user is problematic if the data are particularly sensitive, as the literature, training and tools for reviewing outputs for statistical disclosure control from a user's perspective are scarce. As such, the confidence of data owners in the user's ability to adequately review their outputs cannot be considered high in the general case. This subsequently impacts on their own responsibility as data owner to have taken reasonable precautions to protect their data. This logic underlies some of the control decisions in the other two models.

In the contemporary example of physical secure-settings, the reviewing of outputs is a process reserved for the data owners. Users carry out their analysis within the secure-setting and then submit research outputs to the data owner for review. It is only once they have been cleared by the data owner that research outputs can be published or presented outside of the secure-setting. eDRIS, NHS Scotland's data access project, summarises the review process in their user agreement. Outputs are submitted to Research Coordinators who carry out tests and, if necessary, apply statistical disclosure control methods before outputs are given clearance. The Research Coordinators can request code, syntax and a summary of how the outputs were generated. If a Research Coordinator denies clearance for a particular output, the user can appeal but the final decision rests with the data owner. In addition, the user agreement adds the clause that abstracts and papers intended for publication can also be subject to review by Research Coordinators (NHS National Services Scotland, 2013a, 7). The Northern Ireland Longitudinal Study (NILS) provides more detail than the eDRIS project on the review process. NILS policy (Northern Ireland Longitudinal Study Research Support Unit, 2013, 4-5) states that a Research Support Officer is responsible for output review; this includes:

- If data are to be released in tabular form, then the NILS support officer must ensure that any information that could potentially identify an individual is aggregated, suppressed or removed as appropriate.

- When releasing tabular data NILS support personnel must ensure that cell counts are 10 or greater. If associated data allows the cell to be split then the support officer must aggregate the data to the highest level consistent with the need to explain the results.

- No data on birth dates of NILS members may be released, with the exception of year of birth. Any analyses which require month of birth or full date of birth will be conducted by NILS-Core staff.

- No data on date of death of NILS members may be released, with the exception of month and year of death. Any analyses which require day of death will be conducted by NILS-Core staff.

- Exposure times may be included in aggregated datasets provided there is more than one event in each cell.

- Sample uniques or individual cases are never allowed.

Introducing the role of Research Support Officer or Research Coordinator as 'gate-keepers' for research outputs allows the data owner to retain full control of published outputs from their datasets. However, the process of review places an extra strain on the users experience of a data set. Neither eDRIS or NILS make a commitment in their user agreements and policies to review outputs within a specific time-frame, or in the eDRIS case, a commitment to a consistent approach to their review. This uncertainty over review workload and time-scale makes it difficult for users to plan their research accordingly. For example, the time taken to review outputs and make any potential appeal against a negative decision would need to be considered ahead of any submission to an academic journal or conference. Indeed, from the qualitative work carried out during this research, it was noted that labour intensive human processes for output review often formed bottle-necks in the research workflow. It can be assumed that current practices for reviewing outputs (the details of which are discussed further in Chapter 6) will not be adequate at scale when the volume of data and the number of users increases still further.

The remote access secure-setting completes the picture of the three models of access. As data are always kept on the server side of the client-server connection, these remote access secure settings operate review procedures in a similar way to the physical secure-settings. Users request outputs via their client machine and these are then reviewed by the data owner before release. HIC operate a remote access model with centralised reviewing procedures. In their model users submit outputs for review, and a commitment is made to review them between 9-11am on the next working day after submission (Health Informatics Centre (Dundee)). The HIC example, alleviates some of the uncertainty highlighted above in the case of the physical secure-setting. This is done by providing a time-scale for output review. However, there is little difference between the physical and remote access secure-setting in terms of reviewing research outputs. As has been shown in the proceeding section, the benefits of the remote access secure-setting are realised in the data access and analysis aspects of a users workflow. What the HIC example does show, however, is that consideration for the user experience can be made and data owners can provide users with details of their processes so that users can be fully informed when planning their research.

In this section, the role of data owners and users were considered in the context of output review for different access models. As with data access, the balance of trust in the user and risks to data confidentiality is calibrated by the data owner. This is done by increasing or decreasing the amount of control over research outputs ceded to the user. For sensitive data, like those in the health field, it would be difficult for the data owner not to review outputs and still fulfil their obligations to uphold the confidentiality of the data subjects. However, the approach adopted by the data owner and the level of detail and communication with the user can have a significant impact on the user's workflow and ability to plan their research effectively. In the next section, the focus shifts to the possible incentives data owners can offer users to counter balance the restrictions of using a secure setting. This is done primarily in the context of remote secure-settings within virtual research environments, but the incentives could also be implemented in a physical secure setting.

5.3 Virtual Research Environments

The Virtual Research Environment (VRE) has been discussed in the literature at least since the mid-nineties and has appeared across disciplines, see Donaldson (1997); McKee (1995); Nikolov and Nikolova (1996). VREs are defined as any system that allows researchers remote access to research data and tools out-with the user's location and local network. This wide definition could now apply to the plethora of 'cloud' services such as Google's Drive and Docs, and Dropbox, which are often used by researchers and teams to manage access to and execute data management for research data. The variety of VREs that are specifically relevant to this research are those that emerged from e-Science projects in the late two-thousands including MethodBox and MyExperiment, see (The MethodBox Project, 2011; De Roure et al., 2008). In addition, the Data Management through e-Social Science (DAMES) project developed a range of web portal based services for researchers to share code and syntax in an attempt to mitigate the volume of repetition in quantitative social science. DAMES also crowd-sourced the creation and improvement of metadata, and although the services deployed through DAMES were classified as Grid Enabled Specialist Data Environments (GESDE), they fit within the above definition. For details of the DAMES services, see Lambert (2010); Doherty et al. (2010).

These types of virtual infrastructure has sought to address many of the barriers to efficient analysis, especially by closing the gap between researchers and the technical infrastructure needed to analyse complex data. For example, the MyExperiment project notes that its objective is to improve the researcher experience (De Roure and Goble, 2007). In their introduction De Roure and Goble provide comment on the need for automation: *"The techniques of e-Science help the scientist deal with increasingly large and increasingly complex scientific applications. Key to this is automation, and several scientific workflow tools have become established as a means of automating the processing of scientific data in a scalable and reusable way"* (De Roure and Goble, 2007, 1). Their particular approach was to adopt a number of the 'web 2.0' social technologies to enhance the user experience through collaboration and sharing, but its the more abstract principle that e-Science approaches can facilitate the researcher in analysis of complex data that is most relevant to this research.

Similarly the DAMES project sought to harness the potential of VREs and e-Infrastructure to support data management processes. Tan et al. (2009) establishes three key reasons that underpin the DAMES approach. First, the recognition, which is also made in this research, that the volume of data sources available to researchers is increasingly vast and that data management processes such as formatting, data linkage and manipulation occupy a significant amount of a researcher's time before any analysis is carried out. Second, Tan et al acknowledge a significant skills gap for researchers when it comes to the computer programming skills required to carry out complex data management tasks[5]. Third, Tan et al noted that a focus on data management using e-Infrastructure could afford the researcher new opportunities for more complex data-linkage and analysis.

At an more abstract level, Keraminiyage et al. (2009) set out "critical success factors for collaborative research" which provide a relevant framework for the benefits that VREs can provide if these success factors can be adequately satisfied. Table 5.1 summarises this framework. Although, the original targets of this framework were researchers working in collaboration, it is equally valid to see the relationship between researchers and data controllers as a collaboration in a similar vein. This framework will be revisited after VREs have been discussed in the eHealth and statistical disclosure control context.

Focus element	**Success factors**
Trust	Mutual respect and trust among partners, Good personal relationships, Simple collaborative agreement, Clear and honest understanding of each other's abilities
Commitment Ability and Leadership	Top managerial commitment from all parties, Active participation on project team by all the parties, Adequate resources, specialist and complementary knowledge and expertise of partners, One agreed project leader with required authorities
Transparency and clarity	Common goals with no hidden agendas, Clear understanding of each partner's responsibilities and tasks, Clearly defined objectives Clearly defined responsibilities, Mutually agreed project plan, Realistic aims, Defined project milestones, Focused project scope
Communication and monitoring	Effective communication, communication and regular contacts with partners, Regular progress monitoring, and Ensuring collaborators deliver, Monitoring project's progress against agreed milestones

Table 5.1: Success Factors of Collaborative Research

Table adapted from Keraminiyage et al. (2009, 61)

[5] The European Commission highlighted the continuing lack of general coding skills five years after Tan et al, see Kroes and Vassiliou (2014).

VREs in SDC Workflows

Despite the above and many other examples of innovation in the VRE sphere, statistical disclosure control has not received the same attention and still presents a significant challenge for these e-Infrastructures. Many rely on the same onerous human processes used in the brick-and-mortar, physical secure-settings that proceeded them and have not heeded the 'automation is key' message from the e-Science programme. What will be set out here is a brief argument in support of VREs in the eHealth sphere (and more generally in quantitative research that utilizes sensitive data). This will be discussed in the context of eHealth projects which could provide a VRE for researchers with tools that benefit their workflow, but also tools that support the data owners own data management processes such as statistical disclosure control.

First, given that the primary focus of this research is data that are subject to particular scrutiny because of their implications for privacy and confidentiality, the earlier definition of a VRE should be made more prescriptive. Therefore, in this context, the scope of a VRE is limited to one which handles data that is not publicly available and poses significant risks to privacy and confidentiality. In the earlier discussions of data access and generating outputs, it was noted that data controllers must strike a balance between the trust placed in individual users and their responsibilities as a data controller. However, it is possible to calibrate that balance so that reasonable steps are taken to protect the data, while also offering incentives to researchers to mitigate the negative aspects of secure-settings. For the researcher, access to social technologies within a secure-setting, where researchers can share metadata and other resources, can mitigate against the closed nature of secure-settings, for example see the types of tools discussed in (Allan, 2009; Keraminiyage et al., 2009). These technologies primarily address the concerns of researchers as opposed to data owners. However, these technologies can also be used to benefit the workflow of the data owner.

In the previous section, the potential bottleneck of output review was discussed. This bottleneck arises when the volume of outputs generated exceeds the data owners capacity to review them within a reasonable time period. If data owners could employ some automation in the process of output checking and disclosure control then their capacity to review outputs would increase. This is because some of the analytical burden is shifted to systems which in turn allows the data owner to concentrate on the results of those analyses. These systems could provide features that ensure good data provenance and the prospect of reproducibility, while also communicating appropriate status information to researchers. These combined processes of automation and communication with researchers help to build trust between data owners and researchers and support research collaborations.

Consider a data owner that controls individual level data on hospital admissions. The data are deemed too sensitive for public release and access is restricted to a secure-setting. The data owner provides computing resources for data analysis and opts to provide a virtual research environment for the data users. Similar to previous VREs, users have access to wikis, forums and other mechanisms to share research objects and metadata. In addition, the data owner's statistical disclosure protocol is semi-automated. When users submit outputs for review, the data underlying their outputs are automatically submitted to pre-determined risk analysis procedures. These procedures are pre-determined through a process of review by data owners of their data's characteristics and the contemporary data environment (as discussed in Chapter 3). Based on the results of these risk analyses, some potential statistical disclosure control methods could be applied automatically to the data. As a result of these automated sequences, the data owner has a body of evidence to review when they review the user's output. Further, analysis and consideration could be required before a final decision is made by the data owner. The system can offer a tracking facility for users to track at what stage their outputs are being considered, and in addition, the system, with the approval of the data owner, can provide access to the evidence base for the data owner's decision on statistical disclosure control measures.

The approach proposed in abstract above provides benefits for the user and the data owner. It serves to mitigate the barriers to efficient analysis highlighted earlier in this chapter as well as provide a novel use of e-Science approaches to the problem of statistical disclosure control within a Virtual Research Environment. Returning to the success factors set out by Keraminiyage et al. (2009), it is possible to specify those areas of the collaborative research process that this new approach addresses. The ability for data owners to clearly demonstrate their SDC workflow to researchers and provide timely output review resonates with three of the focus elements: status updates on the review process and documentation provide a level of communication between the researcher and data owner so that research activity can be adequately planned and revised. Metadata concerning data provenance and a body of evidence to support SDC decisions help to build trust between the two parties and also demonstrate the data owner's ability and specialist knowledge.

Options for the Operationalisation of the User Experience in SDC workflows

At the outset of this chapter it was argued that the user experience is an important factor in the SDC workflow. As the demand for data increases it is important that the resources deployed in support of the data can adequately cope with the increasing number of users and their increasingly complex analytical demands. As Deelman and Gil (2006) discuss in their work on user experience and workflows for e-Science projects, providing users "with satisfying experiences and enabling them to conduct their science efficiently and effortlessly stem from the fact that user expectations vary greatly" (Deelman and Gil, 2006, 146). This variance in user expectations is also true of the research environments on offer when analysis of sensitive data is carried out. In our treatment of data utility in 4 the focus is on the data behaving as close to the original as possible (without putting the data at unacceptable risk of disclosure), here the focus is on research environment behaving as closely to a users own local environment as possible, and potentially enhancing that environment through the access to new resources.

In to preceding sections the approach of data owners and controllers was presented through this prism of user experience. The importance of communicating information to users at the right level of detail and at a frequency that enables their workflow to proceed efficiently is a common factor in user experience approaches, and Deelman and Gil (2006) also highlights this in the e-Science context ensuring that users receive feedback from systems if a workflow has stalled or needs further input from them to continue. Deelman *et al* also acknowledge that users analysis develops over time, through testing hypotheses, applying different methods or partitioning data in myriad ways. This means that enhancements, like the sandbox a VRE can provide, as well as providing users with 'dummy' data to develop their workflow outside of the secure environment can be important.

In making the case for greater consideration of the user experience, it is important for data controllers to understand what success in this field might look like. At a superficial level this could take the form of aggregate statistics on the number of users, the number of research outputs or statistics on average time taken to complete a workflow. This type of measurement provides information that standard reporting might produce for the use of a data resource. For example, the UK Data Archive reported this sort of information in their annual report and attempted to draw some further conclusions by linking the usage of their services with the rating of user institutions in the now defunct Research Assessment Exercise (RAE) (UK Data Archive, 2002).

However, in the scenarios one could envisage in the eHealth field, the total number of users could be quite small (compared with the UKDA) and the complexity of their usage could vary greatly. This would make extracting good measures of the user experience from these types of metrics problematic as the numbers would carry a lot of noise. The total number of users, and the quality of research outputs, would provide a good indication of the return on investment made to improve access but still does not capture the user experience per se. Data controllers could take a more qualitative approach to the user experience during the development of data access services and research environments, a good practice example seen above is the Health Informatics Centre in Dundee where users contributed to the development process. This approach could be developed further if data owners draw on the literature of software development, especially areas such as user-centred systems design (Gulliksen et al., 2003).

In production, the systems designed to facilitate access and analysis could provide data controllers with monitoring of key metrics. These metrics would probe potential bottlenecks in the SDC workflow such as output review or the granting of access privileges. If enhancements, like those discussed above (wikis, forums, etc.), are introduced the uptake of these tools could be monitored and also provide mechanisms for informal feedback between users and data controllers. It is in this qualitative feedback from users, and perhaps a study of potential users - those that produce similar research but that have not applied for access, that would yield the most potentially valuable results. This type of qualitative evaluation has been seen elsewhere in this thesis and is tool already used by studies of data subjects and their attitudes toward data access and sharing (Aitken et al., 2011).

In summary, data controllers can use existing approaches from the fields of software development (user-centred systems design), virtual research environments, and quantitative and qualitative evaluation in junction with the technological advances of Web 2.0 infrastructure to assess and better incorporate the user experience in their SDC workflows. The next Chapter will provide an example design for a tool-set that can be integrated with the existing workflows of data owners and other e-infrastructure in the research data environment to further the approach suggested here.

Chapter 6

An e-Science Model for Disclosure Control

This Chapter draws together the theory and literature of the proceeding chapters and posits that an e-Science approach to the problem of statistical disclosure could address the problems highlighted, especially in Chapter 5. It begins with a discussion of the NIAH set of tools. As will be shown, these tools were created in part through an ESRC funded internship with the Scottish Government which provided access to potential users of these tools. Access to stakeholders in the disclosure control process allowed for the collection of qualitative data to inform further development of initial tools. The discussion begins with a brief summary of the first iteration of NIAH and then proceeds to summarise and build upon qualitative data collected from semi-structured interviews with stakeholders in the statistical disclosure control process. Once the motivations for the tool-set has been described in some detail, the attention shifts to how these tools can be integrated with existing or proposed e-Science infrastructures such as STAT-JR and VANGUARD.

6.1 NIAH - a k-anonymity implementation

In this section, a justification for the initial development of NIAH (see Chapter 3 for details of the NIAH methodology) as a separate piece of software, as opposed to scripts for statistical packages, is set out. The purpose of this research is to provide evidence that algorithms that invoke a range of statistical disclosure control methods in a semi-automatic way will enable data providers to release data with a higher level of data utility without any increase in disclosure risk when compared to existing methods. As such, it was important to be able to develop research tools that could test this functionality. Existing methods, as discussed in Chapter 5, are often provided by *ad hoc* user written scripts for statistical packages such as

SAS, SPSS, stata or by other existing software specifically designed for statistical disclosure control, for example μ-argus (Hundepool and Willenborg, 1996).

The context for NIAH's initial development highlighted problems with existing methods used in the day to day operations of government departments. During work carried out in collaboration with the Scottish Government's Health Analytical Services Division, and the Office for National Statistics' Methodology Directorate, a statistical disclosure control methodology was developed for the Scottish Government's Home-care Census (Scottish Government, 2003). Similar methodological work and data analysis had been carried out for the 2010 dataset. In 2010, a k-anonymity assessment of the home care census variables had been made using user generated scripts in the statistical package SAS. Variables were hard-coded into these scripts and no metadata were provided to explain the relevant sections of SAS code. Using these scripts on the latest data available (2011) proved to be difficult without forensic examination of the SAS code and substantial re-coding to reverse-engineer the process. In addition, the output generated from the SAS scripts still required a significant amount of user analysis to interpret the results. This was in part due to SAS not being explicitly designed for k-anonymity assessments and in part the type of reports generated by the scripts. These issues also contributed to resource constraints. The Scottish Government operate a centralised SAS statistics server with user terminals running tasks on the server from local SAS clients, the output reports generated by the user scripts often put considerable strain on the client side machine, and on occasion, rendered it inoperable.

Given the research statement above, NIAH was designed to be portable and usable without knowledge of a particular statistics package. It is the position of this research that this approach provides the following benefits over user generated scripts:

- NIAH provides a generic implementation of a k-anonymity assessment. This means it can be run on any dataset in the specified CSV format, without adjustment of package specific syntax files.
- The analysis provided by NIAH is replicable. The same analysis can be re-run, and changes to the k-anonymity assessment are trivial to make. In the further development use cases provided, it is also envisaged that these tools will keep a log of all activity in a format that will provide users with an audit trail of how data have been treated to ensure good data provenance.
- NIAH is portable. Written in Java, NIAH can be run on a range of systems without compatibility issues.
- NIAH scales with the available resources. NIAH can process large datasets efficiently if sufficient RAM is available - many statistical packages have record/variable limitations.

- NIAH supports remote execution. The implementation of NIAH used in Health Analytical Services Division uses a web service to connect to the tool and run it remotely on a central server (this is discussed further at the end of this chapter). This provides the opportunity for developing statistical disclosure control processes that fit with the virtual research environments discussed in Chapter 5, where data sources might be distributed across a number of physical locations (for an example, see the VANGUARD system in section 6.4).
- The GPL v3[1] license means NIAH can be used and modified freely within the terms of license; this, combined with the portability, enables data providers to set up their own processes across different systems This again removes some of the barriers from data controllers, as specific licenses for statistical packages are not required to carry out k-anonymity assessments.

6.2 The data users' perspective

As has been outlined elsewhere in this dissertation, statistical disclosure control forms part of the overall exploration process when working with quantitative data. It is essential that data owners are not only aware of, but also that they employ strategies such as those that have been summarised in Chapter 4 to understand, the disclosure risk their data poses. It is also important for data owners to ensure that their strategies balance not only data utility, but also the data's usability as discussed in Chapter 5. As outlined in the preceding chapter, data owners should attempt to gather qualitative data from users and other actors within the SDC workflow in order to limit the barriers to efficient data utilisation. Data controllers can incorporate the users perspective through system and process design methodologies like user-centred systems design (again see Chapter 5) and also through the implementation of informal and formal feedback mechanisms.

As a pilot study, that might satisfy in part the user experience requirement set out above, a small-scale interview study was undertaken. Between March and June 2012 five semi-structured interviews were carried out with users of quantitative data, with a focus on linked data. The users were drawn from academic, health board and local authority backgrounds in order to offer evidence which provided coverage both relevant to this research and the Scottish Government's data linkage project; linking health, housing and social care data (HHSC) (Scottish Government, 2011a) for more details on the methdology of this study see Chapter 3.

These interviews covered a range of topics including metadata, remote access, disclosure

[1] details of the licence can be found at http://www.gnu.org/copyleft/gpl.html

control methods, and data analysis workflows. There was a clear distinction between academic users and other analysts in terms of their expectations regarding data quality and level of available detail across variables in a dataset. The academic participants had been working with linked data for some time, and this had occurred under a number of different information governance regimes. Previously, linked datasets had been provided to researchers directly on CD or by FTP transfer direct to their computers. This earlier model placed the majority of the responsibility burden for security and disclosure control upon the individual user, not the data owner. Therefore, the introduction of physical infrastructure such as safe havens, or other restrictions on data accessibility, have been met with some scepticism by researchers. Academics discussed the negative impact of accessing data for a prescribed amount of time in a location away from their own work environment. These included lack of access to good metadata or previous analytical work, as well as other local resources.

However, they acknowledged that the changing information governance landscape has meant that these changes were necessary. They also discussed the positive aspects of working within safe haven environments. These largely centred on the capacity for collaborative working supported by emerging virtual research infrastructures that accompany some safe havens (e.g. websites, forums, shared drive spaces). This topic is revisited in part in a later discussion regarding the metadata from statistical disclosure control process.

For the analysts, whose primary role involved service planning and analysis used for operational purposes, these streams of linked data were quite new. As a consequence, expectations were not as concrete as those expressed by academic users. A general feeling at the local operational level was that they supply data to a number of projects and very rarely see any return or engagement beyond the initial transaction.

Analysts expressed some concern about their capacity to manage and analyse this new data effectively. Informal feedback from different local authorities (LAs) suggested that the resources allocated for analysis, especially in terms of staffing, and the use of statistical packages, varies widely across LAs. The Scottish Government, for its part, has acknowledged that some analytical support will need to be in place to help LA's and health boards make the most of new data on offer. Statistical disclosure control was highlighted as a particular issue in this context. Analysts reported a lack of resources both human and technical to support statistical disclosure control work beyond the *ad hoc* solutions developed by individual analysts or analytical teams.

The areas of convergence for all users lay in the work environment and tools to aid efficient analysis. The theme that emerged most strongly from these discussions was what work could be done before entering safe havens or other physical infrastructure. Metadata was a popular topic in this vein. Respondents expressed a view that good metadata is often discussed but rarely implemented successfully — it is important that data be detailed and accurate, but also accessible, a point emphasised by a number of respondents. Projects often opt for comprehensive data dictionaries which can be inaccessible to novice users without the right tools for self-navigation. For example, the NHS ISD's data dictionary for the Scottish Morbidity Record (SMR) datasets contains a significant amount of detail, but few tools for navigating it by topic of interest, keyword, etc. In contrast, a good example of metadata is provided by the British Household Panel Survey (BHPS) which is now part of the Understanding Society Survey (Boreham and Constantine, 2008).

Extending metadata from a flat document into a more interactive format was also an idea that both groups discussed. This was considered in a similar vein to the transition of websites from flat html documents to Web 2.0 structures (such as forums, social networking sites, etc.). Participants expressed an interest in the development of user-led metadata that could complement the 'official' metadata. They suggested that this could include space for users to work collaboratively on a piece of analysis, or where users can post their own comments, derived variables and pieces of code. It was noted that projects like Method Box (The MethodBox Project, 2011) and MyExperiment (De Roure et al., 2008) appear to include the type of virtual infrastructure that users expressed support for. Both of these projects were created through e-Science funding calls by UK research councils.

In addition to good metadata, it was also noted that access to good training data could help make user's time in the safe haven more efficient. Training data would give users a safe dataset, formatted in the same way as the real data, upon which they could develop their analysis prior to accessing real data. Both groups of users expressed an interest in this idea, with different motivations. For academic users it would negate some of the constraints that accessing a physical safe haven might place on their workflow. For the local analysts it would be a useful tool for learning about a dataset that they had not used before, so they know what to expect from the real data.

It was also observed that this combination of training data, good metadata and a view to more collaborative working would reduce the amount of 'reinventing of the wheel' that participants suggested occurs in this field. For example, without communication between groups of analysts, the same derived variable or cross classification will be created using similar code numerous times. Lastly, participants identified the impact that human resources would have on their own work. Issues were raised such as potential bottle necks in output checking, and general response times to queries if specific knowledge was concentrated in a few individuals. In the worst cases, some participants suggested that these issues posed a threat to the data being used at all.

From these findings, general themes can be drawn out to inform the development of tools that advance this research but also address some of the needs of the user community. The general themes identified here are: modes of access to data; the responsibility for disclosure control; communication and collaboration between users, and data controllers; the specialist knowledge and expertise burden for analysis; and access to good metadata and training data. Some of these themes are picked up elsewhere in this research. These will also form hooks for the narrative in the remainder of this chapter.

The domain knowledge of the participants in the above was combined with the literature reviewed in Chapter 4 and observations of disclosure control workflows during work carried out at the Scottish Government. From this a number of potential use cases were summarised to inform the development of research tools beyond just an implementation of *k*-anonymity (see Section 3.7).

6.3 Comparison with *ad hoc* user generated scripts

It is important to understand how the approach in this research performs when compared to existing methods. In order to compare the two approaches, the NIAH tool-set and an *ad hoc* script provided by NHS Information Services Division were compared directly. This script is typical of the type of user generated statistical package script that is used when assessing disclosure risk in a real-world analysis.

The basis for comparison is a k-anonymity assessment of a series of randomly generated comma separated datasets that range in size from 100,000 records to 4,000,000 records (file sizes in MB range from 10 to 382). To establish a point of comparison, the NIAH k-anonymity algorithm was implemented in both Java and C programming languages. These implementation were tested against each other using a large memory instance provided by Amazon cloud services. The specifications of the instance are provided in Table 6.1.

17.1 GB of memory
6.5 EC2 Compute Units (2 virtual cores with 3.25 EC2 Compute Units each)
420 GB of instance storage
64-bit platform
I/O Performance: Moderate
EBS-Optimized Available: No
API name: m2.xlarge

Table 6.1: Specification of the Amazon Cloud Instance used to test the Scalability of the implementations of NIAH in Java and C

For the C and Java implementations of NIAH, both execution time and the amount of memory used were recorded. These results are provided in figure 6.1 and 6.2. These results are not unexpected given the overheads associated with the Java Virtual Machine and Java's compliance with standards to ensure portability across platforms. The Java implementation only has results up to the 2M record test data as at 4M there was insufficient memory to complete the test. In both time and memory usage, the C implementation outperforms the Java implementation. As discussed earlier in this chapter, the algorithm trades off memory usage for faster execution times which explains the consistent scaling in the amount of memory required (average of ˜28 times the input file size for the C version and ˜62 times for the Java version).

As the C version of NIAH is a compatible version of the Java implementation, it is the C implementation that is tested against the *ad hoc* user scripts. The example script used here was provided by NHS Scotland Information Services Division. It is written for the SPSS statistical package. The same parameters that NIAH's command line interface requires are required to be hard coded into the script. Further to this, extra lines of code must be added for the number of key variables specified. Due to the extra set up costs in configuring the script per input file the comparison here should be seen as conservative, at best. This is because the execution time provided does not include these extra time overheads. Figure 6.3 shows the execution time comparison of the the C version of NIAH and the SPSS script for the same datasets, increasing in size. Due to the requirement for SPSS to be installed, and only a window licensed copy being available, a comparison of memory usage is not provided. At an anecdotal level, using the windows system monitor as a guide the SPSS runtime uses less memory to access and analyse datasets. Further information about the SPSS data structures and memory usage are not publicly available. As Figure 6.3 shows the C version of NIAH out performs the SPSS script for the range of dataset sizes provided. In this comparison, the limitation on dataset sizes was local memory on the test machine (8GB RAM). For the purposes of this research, the comparison up to 2 million records is adequate, given more resources and further time this could be explored further. However, as indicated earlier, the SPSS execution time values are quite generous given the extra script editing overhead.

What has been shown in this section is that the NIAH tool set can perform faster than *ad hoc* users scripts commonly used by data owners at the present time. Also the design decisions made in the development of the tool set are based on qualitative evidence drawn from both the potential data end users and data owners. In addition, it has been shown that NIAH has been developed with a view to added functionality and the integration of these tools within existing workflows for statistical disclosure control. In the next section, the latter of these two aspects, integration, will be discussed in the context of other emerging e-Science tools in the field of data analysis.

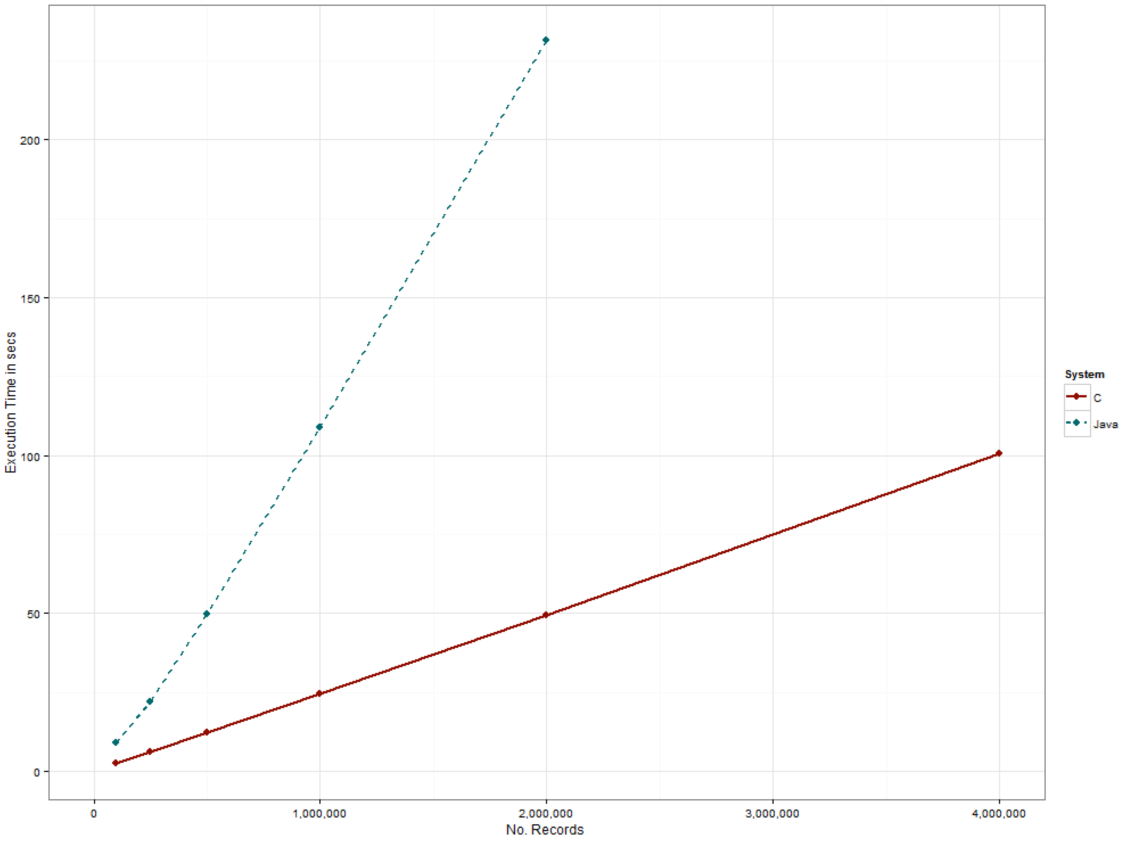


Figure 6.1: Execution Time for the C & Java Implementations of NIAH

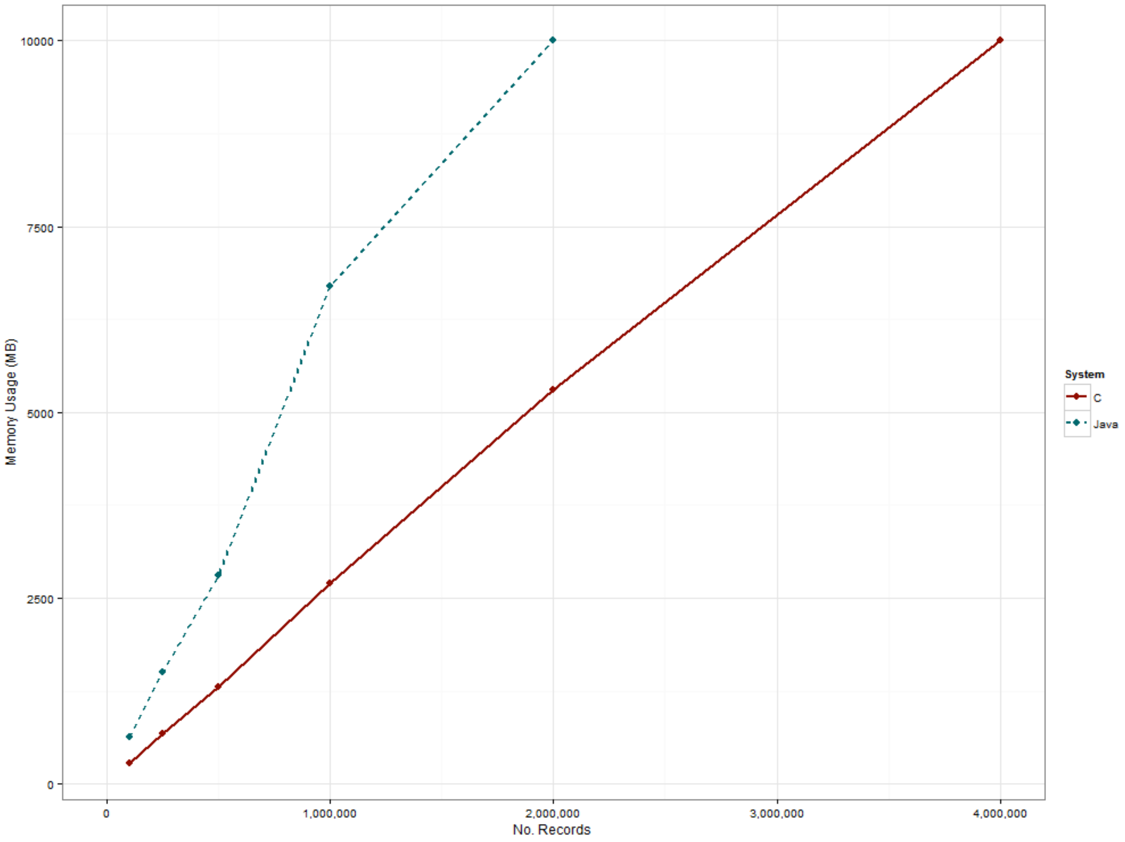


Figure 6.2: Memory Usage for the C & Java Implementations of NIAH

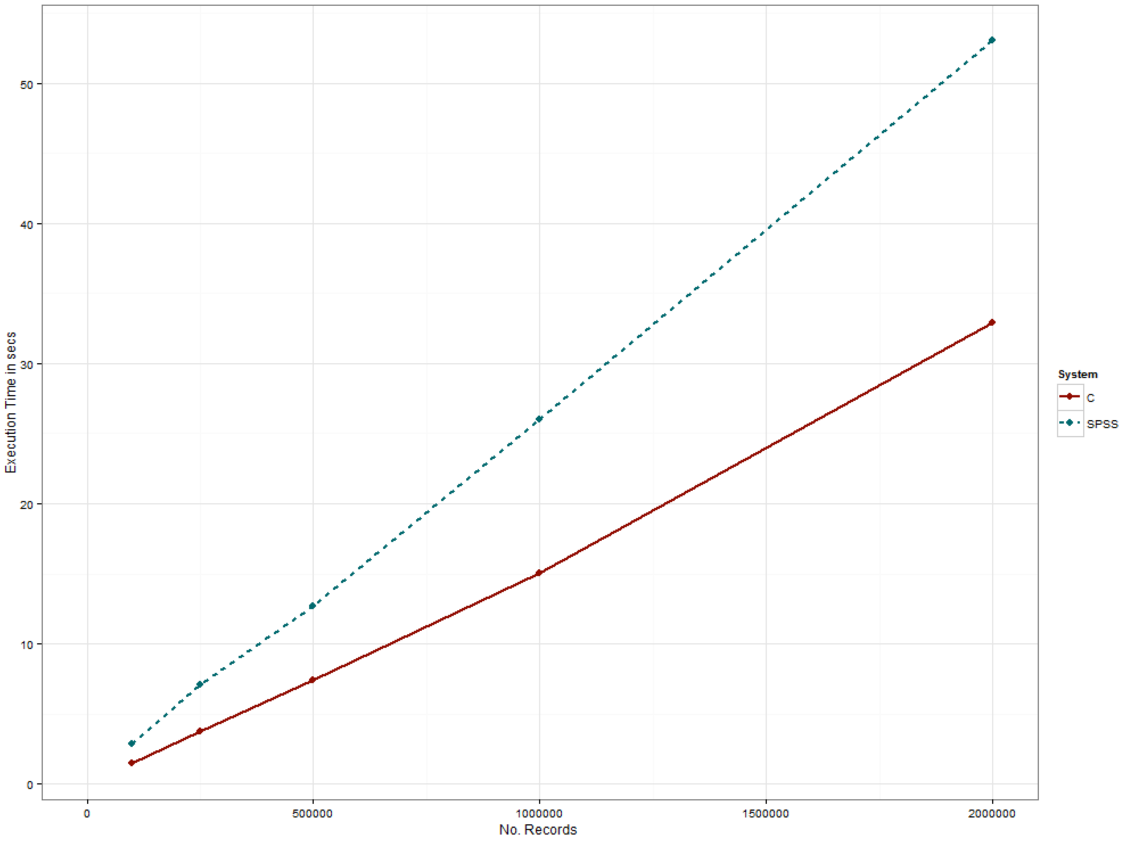


Figure 6.3: Execution Time for the C & SPSS *k*-anonymity Assessments

6.4 Integration of NIAH with existing e-Science infrastructure

As has been discussed in this chapter, the motivation for a tool set like NIAH is the provision of tools that can be easily integrated with existing workflows. The decision to follow a CSV data format allows for the transfer of inputs and outputs to and from the tool set via other data analysis packages. For example, in the overview of the broader range of NIAH's functionality, the stata command is defined. This command would pass the NIAH outputs to the stata package using stata's own command line interface to import the CSV data. This type of interoperability provides a simple example of the potential for integration. What is set out in this next section is a more complex integration of NIAH with existing e-Science infrastructures; STAT-JR and VANGUARD.

STAT-JR

STAT-JR[2] is described as "a software environment for promoting interactive complex statistical modelling" (Charlton et al., 2012, 2013). This Python based system allows users of different abilities to run statistical tests from their browser. It also provides inter-operability by providing a wrapper for complex methods provided by packages like MLwiN (Rasbash et al., 2009), JAGS (Plummer) and stata (StataCorp, 2013). STAT-JR fits a common e-Science model by providing a browser-based solution that allows users to easily access a range of statistical methods and tools to complement existing science workflows. To achieve this STAT-JR uses a series of Python templates which when served with parameters by the user through the browser interface calls the required tools and executes the task. The results are then presented through the browser and the user is given the option to download associated outputs.

Given this range of interoperability, STAT-JR would seem an ideal candidate for NIAH integration. This gives the users of STAT-JR the added functionality of carrying out k-anonymity assessments on their datasets prior to or after other operations have been carried out using the STAT-JR system. Due to the templating system used by STAT-JR the technical aspects of integrating NIAH are relatively trivial. A Python template was constructed that takes user parameters from the browser interface and enters them into an appropriate command line arguments string for NIAH. NIAH is then executed using the those command line arguments. The returned outputs are then displayed in the browser and the option to download the data is given. Below is a worked through example of how NIAH operates within the STAT-JR environment (including figures 6.4,6.5,6.6 & 6.7).

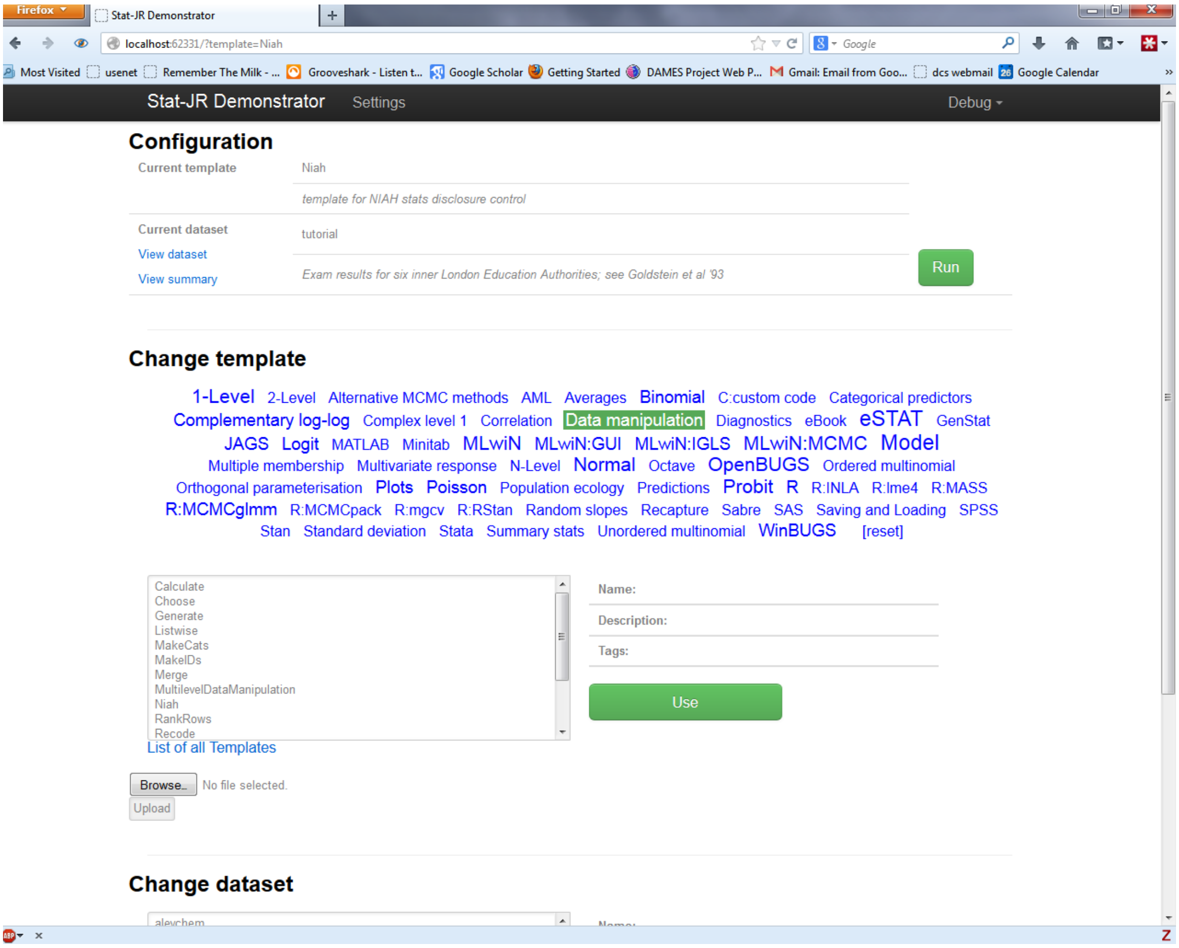
1. **Select the NIAH Template:** The NIAH template is selected from the Data Manipula-

[2] More information about STAT-JR can be obtained from http://www.bristol.ac.uk/cmm/software/statjr/

tion category. It should be noted that the data has already been selected at this stage.

2. **Input NIAH Parameters:** The value of k is set for the assessment, and the key variables are chosen from a list generated from the input data set.
3. **Review Input:** Before the operation is executed STAT-JR displays the input parameters and information about the system.
4. **Review Outputs and Export:** Datasets generated by NIAH can be viewed in the browser window, and files associated with this execution of NIAH can be downloaded.

As the outputs are generated by NIAH within the STAT-JR environment, those outputs can then subsequently be selected and used in the other statistical functions that are available within the environment.



Configuration

Current template: Niah

template for NIAH stats disclosure control

Current dataset: tutorial

View dataset

View summary

Exam results for six inner London Education Authorities; see Goldstein et al '93

Change template

1-Level 2-Level Alternative MCMC methods AML Averages Binomial C:custom code Categorical predictors Complementary log-log Complex level 1 Correlation Data manipulation Diagnostics eBook eSTAT GenStat JAGS Logit MATLAB Minitab MLwiN MLwiN:GUI MLwiN:IGLS MLwiN:MCMC Model Multiple membership Multivariate response N-Level Normal Octave OpenBUGS Ordered multinomial Orthogonal parameterisation Plots Poisson Population ecology Predictions Probit R R:INLA R:lme4 R:MASS R:MCMCglmm R:MCMCpack R:mgcv R:RStan Random slopes Recapture Sabre SAS Saving and Loading SPSS Stan Standard deviation Stata Summary stats Unordered multinomial WinBUGS [reset]

Calculate
Choose
Generate
Listwise
MakeCats
MakeIDs
Merge
MultilevelDataManipulation
Niah
RankRows
Recode

List of all Templates

Name:

Description:

Tags:

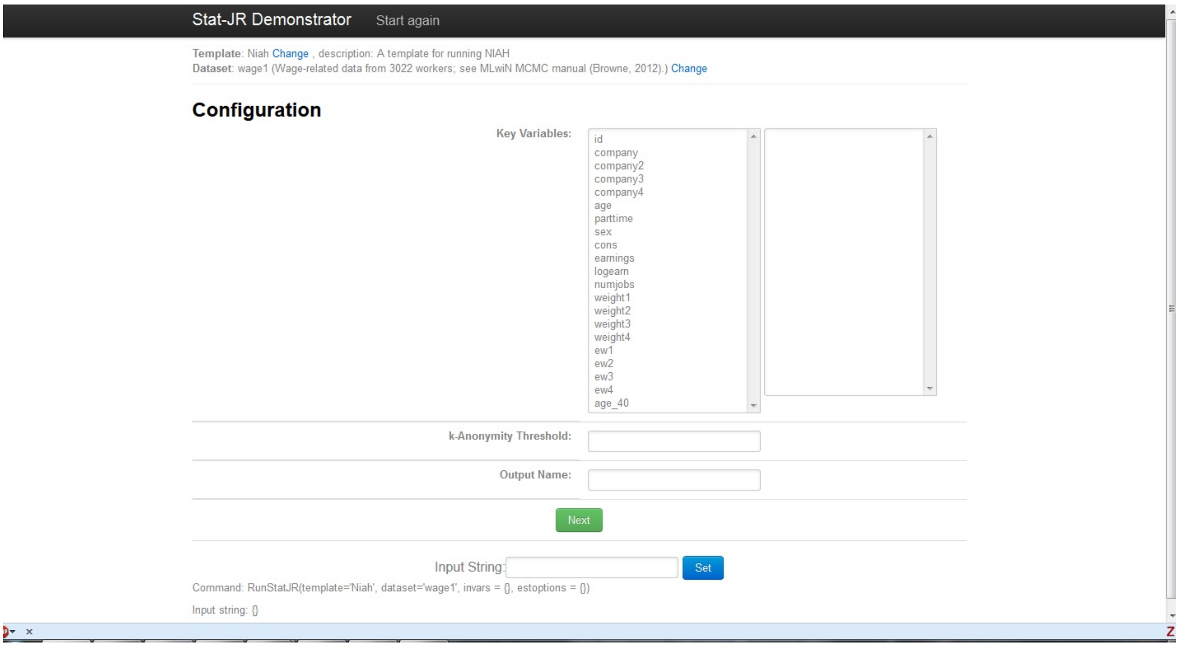
No file selected.

Change dataset

alevchem

Name:

Figure 6.4: STAT-JR: Template Selection



Template: Niah Change , description: A template for running NIAH
Dataset: wage1 (Wage-related data from 3022 workers; see MLwiN MCMC manual (Browne, 2012).) Change

Configuration

Key Variables:

id
company
company2
company3
company4
age
parttime
sex
cons
earnings
logearn
numjobs
weight1
weight2
weight3
weight4
ew1
ew2
ew3
ew4
age_40

k-Anonymity Threshold:

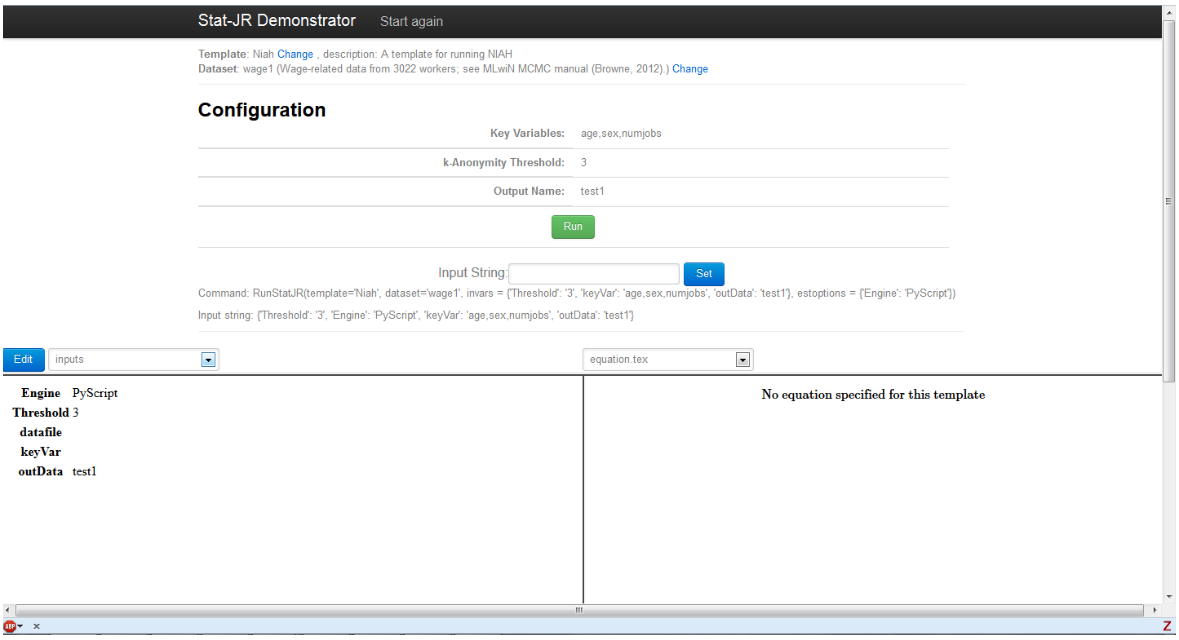
Output Name:

Input String:

```
Command: RunStatJR(template='Niah', dataset='wage1', invars = {}, estoptions = {})

Input string: {}
```

Figure 6.5: STAT-JR: Setting NIAH Parameters



Template: Niah Change , description: A template for running NIAH
Dataset: wage1 (Wage-related data from 3022 workers; see MLwiN MCMC manual (Browne, 2012).) Change

Configuration

Key Variables: age,sex,numjobs
k-Anonymity Threshold: 3
Output Name: test1

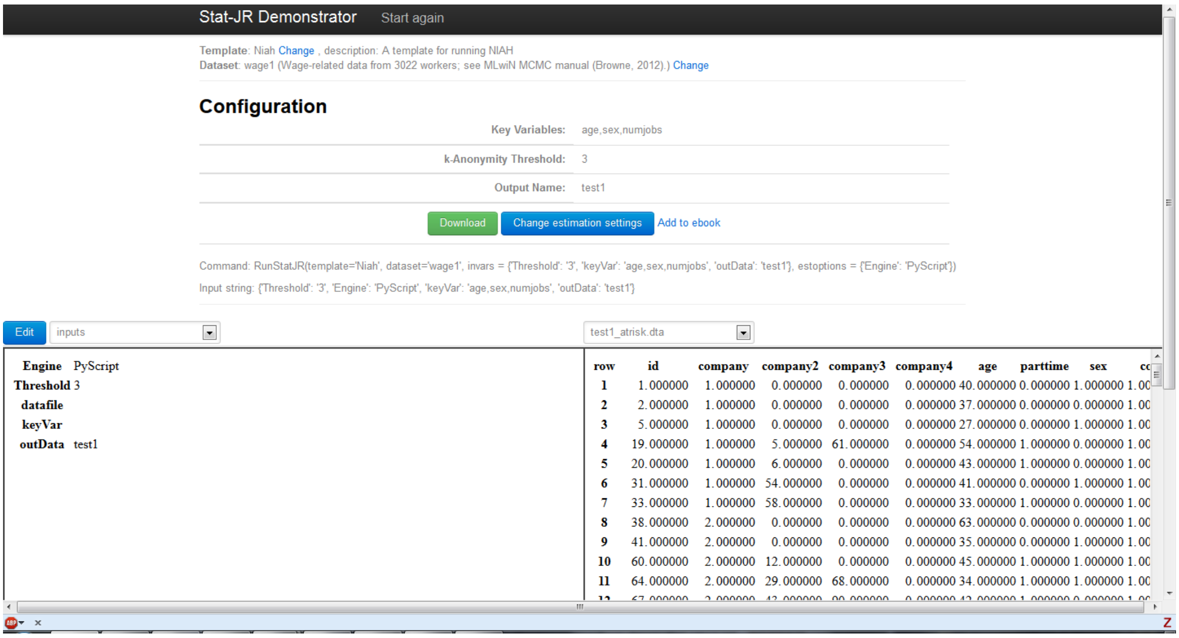
Input String:

```
Command: RunStatJR(template='Niah', dataset='wage1', invars = {'Threshold': '3', 'keyVar': 'age,sex,numjobs', 'outData': 'test1'}, estoptions = {'Engine': 'PyScript'})
Input string: {'Threshold': '3', 'Engine': 'PyScript', 'keyVar': 'age,sex,numjobs', 'outData': 'test1'}
```

Engine PyScript
Threshold 3
datafile
keyVar
outData test1

No equation specified for this template

Figure 6.6: STAT-JR: Review of NIAH Inputs



Template: Niah Change , description: A template for running NIAH
Dataset: wage1 (Wage-related data from 3022 workers; see MLwiN MCMC manual (Browne, 2012).) Change

Configuration

Key Variables: age,sex,numjobs
k-Anonymity Threshold: 3
Output Name: test1

```
Command: RunStatJR(template='Niah', dataset='wage1', invars = {'Threshold': '3', 'keyVar': 'age,sex,numjobs', 'outData': 'test1'}, estoptions = {'Engine': 'PyScript'})
Input string: {'Threshold': '3', 'Engine': 'PyScript', 'keyVar': 'age,sex,numjobs', 'outData': 'test1'}
```

Engine PyScript
Threshold 3
datafile
keyVar
outData test1

row	id	company	company2	company3	company4	age	parttime	sex	cc
1	1.000000	1.000000	0.000000	0.000000	0.000000	40.000000	0.000000	1.000000	1.00
2	2.000000	1.000000	0.000000	0.000000	0.000000	37.000000	0.000000	0.000000	1.00
3	5.000000	1.000000	0.000000	0.000000	0.000000	27.000000	0.000000	1.000000	1.00
4	19.000000	1.000000	5.000000	61.000000	0.000000	54.000000	1.000000	0.000000	1.00
5	20.000000	1.000000	6.000000	0.000000	0.000000	43.000000	1.000000	0.000000	1.00
6	31.000000	1.000000	54.000000	0.000000	0.000000	41.000000	0.000000	1.000000	1.00
7	33.000000	1.000000	58.000000	0.000000	0.000000	33.000000	1.000000	0.000000	1.00
8	38.000000	2.000000	0.000000	0.000000	0.000000	63.000000	0.000000	0.000000	1.00
9	41.000000	2.000000	0.000000	0.000000	0.000000	35.000000	0.000000	1.000000	1.00
10	60.000000	2.000000	12.000000	0.000000	0.000000	45.000000	1.000000	1.000000	1.00
11	64.000000	2.000000	29.000000	68.000000	0.000000	34.000000	1.000000	1.000000	1.00

Figure 6.7: STAT-JR: Review of NIAH Output

VANGUARD

The VANGUARD (Virtual ANonymisation Grid for Unified Access of Remote Data) system is a secure data transfer system specifically designed for use in data-linkage scenarios where data sources are distributed across external locations and organisations (Stell et al., 2009; Sinnott et al., 2008b). It was developed by the National e-Science Centre in the context of e-Health projects that considered the infrastructure requirements for the sharing and linking of electronic health records for research. These projects (for example Sinnott et al. (2008a); Scottish Health Informatics Programme (SHIP)) also identified the need to combine ethical and technological approaches to address issues of data access, which resonates with the theoretical framework set out in Chapter 2 (Sinnott et al., 2006, 5). More specifically, the Virtual Organisations for Trials and Epidemiological Studies (VOTES) project which sought to streamline the clinical trials process across a number of participating partner institutions, added a substantial caveat that their grid computing based system would be subject to "strict ethical, data protection and security constraints" (Sinnott et al., 2006, 2). However, these constraints were only included in the design as a signpost to human processes within data-owning organisations. What is set out below is a proposal to integrate the NIAH tool set within the VANGUARD system to provide an operationalisation of these constraints to further streamline the research process.

At the conceptual level of VANGUARD's design are the following components (Stell et al., 2009, 332):

1. **Viewers** are used to access potentially remote datasets (typically this is associated with a specific clinical research study that has been approved by an independent ethics body).

2. **Guardians** protect the data resources being provided to the virtual organisation.

3. **Agents** mediate the exchange between the guardian and the viewer.

4. **Bankers** maintain a record of all transactions that have taken place and limit resource data exchanges based on accountability information.

In the context of this research, the Guardian component is of particular interest. Stell *et al* define the Guardian as the protector of the data sources. There could also be multiple Guardians for each discrete data source. The specific tasks of the Guardian are defined as: "The Guardians periodically check for requests (queries) they should respond to; the Guardian pulls requests that they should act upon; the Guardian decrypts and makes a locally defined authorization decision on the query, and when satisfied that it meets all local policy criteria on data access and usage, executes the query" (Stell et al., 2009, 333). Current practice would indicate that the Guardian is akin to the data owner role more familiar in the context of this research. Therefore, they would need to perform their own queries to ascertain if the required policy criteria have been satisfied. As has been argued elsewhere, existing approaches to this role are potential sites for bottlenecks in the data access and analysis workflow.

If NIAH were integrated with the VANGUARD system, a Guardian could use the tools set out earlier in this chapter to facilitate their decisions. Further to this, if their policy criteria could be framed and defined in he context of parameters to match the input requirements of NIAH then a degree of analytical automation can be established to make the Guardian workflows more efficient. A high-level comparison, using the methods at the disposal of a Guardian of NHS records in Scotland versus the same Guardian using an integrated NIAH tool set will be considered. Using the VANGUARD system within a common existing workflow, a query is created by researchers to link and analyse two datasets (e.g. The Scottish Record of Morbidity and data from the GPASS system used by general practitioners). That query is received by the Agent, which sends the relevant part of the query to the Guardians for each specific data set. The Guardian must periodically check for queries from the Agent to approve. When a Guardian becomes aware of a query, time and resources must be allocated for testing the query against their policy criteria. If the model of *ad hoc* scripts is used then the Guardian is required to provide a significant level of input (adjusting code to meet the relevant dataset format and variables). While this process is being carried out the researcher is waiting for data to progress their research. To use a software engineering analogy, to some extent the research project is ’busy-waiting’ until the Guardian releases the ’lock’ on the query.

It should be remembered that this research does not hypothesise that a fully automated ‘Guardian’ role could or should be implemented. However, if NIAH and the associated tool set were integrated into the VANGUARD workflow as set out above, some of the wasted time between query submission and approval could be reclaimed. This can be achieved by modifying the data owner’s workflow to include the following: a data owner should have ostensibly developed their policy on data access and statistical disclosure control following good practice as the literature sets out in chapter 4. In doing so they have knowledge of their data environment, their data’s characteristics and the views of relevant stakeholders. These policy considerations can be expressed in terms of values of k and a series of potential quasi-identifiers which can be used as input parameters for NIAH. This then allows for the semi-automation of the Guardian process: a query is created by researchers to link and analyse two datasets; that query is received by the Agent, which sends the relevant part of the query to the Guardians for each data set to be linked; the Guardian’s system could then carry out a range of k-anonymity assessments and implement different SDC algorithms; the Guardian must periodically review the results of these assessments to carry out any further SDC work.

This modified workflow encompasses a number of benefits over existing approaches. The systematic approach to creating policy, which is then applied in a consistent and replicable way, ensures consistency within a data owning organisation. This has the added advantage that this process can be well documented and creates additional metadata that the data users can review to encourage trust in the data owners approach. The semi-automatic application of SDC methods could remove the first pass of data analysis from the Guardian's workload allowing them to concentrate on reviewing the results of these methods and making an informed decision on data access and linkage. This also increases the volume of queries that a Guardian could process over time. The added logging, documentation and ability to replicate analytical work that the NIAH tool set could provide, allows the Guardian to demonstrate the processes that have been carried out if a data user were to appeal against a denial of access and this documentation could form the basis for a discussion with the data users about their requirements and what revisions might need to be made to their original query. For example, perhaps requesting fewer variables, or requesting a variable at a more coarse level of detail.

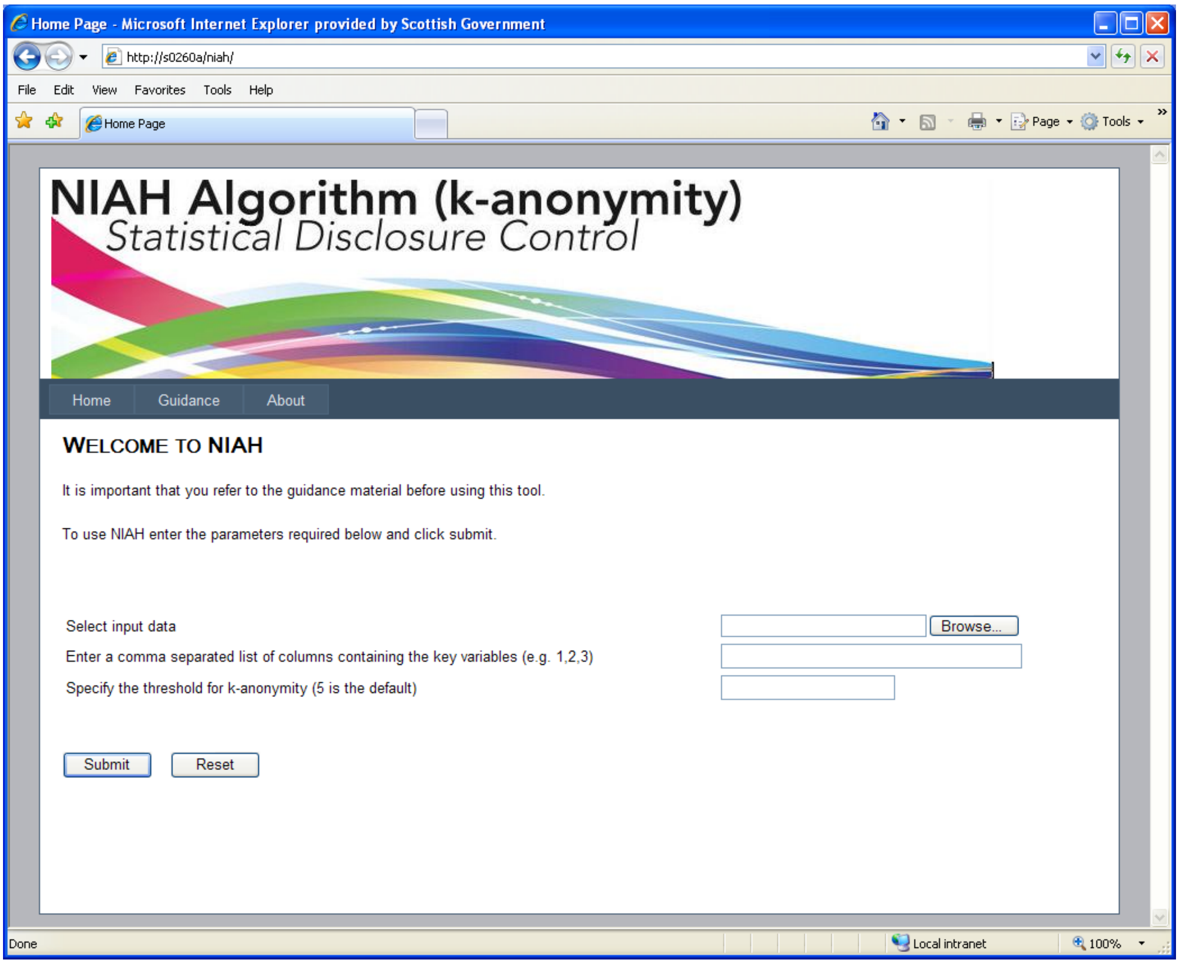
Chapter 7

Case Study: Scottish Government Deployment and Data Analysis

This Chapter presents a case study of the implementation of the tools and methods discussed in the preceding chapters in the context of the Scottish Government's analytical services. The first section of the Chapter discusses the Scottish Government's own deployment of NIAH as a web application. The Chapter then continues with a demonstration of NIAH in use in a secure setting at the Scottish Government. This demonstration provides a real-world analysis of linked education data, and data regarding children looked after by the state, in which the NIAH toolset is used to assess the potential risks and implement disclosure control measures before the resulting data utility is assessed.

7.1 NIAH in use at the Scottish Government

Having shown how the NIAH tool set can be integrated into existing e-Science infrastructure, one last example is presented to show how organisations have adapted NIAH to fit with their own infrastructure and workflows. The Scottish Government through Health Analytical Services and the Scottish Exchange of Data Unit (ScotXed) have implemented a system for using NIAH to carry out k-anonymity assessments on datasets they hold before publishing data or distributing data to third parties. They use a web interface (see figures 7.1 & 7.2) which communicates with NIAH installed on a centralised server. This allows analysts to use the NIAH tool set without the limitations of their local machines resources and offers the Scottish Government the ability to store logs of NIAH activity centrally so this activity is resilient to changes in personnel. This model of implementation is also suitable for a wider roll-out to local authorities. As identified in the discussion of the data users' perspectives, a problem for both local and national governments is a lack of consistency in analytical expertise, combined with the use of different software. Respondents highlighted the vastly different resources available to Local Authorities for statistical disclosure control and data publishing processes. Use of the NIAH tool set, as part of their data management workflow, could ensure at least some consistency in approach between institutions.



Home Page - Microsoft Internet Explorer provided by Scottish Government

http://s0260a/niah/

NIAH Algorithm (k-anonymity)

Statistical Disclosure Control

WELCOME TO NIAH

It is important that you refer to the guidance material before using this tool.

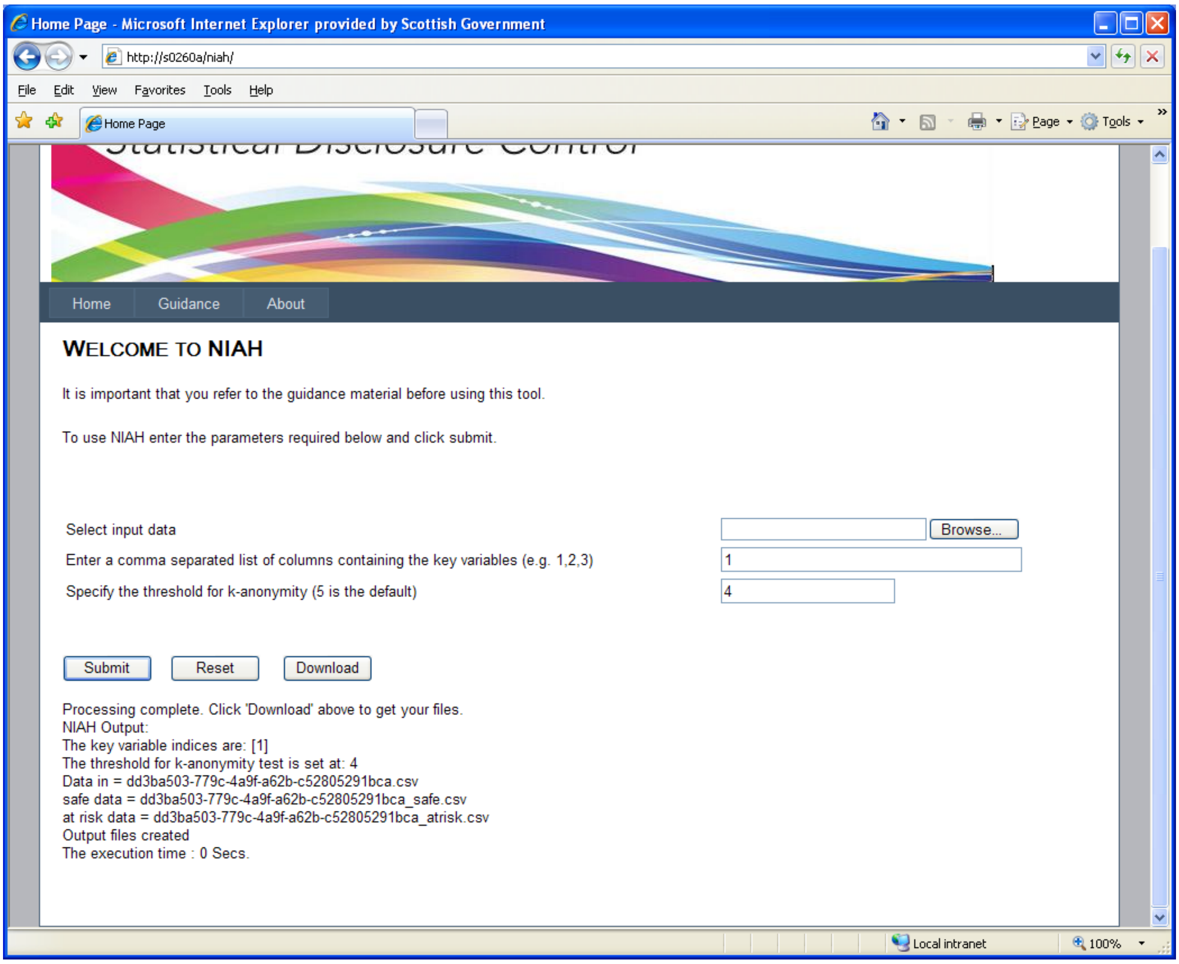
To use NIAH enter the parameters required below and click submit.

Select input data

Enter a comma separated list of columns containing the key variables (e.g. 1,2,3)

Specify the threshold for k-anonymity (5 is the default)

Figure 7.1: The Scottish Government’s landing page for their implementation of NIAH


Home Page - Microsoft Internet Explorer provided by Scottish Government
http://s0260a/niah/
File Edit View Favorites Tools Help
Home Page
Page Tools
Statistical Disclosure Control
Home Guidance About
Local intranet
100%

WELCOME TO NIAH

It is important that you refer to the guidance material before using this tool.

To use NIAH enter the parameters required below and click submit.

Select input data [ ] Browse...

Enter a comma separated list of columns containing the key variables (e.g. 1,2,3) 1

Specify the threshold for k-anonymity (5 is the default) 4

```
Processing complete. Click 'Download' above to get your files.
NIAH Output:
The key variable indices are: [1]
The threshold for k-anonymity test is set at: 4
Data in = dd3ba503-779c-4a9f-a62b-c52805291bca.csv
safe data = dd3ba503-779c-4a9f-a62b-c52805291bca_safe.csv
at risk data = dd3ba503-779c-4a9f-a62b-c52805291bca_atrisk.csv
Output files created
The execution time : 0 Secs.
```

Figure 7.2: The Scottish Government's output page for their implementation of NIAH

7.2 A Demonstration of NIAH Within a Secure Environment: Scottish Government Education Data

Having discussed the development of the NIAH tool-set earlier in Chapters 3 and 6, the focus now moves to a case study using NIAH in a real-world setting as part of a statistical disclosure analysis workflow. The subject of the case study are data from the Scottish Government, drawn from administrative sources about 'looked after children'. Looked after children in this context are defined by the Children (Scotland) Act 1995[1]:

> 'Looked After Children' are defined as those in the care of their local authority. The majority will come into one of these categories:
>
> Looked after at home:

[1] These details are taken from the Scottish Government, see http://www.Scotland.gov.uk/Topics/People/Young-People/protecting/lac/about Accessed:03/07/2013

> Where the child (or young person) has been through the Children's Hearings system and is subject to a Supervision Requirement (regular contact with social services) with no condition of residence. The child then continues to live in their regular place of residence (i.e. the family home).
>
> Looked after away from home:
>
> Where the child (or young person) has either: been through the Children's Hearings system and is subject to a Supervision Requirement with a condition of residence; is subject to an order made or authorisation or warrant granted by virtue of Chapter 2, 3 or 4 of Part II of the 95 Act; is being provided with accommodation under Section 25 (a voluntary agreement); or is placed by a local authority which has made a permanence order under Section 80 of the Adoption and Children Act 2007. In these cases the child is cared for away from their normal place of residence, by foster or kinship carers, prospective adopters, in residential care homes, residential schools or secure units.

The data are recorded at the local authority level by social work departments and then collated centrally to form an annual Looked After Children Survey by the Scottish Government. Publications are then generated centrally and these include aggregate statistics, for example see Scottish Government (2013) and Scottish Government (2011d). In keeping with the Scottish Government's position on opening-up administrative data and offering data linkage opportunities across sectors, data such as the Looked After Children Survey could be a candidate for release in some sanitised form. For example, this data combined with other education data sources such as the Pupil Census or Scottish Qualifications Authority (SQA) data could prove to be an important resource for education researchers, local authorities, and central government. However, this data provides a realistic scenario for the work described in this dissertation because the subject matter is particularly sensitive in that it concerns children, and in particular vulnerable children looked after by the state.

Before proceeding to describe the dataset further and beginning the analysis, it is worth noting the technical and administrative configuration used in this instance, as it demonstrates a possible model for the integration of NIAH with existing work-flows in an applied setting. The data are held centrally on servers controlled by the Scottish Exchange of Education Data (ScotXed) unit which is part of the Education Analytical Services Division within the Learning and Justice Directorate of the Scottish Government. The data are stored in a Microsoft SQL Server database and were exported as comma separated values for the NIAH processes. NIAH was installed on the central ScotXed server with the following specification: Dell PowerEdge R710 with 8 EM64T 2.9MHz Intel Processors, 32GB RAM, running Microsoft Windows Server 2003 and Java version 1.6 64bit server edition. The server was accessed by the researcher via a remote desktop connection and NIAH was operated from the command line interface. Any post-processing of output data was done via the researcher's local machine connected to the Scottish Government's central SAS statistics server. All data, outputs and analysis were contained within the Scottish Government's systems and final outputs were audited by the Education Analytical Services Division. Figure 7.3 shows a graphical representation of this configuration.

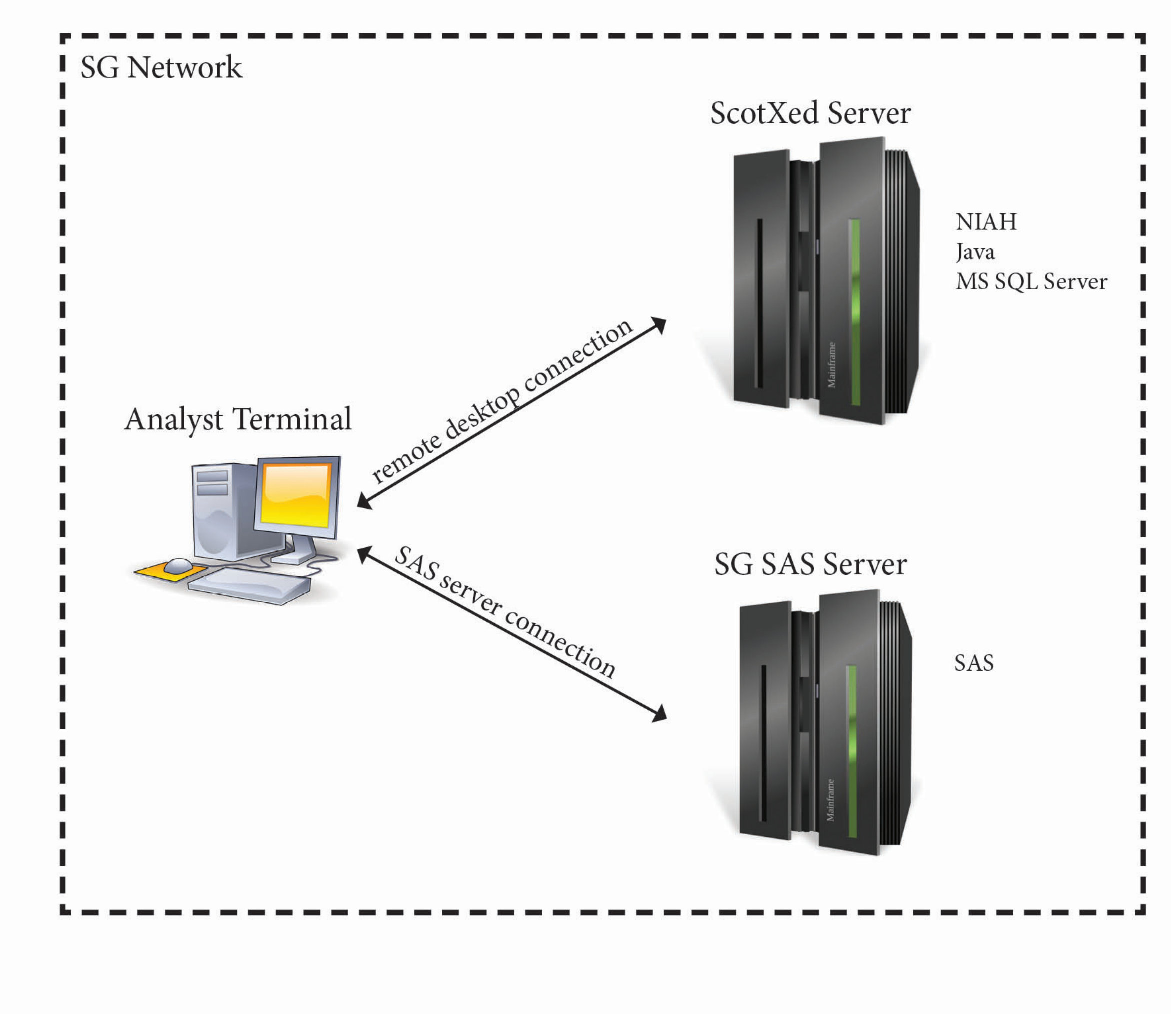


Figure 7.3: The Technical Configuration for Analysis while Analysing the Education Dataset at the Scottish Government

The dataset generated from administrative sources for the case study contained 8, 185 records from the 2010 Looked After Children Survey and contained a range of demographic information combined with data on education institutions, student attainment, and looked after placement details. The subset of variables that either form part of the quasi-identifier key variable analysis or are highlighted as sensitive variables that need to be protected will be the main focus.[2] Following the mixed qualitative/quantitative methodology detailed in chap-

[2] For more details on the looked after children data collected by the Scottish Government, including data specifications and collection methods; see the ScotXed website www.ScotXed.net.

ter 4.2, 2 possible intruder scenarios for the data were considered; this approach is based on the work of Elliot and Dale (1999). Differing from Elliot and Dale's work in the level of detail, because of the available variables and metadata, two major groups of possible attackers are defined — those known directly by the target and those that are not. These are labelled 'relatives' and 'strangers' and these were discussed informally with government education analysts to assure their relative validity. The two intruder scenario groups are:

1. **Relatives**

   Our first intruder scenario was based around a relative of a particular child attempting to re-identify the child's record to discover some details about their placement in care. To this intruder scenario the following potential key variables are attributed that could be mapped to the data: date of birth, gender, ethnic group, national identity, main disability and datazone (low level geography).

2. **Strangers**

   A stranger could be seeking information for a number of reasons, including journalism, activism or political motivations. In legal cases involving young people for example, efforts are made to preserve their anonymity and therefore disclosure from anonymised sources could be an avenue of attack. To this intruder group the variables: gender, age (specifically not date of birth), ethnic group (potentially at an abstracted level of detail), and local authority are attributed.

From these two intruder scenarios the key variables are drawn out and presented in table 7.1; Scottish Parliamentary Constituency has also been included so that some comparative work can be conducted across the three geographical variables present in the dataset. Student stage and an urban/rural classification are also included to discuss the problems of proxy indicators. Table 7.2 is a list of the potentially sensitive variables present in the subset of the data. These represent information about the data subjects that might be considered especially private. With only a cursory examination, variables that might indicate an individual's socio-economic status, their educational attainment and their school type (specifically with reference to those that attend a special school) have been chosen. Each key variable will be described and some univariate statistics will be provided to illustrate how they are positioned within the dataset as a whole.

Geographical Variables

The geographical variables give us an indication of where risk might be concentrated. It should be noted that it would be unusual to release data with geographic areas smaller than local authority for this type of data, however the presence of 3 geographic variables gives us

Variable	Description
Lacode	Local Authority Code (3 digit)
Lacdob	Date of Birth (YYYY-MM-DD)
Lacgender	Gender (M/F)
Lacethnicgroup	Ethnic Group (17 categories - 2 digit)
Nationalidentity	National Identity (9 categories - 2 digit)
Datazone	Datazone (small area geography)
SParlCon	Scottish Parliamentary Constituency (2 digit)
MainDisability	Main Disability (2 digit)
StudentStage	Student Stage e.g. Primary 1 (coded as 2 characters)

Table 7.1: Education Data: Key Variables & Descriptions

Variable	Description
FreeSchoolMealRegistered	Whether a student is registered for free school meals
StudentLookedAfter	Whether the student is looked after by the local authority either at home or out of the home
SIMD	Scottish Index of Multiple Deprivation
Attainment Variables(Multiple)	Various variables detailing a students level & quantity of qualifications
SchoolType	Type of School (Primary, Secondary, Special)

Table 7.2: Education Data: Potentially Sensitive Variables

the ability to see what the effect of different levels of geography have on disclosure risk in this particular case.

Local Authority (LA)

Figure 7.4 provides a breakdown of the number of records by local authority. From these initial descriptive statistics the number of records with a given value is considered as a crude indication of disclosure risk. For example, consider LA 330 (Orkney) with only 23 cases; the potential for re-identification given knowledge of the LA alone is greater than for LA 260 (Glasgow City) that carries 1604 cases.

Scottish Parliamentary Constituency

From Figure 7.5 it is shown that the geographical partitioning on constituency lines has spread the distribution of cases across a larger number of geographies. However constituency 67 (Shetland) has the lowest number of cases but at 16 this does not look very different from the 23 cases for Orkney at the LA level. It should be noted that different geographical

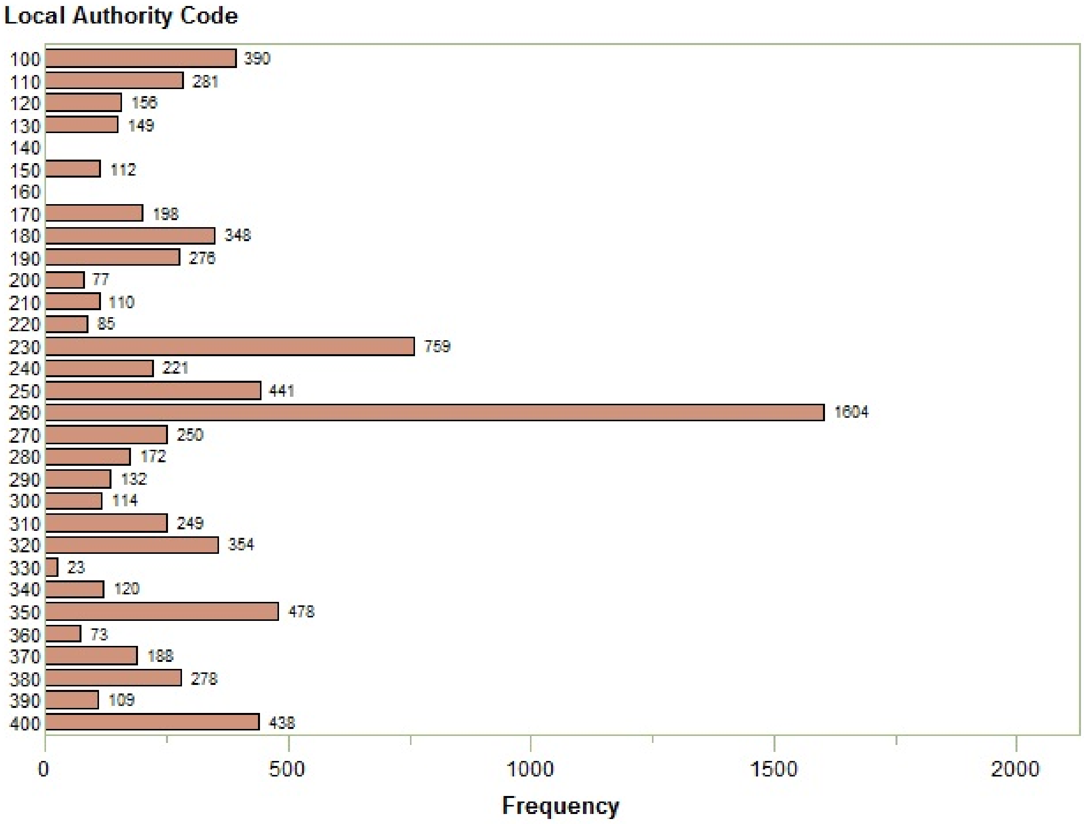


Figure 7.4: Education Data: Frequency Distribution of Local Authority (N=8185)

partitioning with a similar number of areas can have completely different risk profiles. With this comparison of LA and Parliamentary Constituency it can be shown that the counter-intuitive effects of geographic detail on disclosure risk highlighted by Elliot et al. (1998) and Witkowski (2008) are partially in affect. This literature focuses on the effects of different scales of geography for which the disclosure risk does not scale similarly. For example, a group of records with particularly unique characteristics present at the datazone level might also be particularly unique at local authority level despite the aggregation of geography.

In addition, if multiple geographical variables were released, or variables that could act as a proxy for geography were included, for example the Scottish Index of Multiple Deprivation (SIMD) or Urban Rural indicators, then particular attention should be paid during any disclosure risk assessment in case records can be better located by an intruder combining their knowledge of geographic details, see Steel and Sperling (2001).

As a small example of the proxy geography problem consider an analysis of the data that includes LA codes and the commonly used 8 fold Urban-Rural indicator. Using the NIAH tool-set's k-anonymity algorithm, 24 records were identified with unique combinations of LA and Urban-Rural Indicator. To achieve this NIAH was executed with the key variables local authority and the urban/rural indicator using a threshold of 2. With this knowledge,

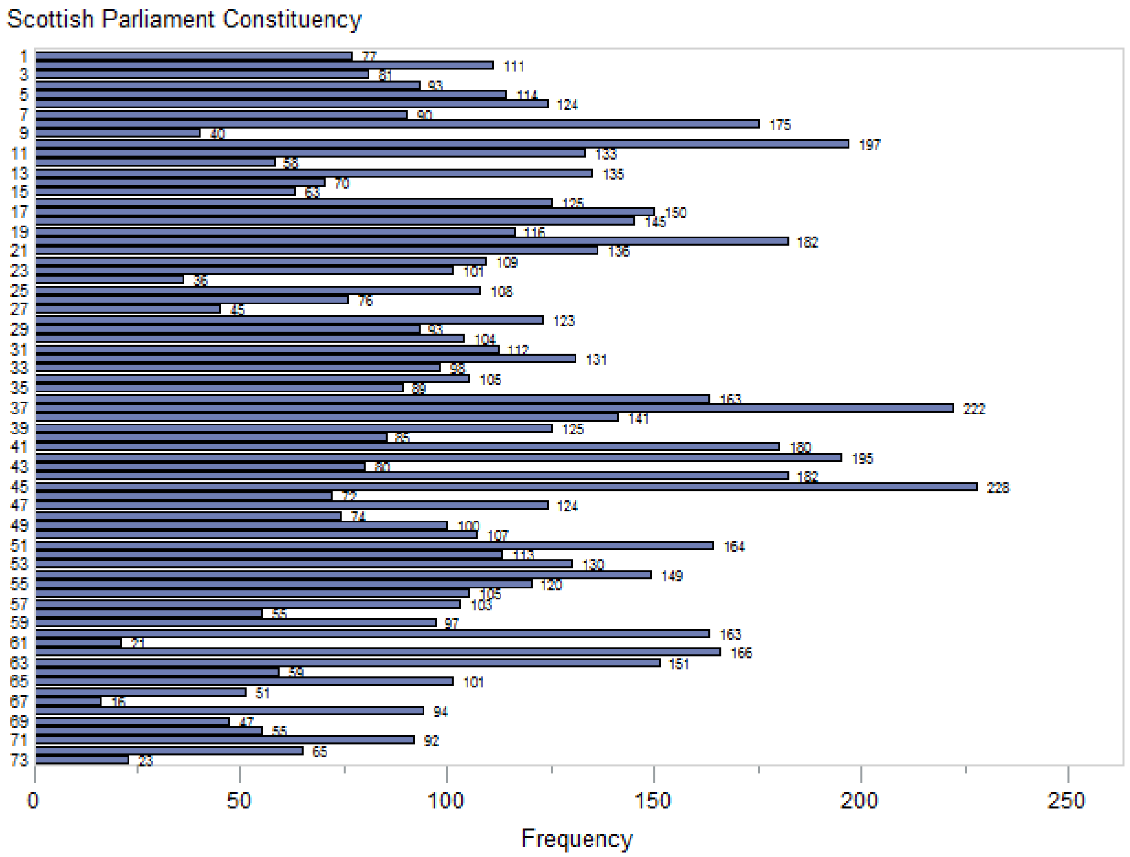


Figure 7.5: Education Data: Frequency Distribution of Scottish Parliamentary Constituency (N=8185)

an intruder could use publicly available maps of the urban rural classifications published by the Scottish Government in order to significantly narrow the geographical area attached to a particular record. To better illustrate this point, Figure 7.6 is a subsection of a map of Scotland colour coded by urban/rural indicator. Consider the area circled in Figure 7.6, this area in the Highland Local Authority north of Inverness is made up of 'Very Remote Small Towns' in the 8 fold urban rural classification and forms a small part of the Highland area as a whole. If an intruder had access to this knowledge beyond simply the local authority it would significantly narrow down the geographical location of the record.

Datazones

The datazones are based on census output areas and form the key small area statistical geography in Scotland. There are over six thousand areas and they have populations of between 500 and $1,000$ individuals. As such they carry a much greater potential disclosure risk than the LA and constituency areas. Table 7.3 gives us an insight into this risk level by showing the relative uniqueness of records in the dataset that belong to a particular datazone. As we can see 60% of records belong to a datazone with less than 5 records associated with it

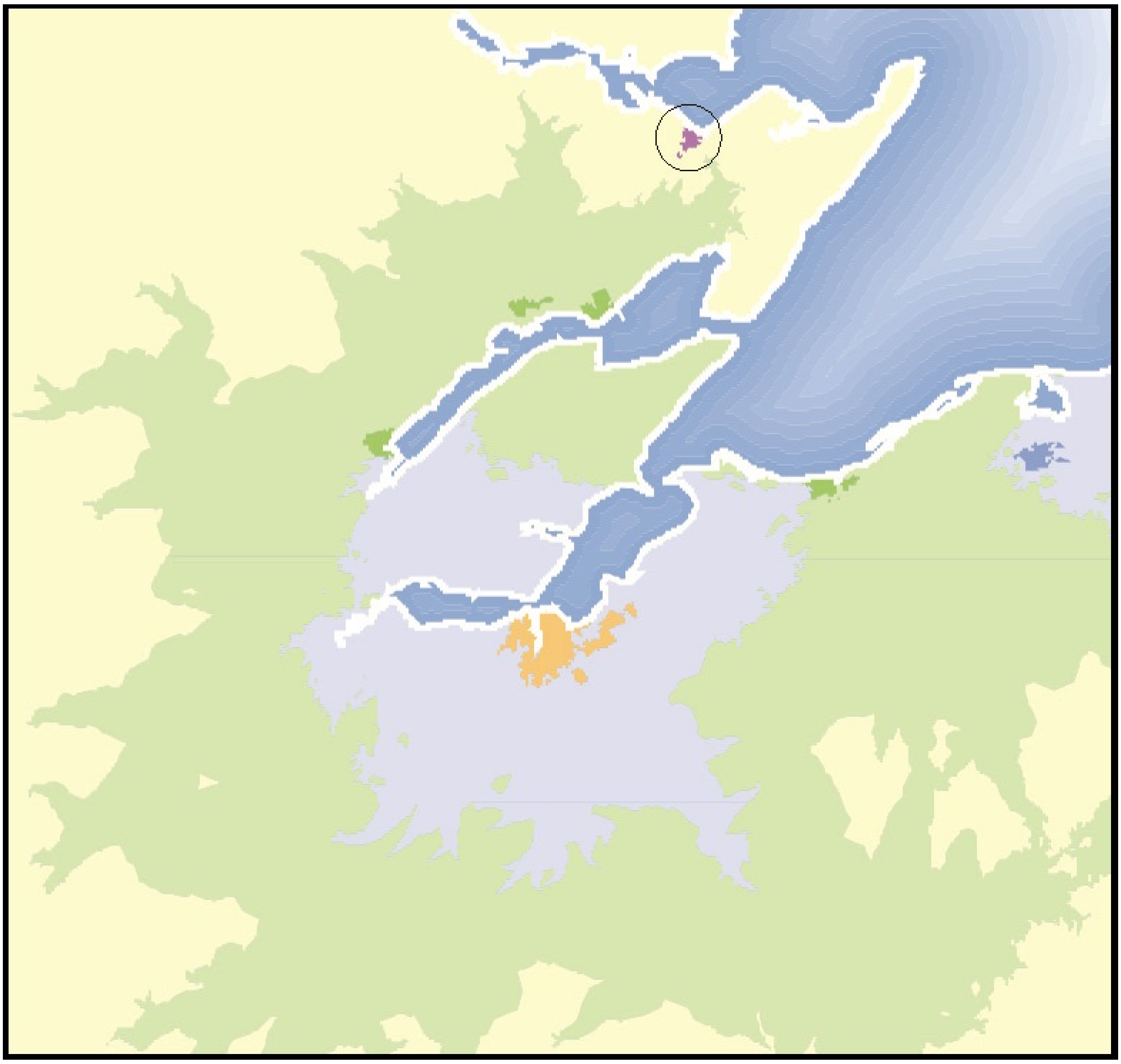


Figure 7.6: Education Data: A Section of the 8 fold Urban-Rural Classification Map

and over 50% of datazones have less than 3 records associated with them. The implication of these percentages is that the ability of an intruder to narrow down the pool of possible correct matches is greatly enhanced if the data contain datazone as a geographic variable, because this increases their chance of finding unique records that could match their *a priori* knowledge.

Demographic Variables

Age (Date of Birth)

Our dataset contains the full date of birth, however it is unlikely this would ever be released without some perturbation because including a full date of birth further increases the chances of uniques in the data, see the treatment of date of birth in Iyengar (2002) and Sweeney (2001) for example. It is of little analytical value, apart from in studies of attain-

	Datazones with only 1 record	Datazones with less than 3 records	Datazones with less than 5 records	Datazones with less than 10 records
% of records (N=8185)	14%	32%	60%	92%
% of datazones (N=2988)	38%	62%	85%	98%

Table 7.3: Education Data: Summary of the Number of Records per Datazone

ment in which where a student's birthday falls in the year is important such as Angrist and Krueger (1992). This highlights, again, the qualitative elements of the analysis, as argued by microdata projects such as the Integrated Public Use Microdata Series (IPUMS) at the University of Minnesota: data utility is highly user specific and can only be determined accurately through dialogue with users. To cover the general use case, an age in years variable is derived for the analysis and this is used below. To clarify here, age in this dataset is the age derived from the date of birth for end of the Looked After Children collection period (31st July). Age is a common key variable across a number of datasets and it is relatively easy to obtain from public sources. Figure 7.7 shows the distribution of the age variable, given that this is education data concerning young people the data is fairly homogeneous. However, immediate areas of concern, with regard to statistical disclosure, are concentrated in the upper end of the distribution with both ages 20 and 21 having counts of less than 100 records. At this stage it can only be speculated upon as to how this potential disclosiveness can be addressed and these records might be candidates for top-coding or suppression, as seen in the Sample of Anonymised Records (Willenborg and De Waal, 1996, 42). This method is suggested as other forms of non-perturbative disclosure control have little effect for outlying values with few records associated with them.

Gender

Gender itself does not normally present a disclosure risk, but it can be a mitigating factor when cross-tabulations with other key variables are taken into account. As with age, gender is a fairly easy variable for an intruder to know *a priori*. Table 7.4 provides the breakdown of gender in the data. All factors being equal it might be expected that the gender balance would match that of the population[3] however, there is a slightly higher number of males than females looked after by the state. To demonstrate the potential problem of combining the gender key variables with others, Figure 7.8 shows Gender in combination with age for the

[3] the 2011 Scottish Census shows a split of 48% male and 52% female — http://www.scotlandscensus.gov.uk/en/censusresults/bulletin.html.

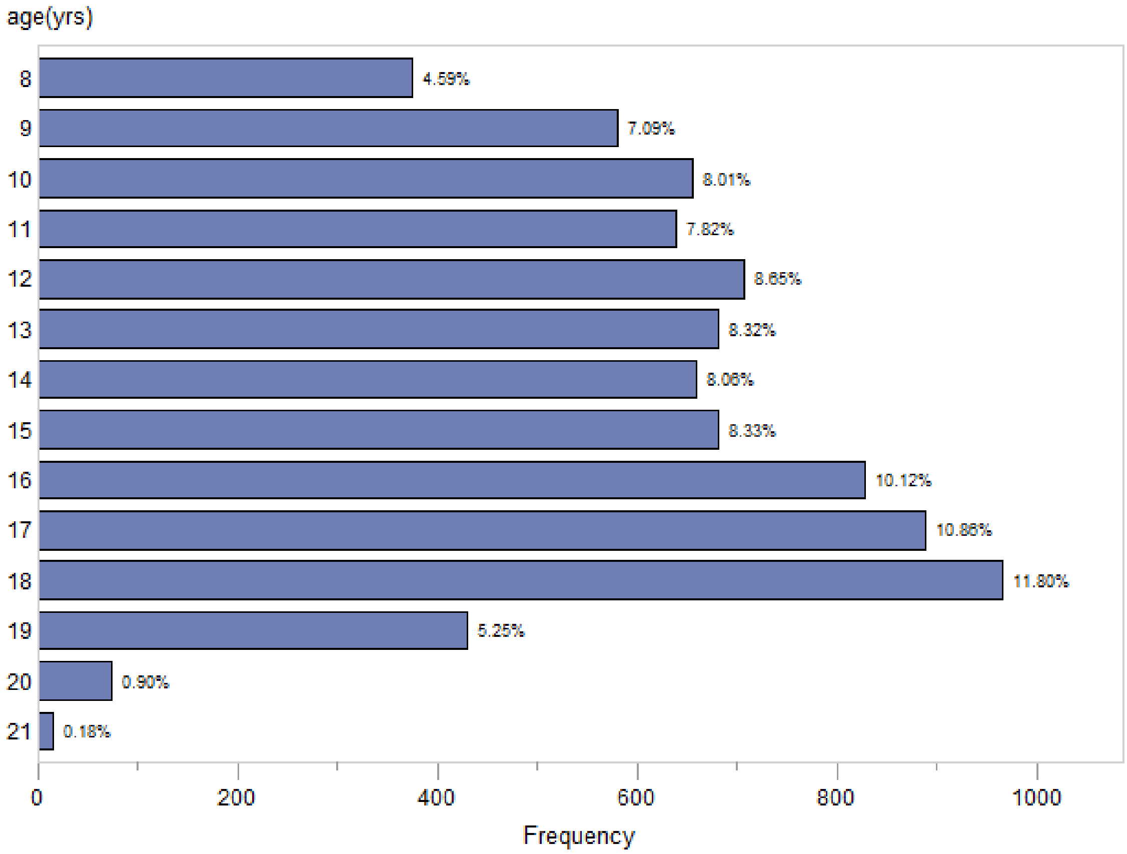


Figure 7.7: Education Data: Age Distribution (N=8185)

N=8185	Female	Male
% of Records	46%	54%

Table 7.4: Education Data: Gender Distribution

City of Edinburgh Local Authority. The population distribution of gender is largely reflected in figure 7.8. In terms of a potential intrusion scenario, knowledge of a target's gender can potentially half the ambiguity associated with a match. For example, from figure 7.8 we know that there are 72 individuals aged 14, if it is known that an intruder is targeting a female, the target group is reduced to 38.

Ethnic Group

Ethnicity data presents a particular problem for disclosure control in Scotland especially, because ethnicity in Scotland is relatively homogeneous, with the majority of the population falling into a 'white' category with few outliers. This data is often removed from research datasets because the disclosure risk outweighs the analytical utility unless the study has a

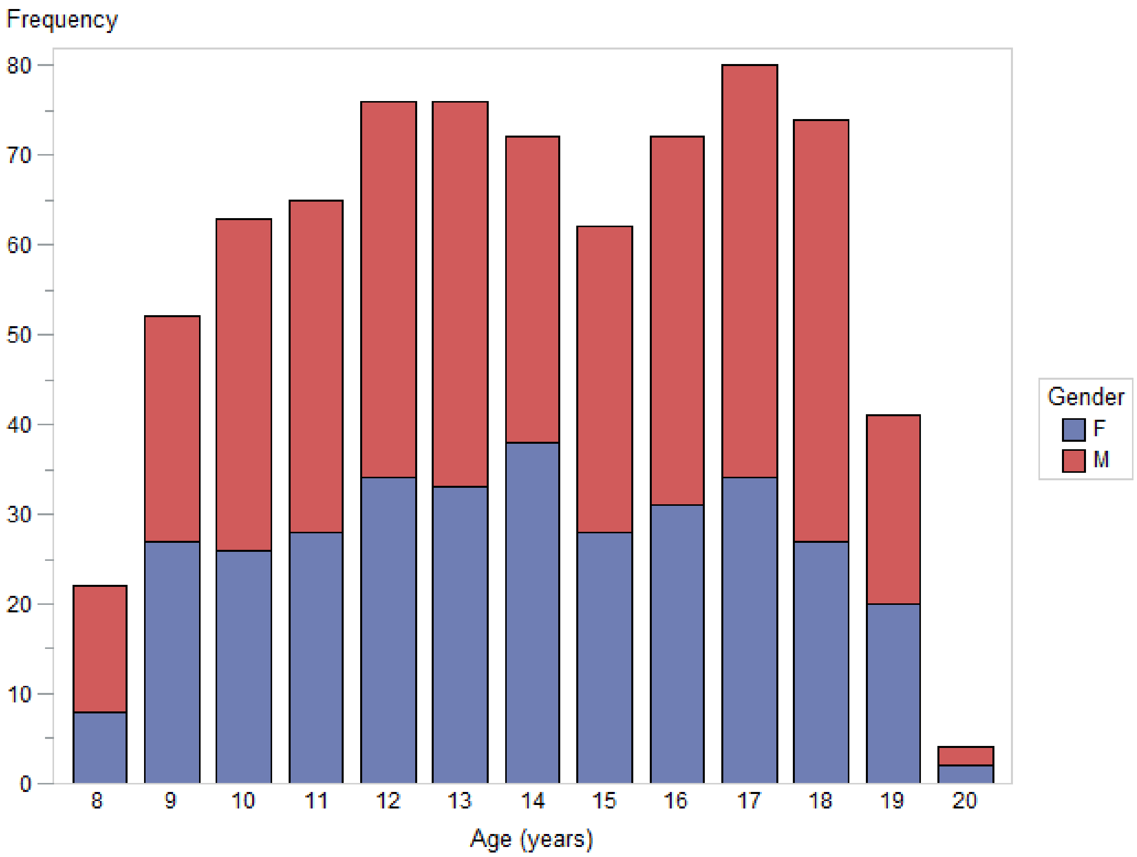


Figure 7.8: Education Data: Gender Distribution by Age for the City of Edinburgh Local Authority (N=759)

particular focus on ethnicity. This lack of data sources has been highlighted, especially in the health field (Bhopal et al., 2011; Ranganathan and Bhopal, 2006). This homogeneity is very apparent in Figure 7.9. However, it is possible to counter this argument about homogeneity The categorisation of ethnicity into a large number of very specific groups, which are subjectively assigned potentially introduces a significant amount of ambiguity for an intruder to navigate. Even if the level of detail is reduced to ‘white’ and ‘non-white’ an intruder cannot be certain to which category their target might be assigned. This also resonates with issues of operationalising ethnicity data in analysis more generally, see Lambert (2010).

National Identity

National identity also presents a similar challenge to ethnic group because of the relative homogeneity of the data. However, the ability for an adversary to know national identity *a priori* is less obvious than for ethnic group because of the greater subjectivity in the assigned value. Figure 7.10 provides a breakdown of national identity.

It is again worth noting that the suppression of a variable such as ethnic group or national identity is not a universal rule. A study interested in the attainment of children from different

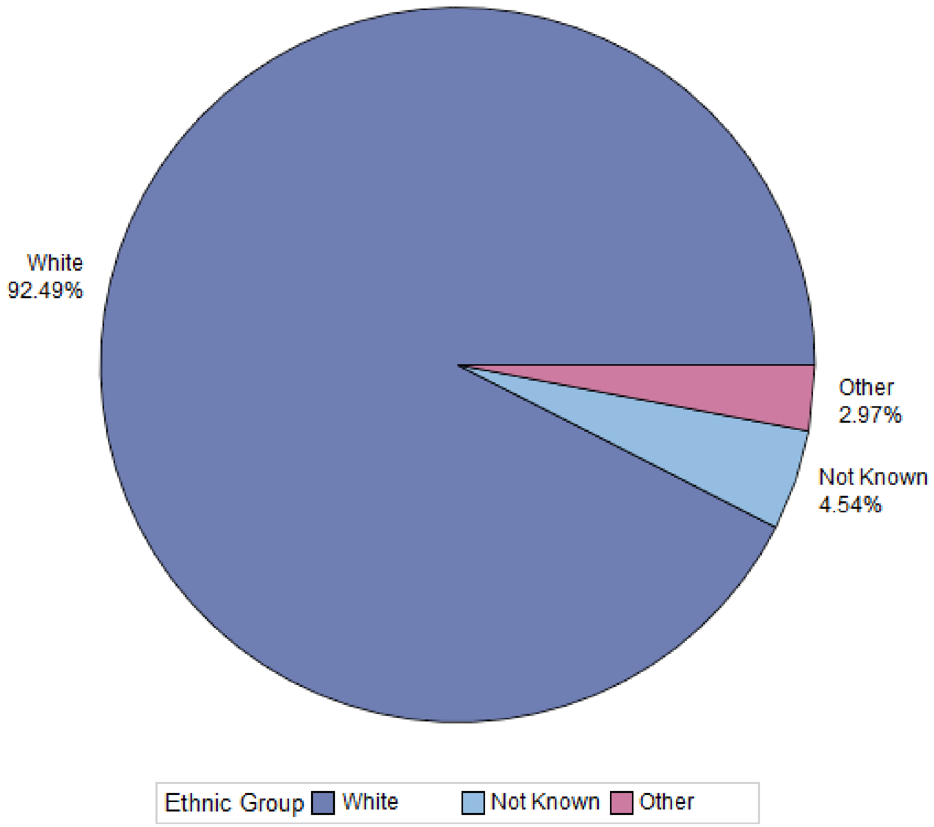


Figure 7.9: Education Data: Ethic Group Distribution (N=8185)

ethnic groups or national identities might not be interested in gender or age for example. Therefore, the level of detail required for effective analysis could be negotiated for specific research projects. Throughout this case study, the focus has been the more general public use case and the typical variables such as age and gender which are common to the majority of analyses, but this need for flexibility should not be forgotten.

Main Disability

Main disability is perhaps not a commonly analysed variable, however given that this is an exploratory analysis it is considered for its disclosive potential. The scenario here is that a target for re-identification has a disability known to the adversary that could aid their attack. Figure 7.11 displays the distribution of main disability categories. The majority of individuals have no disability and besides social, emotional and behavioural difficulties the other disability categories are too small to be represented here. These include visual and hearing impairments, physical disabilities and learning disabilities. These form 'non-obvious' categories and it is unclear that they would be significant in an intruder's attempt to disclose sensitive data or re-identify an individual, however consider the scenario represented in Figure 7.12, in which the intruder knows the age, local authority and that the individual

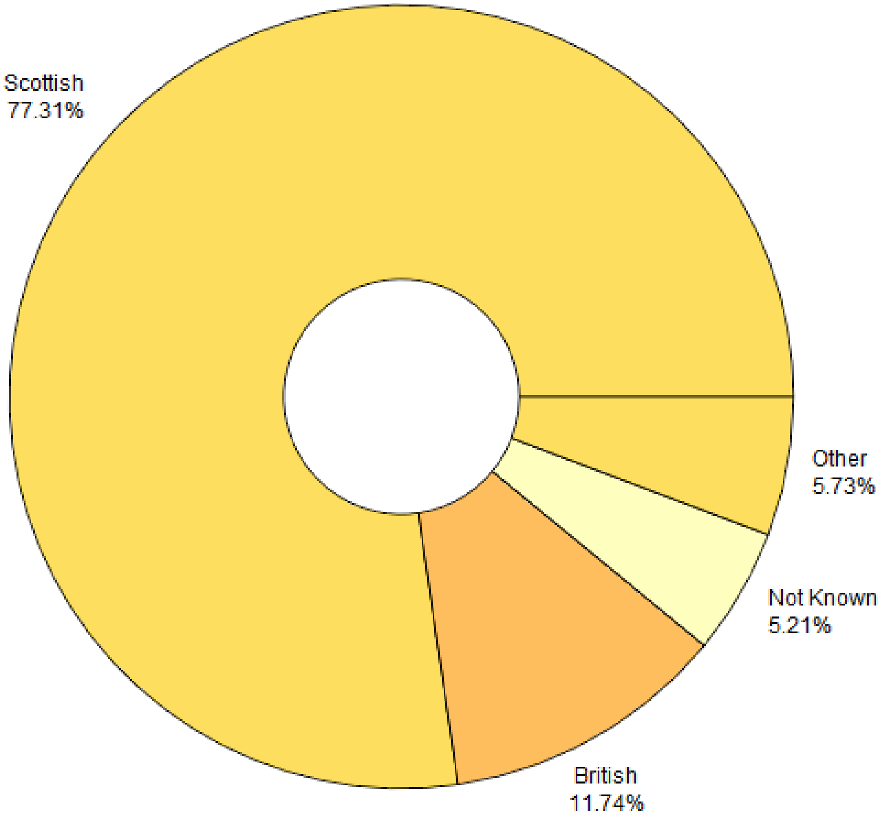


Figure 7.10: Education Data: National Identity Distribution (N=8185)

has a physical disability. The ambiguity is reduced significantly, in some cases down to one record.

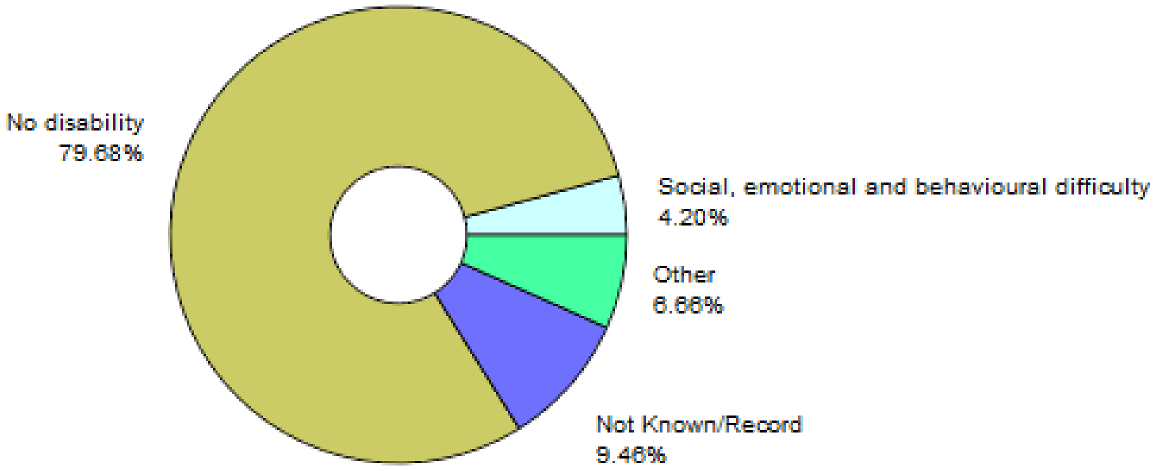


Figure 7.11: Education Data: Main Disability Distribution (N=8185)

Student Stage

Our last key variable is student stage, an indicator of what level of schooling the individual receives. This has been included as a proxy for age, because they are strongly correlated; if student stage and age were released together it could undermine any disclosure control carried out on the age variable. Figure 7.13 demonstrates the proxy for age problem by stacking student stage against age when age is recoded into 5 year bands. The banding was carried out using the ‘Banding’ command in the NIAH tool-set, bands were created using the ‘BandingCreate’ command (Bands are created by specifying a banding schema in advance or interactively; users can chose to automatically divide the age range into equally sized bins or by manually entering the age groups). Using student stage it is possible to partition the age bands into their constituent parts, this is more ambiguous than releasing age in years because of the possibility of individuals with different ages from the norm for their stage of education, however it demonstrates the ability to undermine disclosure control methods using proxy indicators.

Risk Assessment Across the Key Variables

Now that there is a better grasp of the dataset and the key variables that were chosen, the demonstration of NIAH can proceed toward completing a risk assessment for the education dataset. Specifically what is proposed to do in the next section is to use NIAH to implement a range of k-anonymity partitions (k=3,5,10) for the key variable combinations. Using the output of NIAH's k-anonymity algorithm as a guide, statistical disclosure control methods

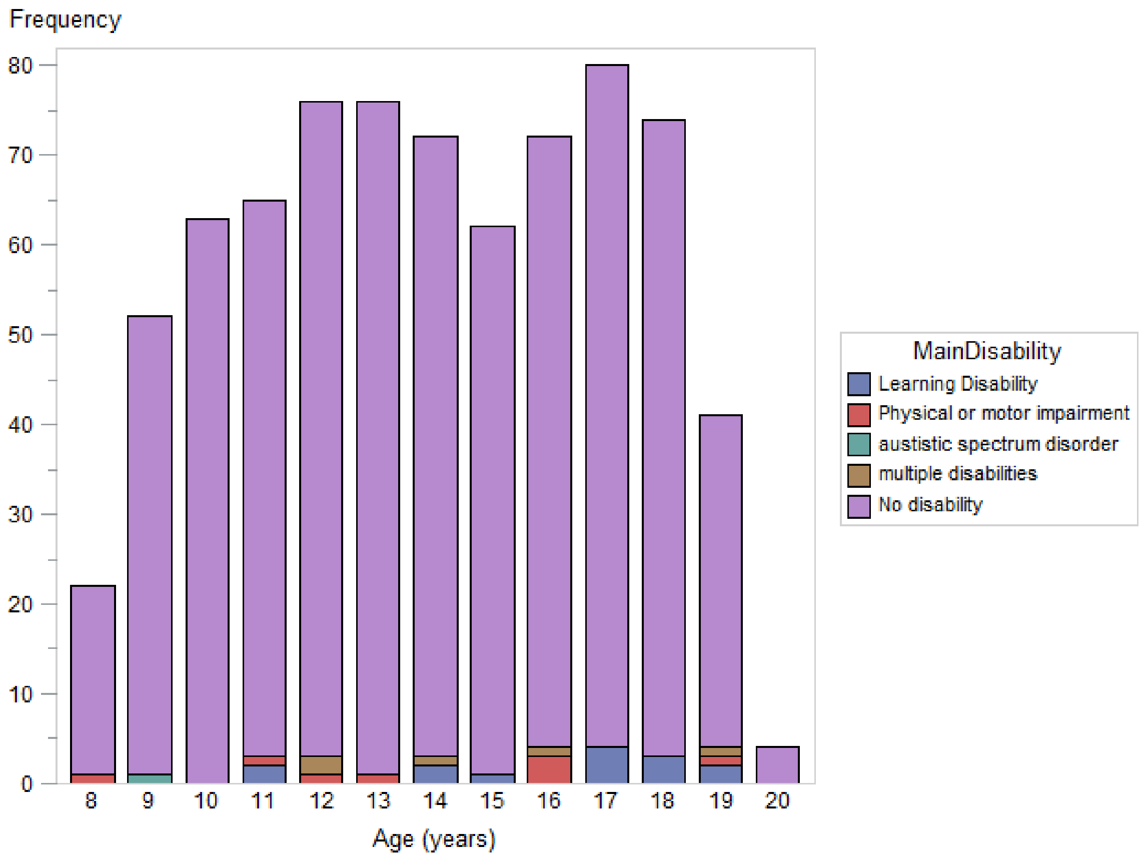


Figure 7.12: Education Data: Main Disability Distribution by Age for the City of Edinburgh Local Authority (N=759)

included in the NIAH tool-set will be implemented. Once a sanitised copy of the original data has been created, the effects of the disclosure control on the data's utility will be examined by conducting a small exploratory analysis for the research question: what factors affect the school attendance of 'looked after children'. This analysis will be carried out on the original data and the sanitised data in order to draw some comparison. This approach to data utility is discussed in detail in Chapter 4.4 and is in keeping with the qualitative need to discuss data utility in terms of research questions and realistic scenarios.

Our first NIAH k-anonymity iteration includes all key variables listed in table 7.1 except we have chosen to use Local Authority as the geographic variable in this first iteration.[4] Later examples including the other geographical variables will follow. Table 7.5 provides an overview of the results generated by NIAH for all key variables by Local Authority. The first point to note is the number of records deemed to be 'at risk' of disclosure, with the threshold set at 3, in this instance 40% of the records are flagged. Remember that k=3 is a relatively loose threshold (i.e. an indicator that the data are not deemed particularly sensitive

[4] The NIAH command used here was: NIAH -kv 1,4,5,6,7,37,41,51,52 -th 3 -o kv-all-k3 ageinyrs.csv — the iterations that follow will be some variation on this command string.

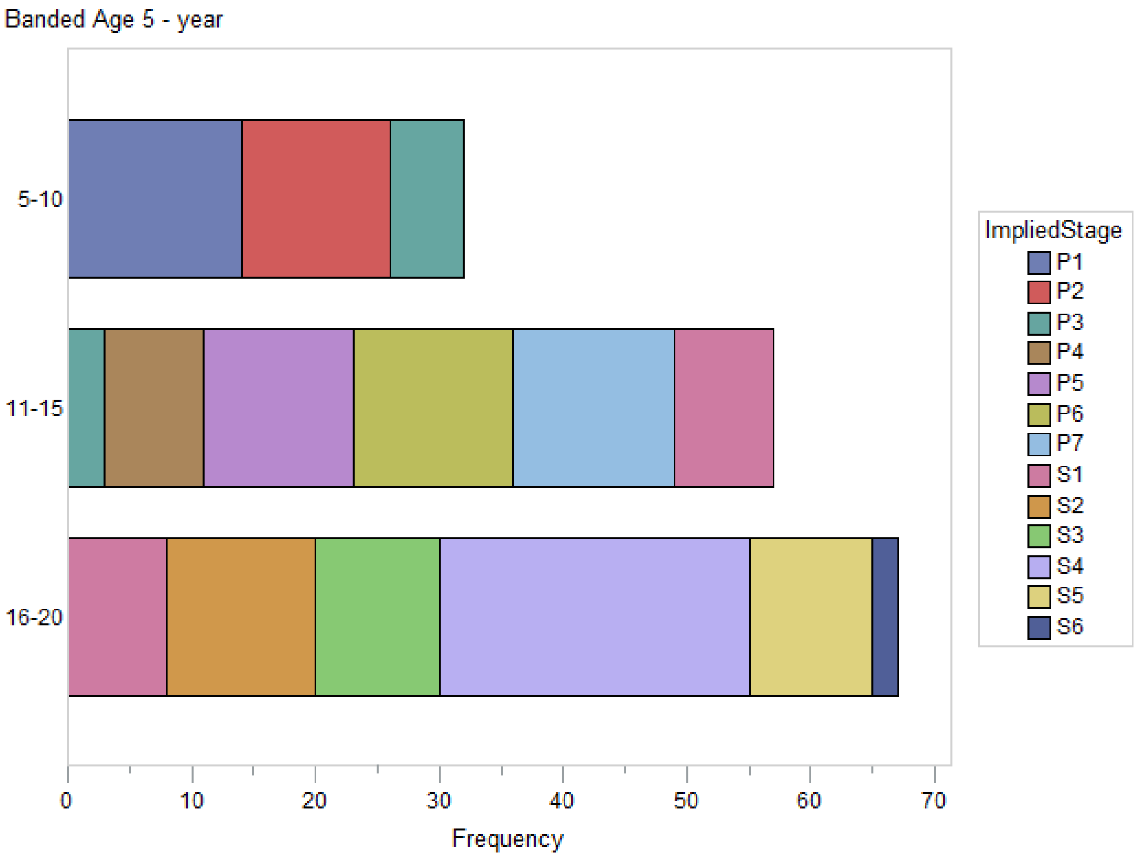


Figure 7.13: Education Data: Student Stage by Age (Recoded to 5yr Bands) for the Angus Local Authority (N=156)

or disclosive), the logic being that for every record there are at least 2 other records with the same values across the key variables. However, at this stage, the raw data are still being used without any disclosure control, therefore this position of 40% should decrease as we continue with the analysis.

Table 7.5 also highlights the problem of identifying risk in a dataset with particularly homogeneous variables. By using the mode as the summary measure, and knowing that the data are heavily skewed, there is little substantive change in the variable values across the partitions. However, the percentage of N represented by the modal value has been included, which does give some indication of the movement of records from 'safe' to 'at risk'. As discussed in the section on key variables, ethnicity is a particular challenge here. However we can see that the partitioning on k=3 has an effect on the homogeneity of ethnic group and national identity. It would not be unrealistic to assume these two variables are strongly correlated. Table 7.6 and 7.7 show the contingency table for these two categorical variables and some association statistics. This supports the assumption about ethnic group and national identity. Given that there are a significant number of cells with low cell counts, the Likelihood Ratio Chi-square result is the more useful association statistic. There is a statistically

Variable (Measure)	Original Data (N=8185)	NIAH Safe Output (N=4922)	NIAH At Risk Output (N=3263)
Age			
Mean($\pm 1\sigma$)	(10.7) 14.0 (17.3)	(10.5) 13.7 (16.9)	(11.0) 14.4 (17.8)
Gender			
Mode (% of N)	Male (54%)	Male(53%)	Male (55%)
Ethnic Group			
Mode (% of N)	White (92%)	White (99%)	White (88%)
National Identity			
Mode (% of N)	Scottish (76%)	Scottish (91%)	Scottish (53%)
Main Disability			
Mode (% of N)	No Disability (80%)	No Disability (91%)	No Disability (63%)
Student Stage			
Mode (% of N)	S3 (10%)	S3 (11%)	SP (12%)

Table 7.5: Education Data: Summary of NIAH Output using all Key Variables and Local Authority, *k*=3

significant correlation here and the strength of the association is indicated by Cramer's V (a measure of association between two nominal variables).

Ethnic Group	National Identity						
	Scottish	English	Northern Irish	Welsh	British	Other	Total
White	79.62	2.51	0.1	0.11	11.77	0.45	96.92
Mixed	1.11	0.08	0	0	0.4	0.08	1.7
Asian	0.15	0.01	0.01	0	0.25	0.04	0.51
Black	0.01	0.04	0	0	0.1	0.14	0.29
Other	0.15	0	0	0	0.03	0.1	0.29
Total	81.3	2.66	0.11	0.11	12.57	0.81	100.0

Table 7.6: Education Data: A Contingency Table of National Identity by Ethnic Group (values are percentages)

Statistic (N=8185)	DF	Value	Prob
Chi-Square	30	1039.69	.0001
Likelihood Ratio Chi-Square	30	239.83	.0001
Cramer's V		0.1688	

Table 7.7: Education Data: Association Statistics for Table 7.6

Having identified a shift in the percentage of records that carry the modal value, this is further illustrated in figures 7.14,7.15,7.16 and 7.17 which show the distribution of ethnic group and national identity for the at risk and safe partitions of the data in this first iteration. These figures illustrate the broader picture and confirm that the shift in the pervasiveness of modal value did indicate a removal of less-well-represented national identities and ethnicities to the at risk partition.

Continuing the comparison of geographical effects on disclosure risk, Table 7.8 provides the same summary of outputs generated by NIAH for the Scottish Parliamentary Constituencies (SPC). In the earlier discussions of the key variables, some coarse indications of disclosure risk were identified by looking at counts across the geographical variables. It was noted that despite the smaller areas at SPC level, the size of the cell counts for each area did not look significantly different from the LA areas. At this level of analysis, taking into account the cross tabulations of key variables, it can be seen that at the SPC level 63% of records are considered to be at risk, which is a large increase on the 40% of records in the LA analysis. Given that the key variables are the same and only LAs for SPCs were exchanged, some confidence can be taken in the assumption that the increase in the at risk records partition can be attributed to the use of SPCs.

Variable (Measure)	Original Data (N=8185)	NIAH Safe Output (N=4922)	NIAH at Risk Output (N=3263)
Age			
Mean($\pm 1\sigma$)	(10.7) 14.0 (17.3)	(10.7) 13.8 (17.0)	(10.7) 14.1 (17.5)
Gender			
Mode (% of N)	Male (54%)	Male(53%)	Male (54%)
Ethnic Group			
Mode (% of N)	1.White (92%)	1.White (95%)	1.White (91%)
National Identity			
Mode (% of N)	1.Scottish (76%)	1.Scottish (95%)	1.Scottish (91%)
Main Disability			
Mode (% of N)	84.No Disability (80%)	84.No Disability (92%)	1.No Disability (72%)
Student Stage			
Mode (% of N)	S3 (10%)	S3 (12%)	S3 (10%)

Table 7.8: Education Data: Summary of NIAH Output using all Key Variables and Scottish Parliamentary Constituency, *k*=3

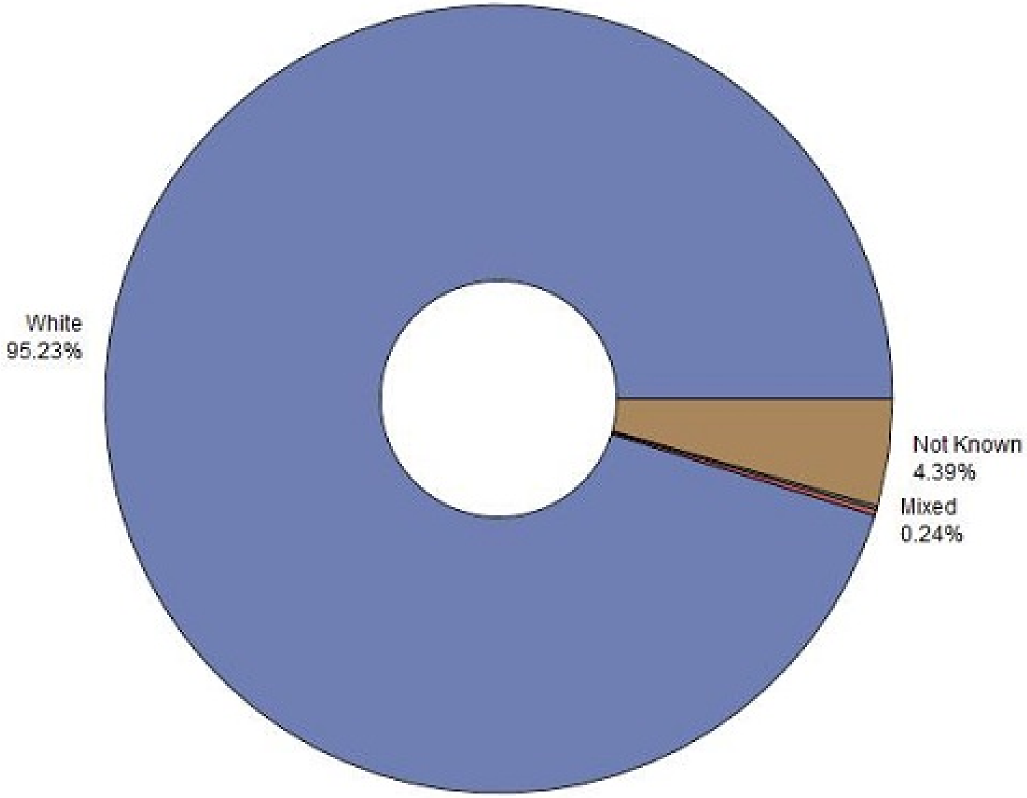


Figure 7.14: Education Data: Ethnic Group for the ‘safe’ Partition (N=4922)

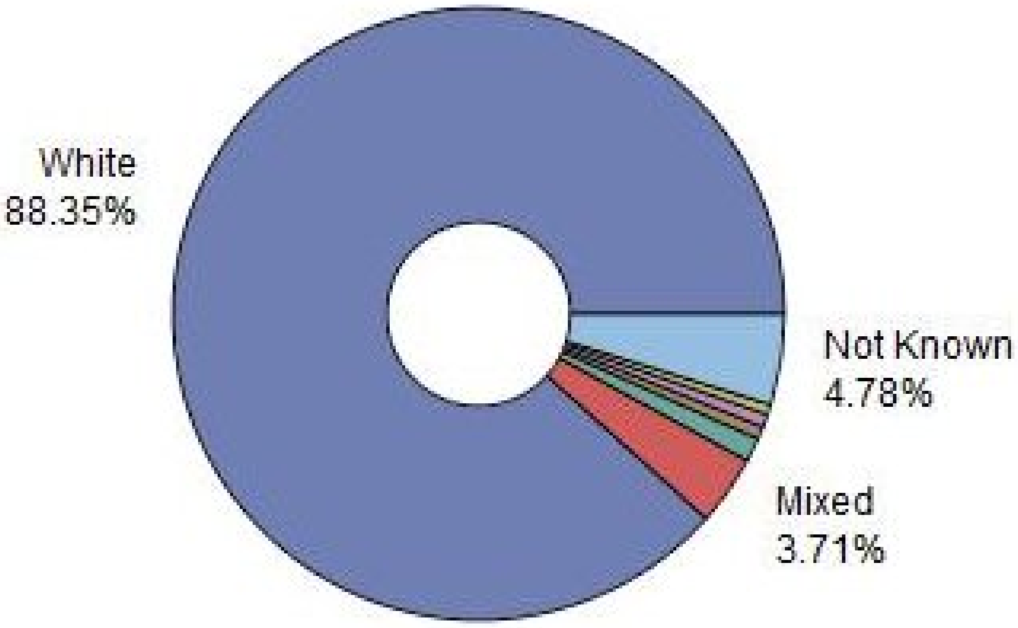


Figure 7.15: Education Data: Ethnic Group for the ‘at risk’ Partition (N=3263)

Figure 7.16: Education Data: National Identity for the ‘safe’ Partition (N=4922)

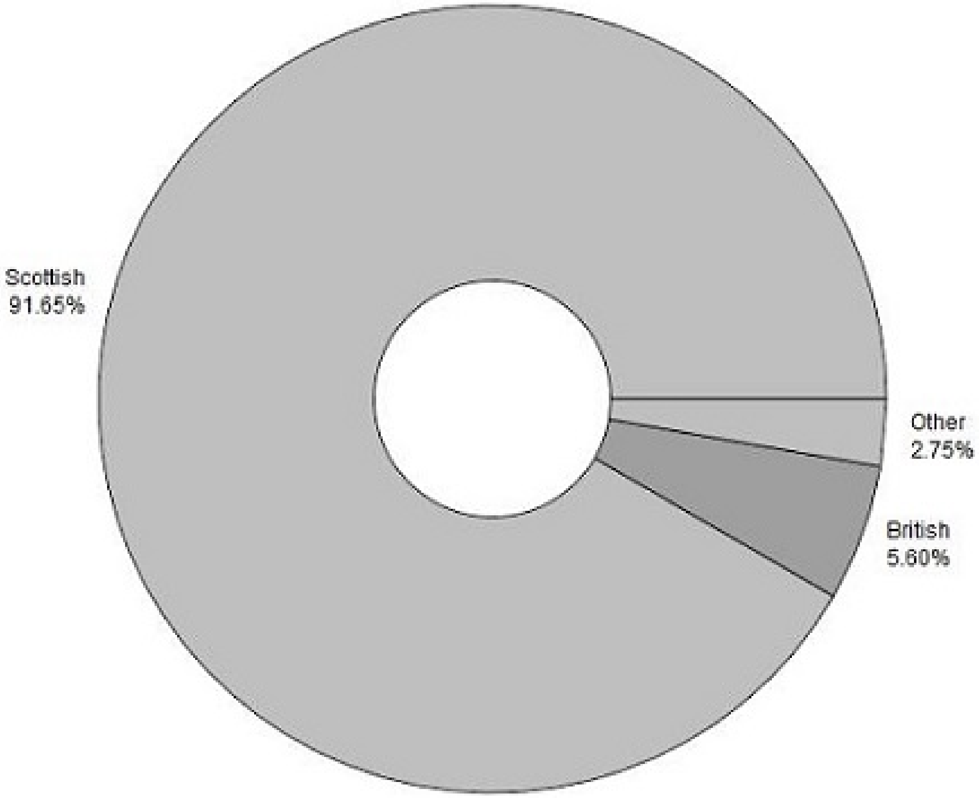
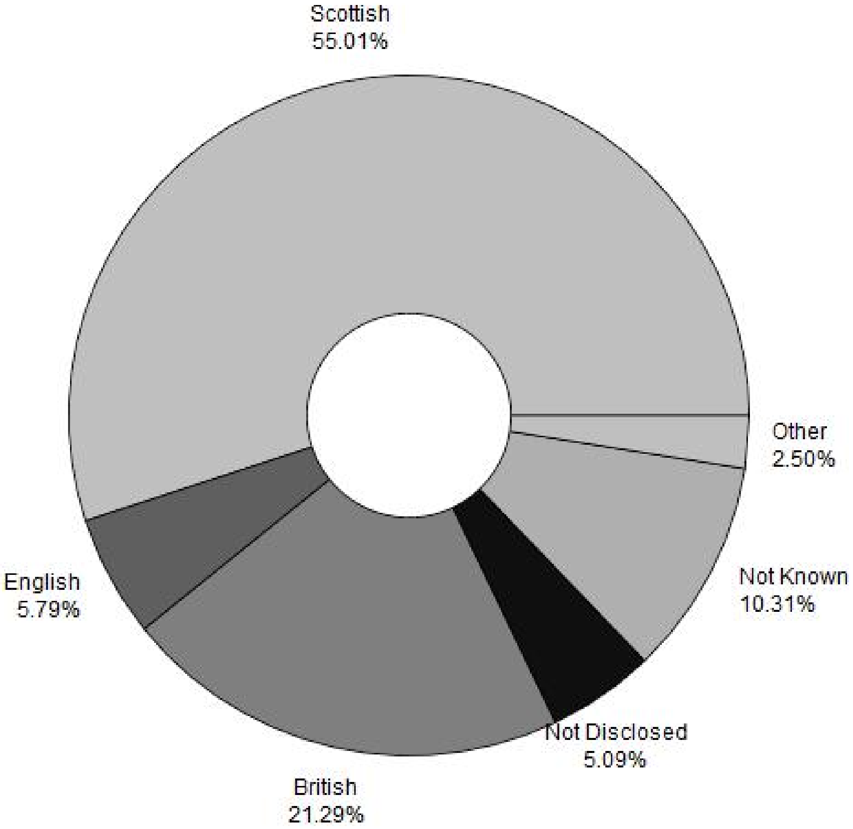


Figure 7.17: Education Data: National Identity for the ‘at risk’ Partition (N=3263)



Due to the low level of geography that datazones represent, the same summary of key variables has not been included because with a threshold of k=3 (the least onerous of the thresholds) 97% of the records are flagged as at risk of disclosure. Referring back to the earlier description of datazones, it can be noted that a figure of 97% is not unexpected as the number of records in each datazone is very small.

From the summary tables, the modes for categorical variables provide an indicator of a shift in the distribution across the partitions for safe and at risk. However, to be able to hone in on areas of disclosure risk, the data can be described in a different way. Figures 7.18 & 7.19 are detailed comparisons of the categorical variable implied student stage for all data and the safe partition, and subsequently the safe partition and at risk partition. If pairs of data points are compared for each respective student stage it can be seen whether there is a significant difference in the proportions. A difference between the safe and at risk partitions was expected, however the original versus safe partition comparison is worth exploring further. Despite 40% of records being flagged as at risk, the proportions for the safe records are relatively unchanged for the student stage variable. All student stages have overlapping confidence intervals until the later stages (S5 onwards). S5 and S6 are poorly represented in the dataset, especially S6 which only has 25 associated records.

Special education (coded as SP) has the largest variation in proportions for safe and original data. Considering only frequency counts, this variation is not expected as there are 500 SP records, which is comparable in size to most of the other student stages. However, there is a possible interaction between the key variables student stage and main disability. For the SP records $\tilde{4}0\%$ have a main disability recorded, compared to less than 20% for most other student stages. What is implied here is that these records are more likely to have been flagged as at risk because they contain a relatively rare main disability, which has the knock on effect that they are more likely to be in special education. Cross-tabulating these two variables for the original data (student stage, recoded into three stages - Primary, Secondary and Special) shows a significant correlation (Likelihood Ratio Chi-Square = 571.71, p<.0001, Cramer's V = 0.246).

Another way the results can be interpreted (figure 7.18) is to consider the analytical validity of analysis were it only to be carried out on the safe partition. The overlapping confidence intervals for the earlier student stages, up to S4, suggests that student stage could be analysed using the safe data with no significant loss of validity from a researcher's perspective. This comparison between the raw data and the, albeit crudely, sanitised data could assist in developing trust in the efficacy of the disclosure methodology of a data controller. This type of analysis would fit well with the arguments of Reiter et al. (2009) on the need for 'verification servers' that offer analysts the opportunity to confirm the validity of their results with reference to the original data. This could also be applied to the difference between the NIAH iterations for LAs and SPCs. Taking the overall number of records deemed at risk, as above,

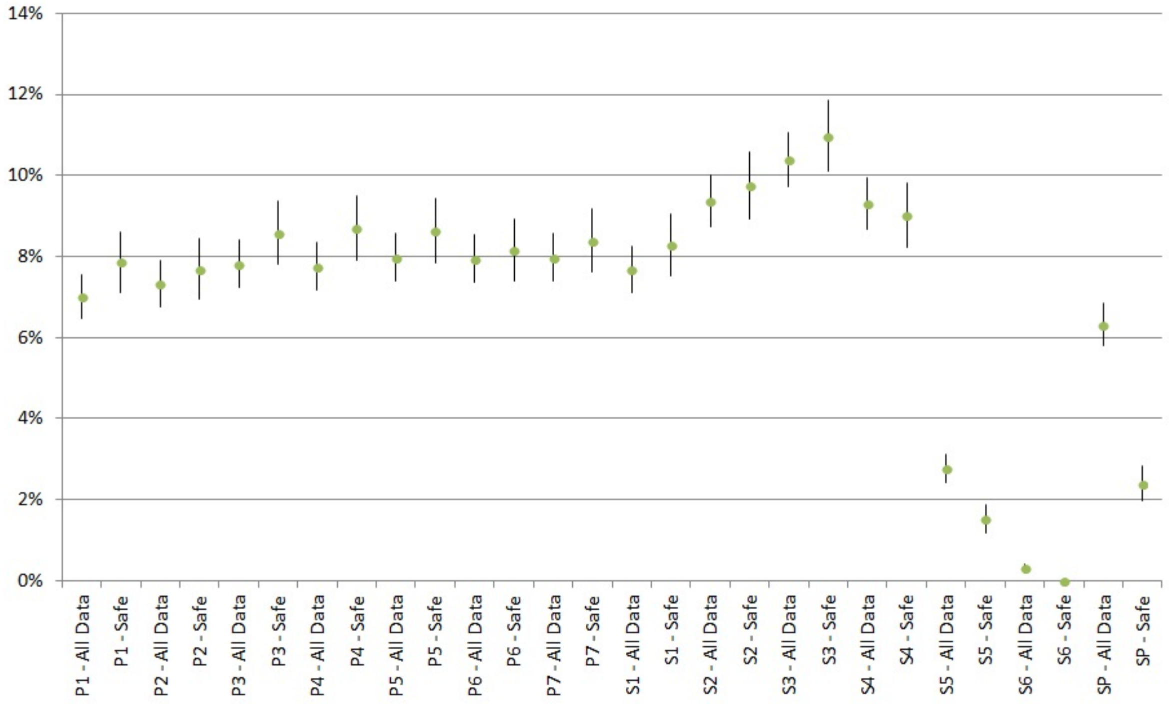


Figure 7.18: Education Data: Student Stage Proportions with 95% CIs for All Records & the 'safe' Partition

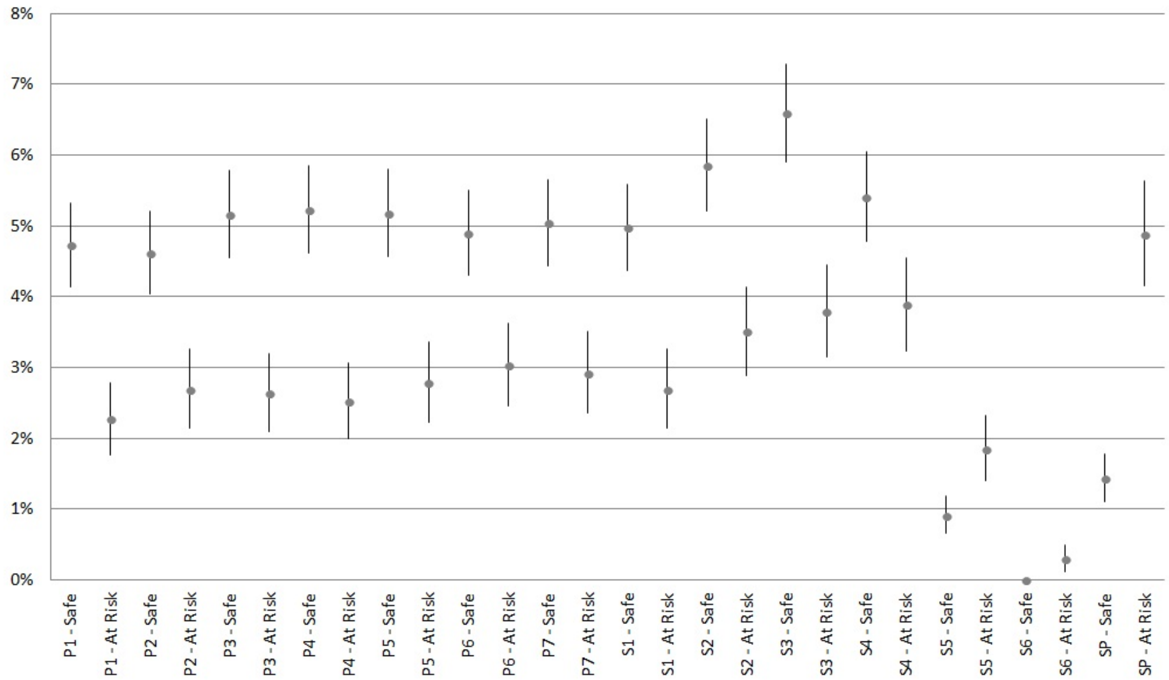


Figure 7.19: Education Data: Student Stage Proportions with 95% CIs for the 'safe' & 'at risk' Partitions

is one indication of difference. However, if Table 7.8 is considered for analytical validity from original records to the safe partition, it is interesting to note that despite the greater number of at risk records the modal values for some of the variables in the safe partition track the original data more closely than the safe partition in the LA summary.

Having considered the relatively generous k-anonymity threshold of k=3, k is now increased to 5. Table 7.9 provides the same summary from the NIAH output as seen in 7.5. As we would expect the number of records considered safe for this iteration is less than k=3, here we have an 18% decrease. From this decrease it can be extrapolated to the line of argument governing the use of k-anonymity; by increasing the threshold, the requirement for privacy, we have reduced the volume of data considered safe, in this case by approximately 18%, and therefore attention should be paid to the resulting data utility. This is done as before, by comparing the summary outputs of k=3 and k=5. Age has remained stable varying only by 0.1 years at the mean, as has gender. Ethnic Group shows that the propensity toward homogeneity that one expects with a threshold approach such as k-anonymity can be misleading. 'White' now constitutes 95% of the safe data as opposed to 99% at k=3. From this result, it can be surmised that given an increase in the threshold, instead of just records with a less well represented ethnicity being transferred into the at risk partition, there are now a number of 'white' records with a combination of key variables that do not meet the k=5 requirement. This counter intuitive trend does not continue through National Identity where the homogeneity increases; 'Scottish' now representing 95% as opposed to 91%, the same is true of Main Disability.

Comparing the summaries of k=3 and k=5 builds a more nuanced picture of the interplay of disclosure risk and data utility. As is the case with ethnicity and the earlier comparison of LAs and SPCs, counter intuitive effects have been shown. These products of the interaction of the different key variables and their ability to satisfy a given threshold make it difficult to create generalised rules of disclosure control, even within the confines of a single dataset or research project. To continue this narrative, the summary results for the k=10 iteration have been included. From table 7.10 it is first observed that there is a significant drop in the number of safe records, only 19% of the records are deemed to satisfy k=10. It is worth noting here that such a strong threshold is not unrealistic, standard operating procedure for Information Services Division of NHS Scotland is to apply a k=10 threshold when assessing statistical disclosure control. Given the size of the at risk partition it is not unsurprising that it is now that partition that begins to look closer to the distribution of the original data. The counter intuitive trend in ethnic group can still be seen if k=10 is compared with both k=3 and k=5, also the homogeneity in national identity and main disability continue.

To reinforce the lack of clear linear rules which can fit the disclosure analysis, consider first the trend in the size of partitions across k=3,5,10. Figure 7.20, with only three data points, still shows that the decrease in the size of the safe partition is not linear. Although no further

Variable (Measure)	Original Data (N=8185)	NIAH Safe Output (N=3546)	NIAH At Risk Output (N=4639)
Age			
Mean($\pm 1\sigma$)	(10.7) 14.0 (17.3)	(10.4) 13.6 (16.8)	(10.9) 14.3 (17.7)
Gender			
Mode (% of N)	Male (54%)	Male(52%)	Male (54%)
Ethnic Group			
Mode (% of N)	White (92%)	White (95%)	White (91%)
National Identity			
Mode (% of N)	Scottish (76%)	Scottish (95%)	Scottish (62%)
Main Disability			
Mode (% of N)	No Disability (80%)	No Disability (93%)	No Disability (69%)
Student Stage			
Mode (% of N)	S3 (10%)	S3 (12%)	SP (11%)

Table 7.9: Education Data: Summary of NIAH Output for all Key Variables and Local Authority, *k*=5

Variable (Measure)	Original Data (N=8185)	NIAH Safe Output (N=1584)	NIAH At Risk Output (N=6601)
Age			
Mean($\pm 1\sigma$)	(10.7) 14.0 (17.3)	(10.6) 13.7 (16.8)	(10.6) 14.0 (17.4)
Gender			
Mode (% of N)	Male (54%)	Male(52%)	Male (54%)
Ethnic Group			
Mode (% of N)	White (92%)	White (97%)	White (92%)
National Identity			
Mode (% of N)	Scottish (76%)	Scottish (98%)	Scottish (71%)
Main Disability			
Mode (% of N)	No Disability (80%)	No Disability (97%)	No Disability (76%)
Student Stage			
Mode (% of N)	S3 (10%)	S3 (12%)	S3 (10%)

Table 7.10: Education Data: Summary of NIAH Output for all Key Variables and Local Authority, *k*=10

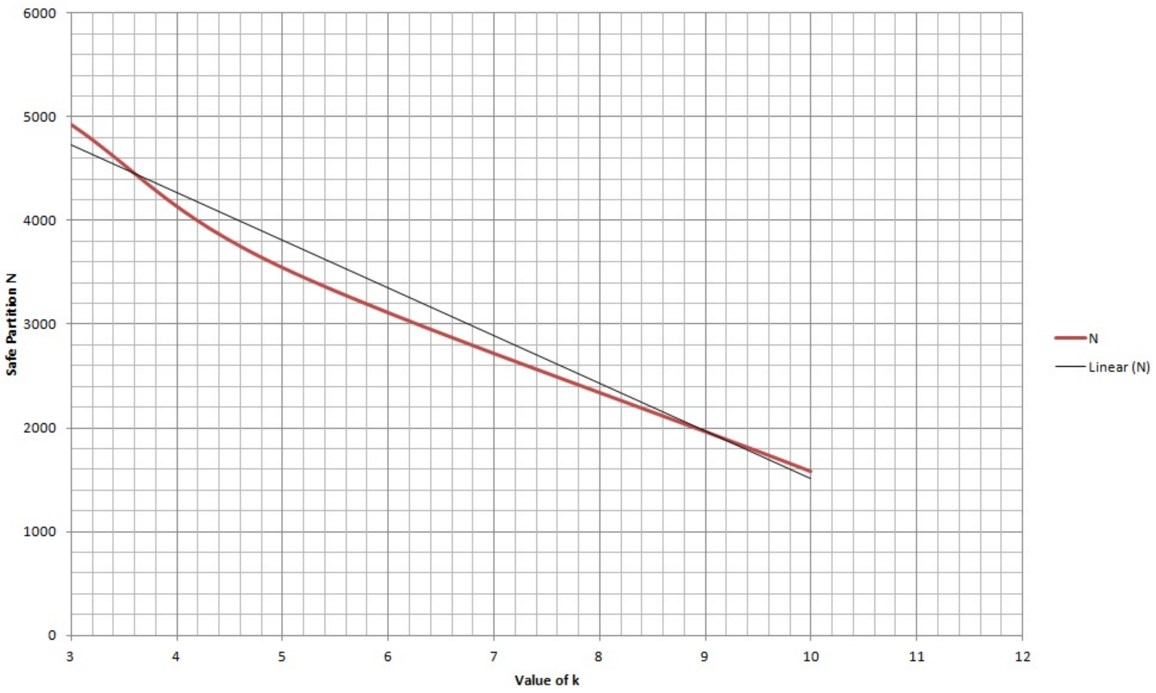


Figure 7.20: Education Data: Size of the Safe Partitions for the *k*-anonymity Thresholds 3,5 & 10

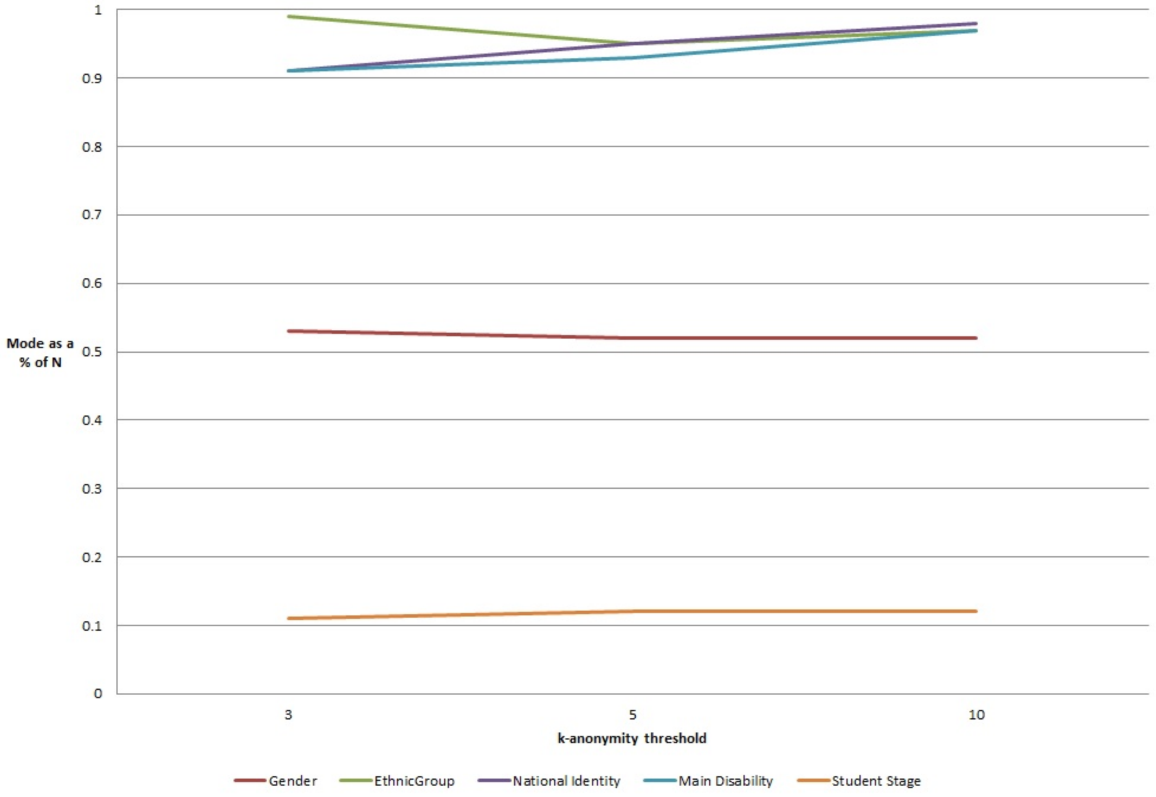


Figure 7.21: Education Data: Modal Values as a Percentage of N over the *k*-anonymity Thresholds 3,5 & 10

increases in k were made during the allocated analytical time at the Scottish Government, it would be expected that k=15 for example would result in none or very few records deemed safe. Figure 7.21 attempts to capture the comparison between the key variables for different values of k. Gender and Student Stage remain flat across the iterations of k, Ethnic Group decreases in a non-linear fashion, while National Identity and Main Disability increase. At this stage the results indicate that the variables ethnic group, national identity and main disability are potential targets for disclosure control alongside age which is often also a target because of it's accessibility. Iterations of k-anonymity with some exploratory disclosure control applied are now explored.

Recoding as discussed in chapter 4.3 is a common strategy for disclosure control. The variable values are recoded to remove some aspect of the variable's detail and therefore reduce the ability of an adversary to make a correct match to a unique record. This also makes the possible matches more ambiguous by providing a coarser level of detail should an adversary take a fuzzy, rather than exact, matching approach. For some institutions, recoding is the default position for disclosure control, for example ISD in NHS Scotland (NHS Scotland ISD, 2012b) defines recoding as the only strategy open to data controllers without explicit permission from senior statisticians. Recoding can often change the variable type, in the case of age in years the numerical variable becomes categorical representing a group of ages. Using the NIAH BandingCreate and Banding commands the age variable was recoded into categories spanning 5 years, e.g. $11 - 15$ and applied bottom-coding to capture all of the lower values— age 10 and below. By choosing to recode age, the student stage variable was suppressed, as indicated earlier student stage acts as a proxy for age as they share a direct correlation. Therefore, the revised re-identification key is LA, age(5yr bands), Gender, Ethnic Group, National Identity and Main Disability.

We then run NIAH's k-anonymity algorithm for k=3 with this key; table 7.11 shows the summarised outputs. First note that 86% of records are in the safe partition, this in an increase of 25% from the original k=3 iteration with age in single years and student stage. Having retained more of the original records the distribution of the key variables should be considered. Age and Gender remain relatively static, the mode representing the same proportion of records $\pm 1\%$. Ethnic Group sees a small increase of 3% in those records with a 'white' value. The biggest shifts in this iteration are seen in the National Identity and Main Disability variables, the safe records representing a more Scottish sample with fewer recorded disabilities. These results would seem to indicate that the last two variables are the potential source for the disclosure risk. In the next iteration Main Disability was recoded to assess these effects.

Recall that the distribution of main disability (figure 7.11) in the raw data had a skew toward no disability; this makes finding a suitable recoding scheme difficult because to retain the relatively rare disability details would leave too many records open to the risk of disclosure.

Variable (Measure)	Original Data (N=8185)	NIAH Safe Output (N=7001)	NIAH At Risk Output (N=1184)
Age			
Mode(% of N)	11-15(41%)	11-15(42%)	16-20(44%)
Gender			
Mode (% of N)	Male (54%)	Male(53%)	Male (54%)
Ethnic Group			
Mode (% of N)	White (92%)	White (95%)	White (80%)
National Identity			
Mode (% of N)	Scottish (76%)	Scottish (83%)	Scottish (36%)
Main Disability			
Mode (% of N)	No Disability (80%)	No Disability (85%)	No Disability (46%)

Table 7.11: Education Data: Summary of NIAH Output for Recoded Age and Suppression of Student Stage by Local Authority, *k*=3

In this example, a binary indicator was created for whether a disability was recorded or not. The aim here is to at least preserve some utility from the main disability variable rather than suppress it entirely. Table 7.12 shows that the number of safe records has increased by 4% and the apparent gap has narrowed in the distribution of the main disability variable across the original and safe records. The biggest gap between the original and safe records in the summary is now national identity, with a gap of 7% in the records represented by the modal value.

In order to contemplate disclosure control for the national identity variable the association effects with ethnicity and the relative distributions of these variables should be considered. As highlighted in the earlier consideration of key variables, ethnicity represents a particular problem for data in Scotland, national identity provides some protection from disclosure risk through its greater subjectivity. Considering the distributions of these two variables, ethnic group is more heavily skewed than national identity; the ‘white’ ethnic group constitutes 92% of the records, whereas ‘Scottish’ is recorded for 77% of the records. For these reasons it was decided to recode national identity into a binary indicator, ‘Scottish’ or ‘Other’ and suppress the ethnic group variable. It is worth reiterating that k=3 is a relatively generous threshold, especially for population level data. It should be noted, therefore, that significant sacrifices were made in respect of data utility to increase the number of records that satisfy the k-anonymity requirement. Table 7.13 provides a summary of the final iteration at this stage. The increase of 6% in the number of safe records is noted, and now 96% of all records are captured. Instead of considering the safe output values, the at risk partition was

Variable (Measure)	Original Data (N=8185)	NIAH Safe Output (N=7328)	NIAH At Risk Output (N=857)
Age			
Mode(% of N)	11-15(41%)	11-15(42%)	16-20(43%)
Gender			
Mode (% of N)	Male (54%)	Male(54%)	Male (51%)
Ethnic Group			
Mode (% of N)	White (92%)	White (95%)	White (73%)
National Identity			
Mode (% of N)	Scottish (76%)	Scottish (83%)	British (24%)
Recoded Disability			
Mode (% of N)	No Disability/Not Recorded (89%)	No Disability/Not Recorded (91%)	No Disability/Not Recorded (71%)

Table 7.12: Education Data: Summary of NIAH Output using Recoded Age and Disability with the Suppression of Student Stage by Local Authority, *k*=3

considered to evaluate the effect the disclosure control had. It can been seen that the at risk records are more likely to be older, of another national identity and are more likely to have a recorded disability.

Thus far the closeness between distributions in the safe partition has been considered. When compared with the original data, this provides the rudimentary measure of the data's utility after disclosure control. The impact of the disclosure control is now considered with respect to a potential research question. The remaining 4% of records that do not satisfy k=3 are suppressed and therefore only the safe partition is used. These results will be compared with the same analysis carried out on the original data. The research question is to explore what factors affect looked after children achieving above or below the median percentage attendance at their educational institution. The variables age, gender, local authority and placement type were used as exploratory variables. A binary indicator was constructed dependent on whether a record has above or below the median attendance (c.92%).

The analysis has been broken down into a series of models. Table 7.14 is a comparison of models 1-6 for reference, however each model is discussed in detail below. These models were generated using the Scottish Governments SAS Statistics Server, using the binary logit regression model as there was a binary dependent variable in the below and above median attendance variable 'GoodAttend'. It should first be noted that to aid the ability to directly compare the models, the age recoding was applied to the original data without any further

Variable (Measure)	Original Data (N=8185)	NIAH Safe Output (N=7821)	NIAH At Risk Output (N=364)
Age			
Mode(% of N)	11-15(41%)	11-15(42%)	16-20(40%)
Gender			
Mode (% of N)	Male (54%)	Male(54%)	Male (51%)
Recoded National Identity			
Mode (% of N)	Scottish (76%)	Scottish (79%)	Other (37%)
Recoded Disability			
Mode (% of N)	No Disability/Not Recorded (89%)	No Disability/Not Recorded (91%)	No Disability/Not Recorded (51%)

Table 7.13: Education Data: Summary of NIAH output using Recoded Age, Disability and National Identity with Student Stage and Ethnicity Suppressed by Local Authority, *k*=3

statistical disclosure control; although this does sacrifice some of the original data's utility, it makes comparison between models easier rather than having to interpret age as a numerical and a categorical variable. To mitigate the effect of this on the overall assessment, Table 7.15 was included, which is a simple linear regression for the dependent variable percentage attendance and age in years. The F-test shows that the model is statistically significant and has R^2 value of 0.1096 suggesting that 11% of the variance in percentage attendance can be explained by age in years. Also the t-test is significant with a parameter estimate of -0.02 - i.e., for every unit increase in age one would expect a 0.02 unit decrease in the percentage attendance. Therefore, the model suggests that as looked after children get older, their percentage attendance in education declines.

Model 1 is the first model using the disclosure controlled data, and recoded age is the only exploratory variable. The N here is different because those records that remained in the at risk partition after the various disclosure control methods were implemented were suppressed. It should be noted that SAS defaults to modelling the 0 value in its binary logit model rather than the 1 value as in stata and R, this behaviour highlights one aspect of the problems associated with working in constrained safe haven conditions where unfamiliarity can lead to errors in interpretation. Therefore, when considering these results, remember that negative parameter estimates are counter-intuitive, a negative value indicates that the parameter is associated with a positive change in attendance above the median.

Consulting table 7.16 it should be noted that N has dropped to 7821 to account for those suppressed records. There is a R^2 of 0.1018 similar to the indicative linear regression for

Dataset	Model 1 Safe	Model 2 Orig.	Model 3 Safe	Model 4 Orig.	Model 5 Safe	Model 6 Orig.
Age 11-15	-0.0258	-0.0405	-0.0251	-0.0402	-0.0249	-0.0400
(standard error)	(0.0320)	(0.0308)	(0.0320)	(0.0308)	(0.0321)	(0.0308)
Age 16-20	-0.6100	-0.5916	-0.6091	-0.5912	-0.6080	-0.5899
	(0.0329)	(0.0318)	(0.0329)	(0.0318)	(0.0329)	(0.0318)
Age 21-25		0.00155		0.000363		-0.00359
		(0.2647)		(0.2648)		(0.2648)
Gender(F)			-0.00140	-0.00728	-0.0150	-0.00815
			(0.0236)	(0.0230)	(0.0237)	(0.0230)
No. of Placements					0.0532	0.0553
					(0.0296)	(0.0291)
pseudo R^2	0.1018	0.0926	0.1019	0.0926	0.1024	0.0932

Table 7.14: Education Data: Models 1-6: Influences on having lower than median attendance, using the ‘safe’ & Original Data

No. Obs: 8185 R-Square: 0.1096

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-Value	Pr>F
Model	1	39.91177	39.91177	1007.02	<.0001

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t-Value	Pr>t
Intercept	1	1.12491	0.00956	117.67	<.0001
Age(yrs)	1	-0.02113	0.00066584	-31.73	<.0001

Table 7.15: Education Data: Linear Regression for Percentage Attendance and Age(in years) using the Original Data

percentage attendance and age in years. However, this figure needs to be interpreted differently given the different models being used. SAS reports the Cox-Snell R^2 which is not directly comparable with the R^2 reported for the linear regression. However, the later models are all binary logits and therefore the R^2 will be more comparable. As a general indication the R^2 value suggests that 10% of the records are explained by this model. For more detail on the differences between R^2 measures see Menard (2000). The bottom age group of less than 10yrs is excluded, so the results are based, on being in one of the older categories as opposed to less than 10yrs old. Age group 11-15 shows a negative estimate, suggesting a small positive effect on good attendance, however it is not statistically significant. Age group 16-20 shows a strong positive effect on good attendance when compared against being less than 10yrs old. The top age group 21-25 is not present as all records with this value have been suppressed as part of the disclosure control. Although the 11-15 estimate is not significant, it is in keeping with the positive trend seen at 16-20, and therefore the first model indicates that as looked after children get older the probability that they will have above average attendance at school increases. This differs from the position seen in the linear model of age and percentage attendance. Although the trend was not strong, the linear model indicated that a decrease in percentage attendance as the data subjects got older should be expected. It is difficult to compare directly these two models as the dependent variable is slightly different — percentage attendance vs. above average attendance — however, Model 1 and 2 which share the same dependent variable and recoding of age can be compared.

No. Obs: 7821 R-Square: 0.1018

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	0.1465	0.0264	30.8955	<.0001
Age 11-15	1	−0.0258	0.0320	0.6494	0.4203
Age 16-20	1	−0.6100	0.0329	344.6688	<.0001

Table 7.16: Education Data: Model 1, Good Attendance and Recoded Age using the ‘safe’ Data

Model 2 shown in table 7.17 repeats model 1 using the original data. The goodness of fit statistics, 0.0926 and 0.1018 can now be compared, so the models explain about the same percentage of the data. The model estimates are also not too dissimilar from model 1. The age group 16-20 still has a significant positive effect on good attendance. 11-15 continues the trend but is not significant, and there are also results from the upper age group but these provide no significant insights. The strength of the effect of being 16-20 rather than less than 10yrs is also similar, only differing by 3%. Thus far the disclosure control has not significantly affected the data utility for this specific research question.

No. Obs: 8185 R-Square: 0.0926

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	0.2282	0.2660	0.7360	0.3910
Age 11-15	1	−0.0405	0.0308	1.7274	0.1887
Age 16-20	1	−0.5916	0.0318	346.5663	<.0001
Age 21-25	1	0.00155	0.2647	0.0000	0.9953

Table 7.17: Education Data: Model 2, Good Attendance and Recoded Age using the Original Data

Model 3 extends the exploratory variables to include gender. In this iteration the safe data are used. The R^2 remains close to 0.10. The inclusion of gender has not significantly changed the model estimates for the different age groups, there is still a strong positive effect from the 16-20 category. Gender modelled as female rather than male has a small positive effect however this is not significant. If this is compared with model 4 in table 7.19, it is observed that the general pattern remains the same, the R^2 values are similar, and the parameter estimates do not change sign or strength. The age group 16-20 still provides the only statistically significant result. Gender is also not significant in model 4. Again, thus far the disclosure control has not adversely effected the analysis; this is discussed further in the conclusion to this section, however it is important to note that the disclosure control did not directly manipulate the dependent variable or the exploratory variables (except age).

No. Obs: 7821 R-Square: 0.1019

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	0.1452	0.0265	30.1275	<.0001
Age 11-15	1	−0.0251	0.0320	0.6125	0.4339
Age 16-20	1	−0.6091	0.0329	343.1089	<.0001
Gender(F)	1	−0.00140	0.0236	0.3499	0.5542

Table 7.18: Education Data: Model 3, Good Attendance with Recoded Age and Gender using the ‘safe’ Data

No. Obs: 8185 R-Square: 0.0926

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	0.2287	0.2660	0.7394	0.3899
Age 11-15	1	−0.0402	0.0308	1.6974	0.1926
Age 16-20	1	−0.5912	0.0318	345.6693	<.0001
Age 21-25	1	0.000363	0.2648	0.0000	0.9989
Gender(F)	1	−0.00728	0.0230	0.1001	0.7518

Table 7.19: Education Data: Model 4, Good Attendance with Recoded Age and Gender using the Original Data

Models 5 and 6 in tables 7.20 and 7.21 respectively now include the further exploratory variable of the number of placements associated with a record. The hypothesis here is that one might see a greater number of placements as having a destabilising effect, which could affect a data subject’s attendance. As the number of placements is numerical and continuous we can interpret the results in model 5 and 6 as the effect of the number of placements increasing. Comparing model 5 and 6 the R^2 values are again similar. The recoded age variable shares the same strength and significance across both models. The change in sign

Table 7.20: Education Data: Model 5, Good Attendance with Recoded Age, Gender and No. of Placements using the 'safe' Data

No. Obs: 7821 R-Square: 0.1024

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	0.0725	0.0483	2.2531	0.1333
Age 11-15	1	−0.0249	0.0321	0.6039	0.4371
Age 16-20	1	−0.6080	0.0329	341.6138	<.0001
Gender(F)	1	−0.0150	0.0237	0.4003	0.5269
No. of Placements	1	0.0532	0.0296	3.2346	0.0721

Table 7.21: Education Data: Model 6, Good Attendance with Recoded Age, Gender and No. of Placements using the Original Data

No. Obs: 8185 R-Square: 0.0932

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	0.1572	0.2687	0.3425	0.5584
Age 11-15	1	−0.0400	0.0308	1.6816	0.1947
Age 16-20	1	−0.5899	0.0318	343.8998	<.0001
Age 21-25	1	−0.00359	0.2648	0.0002	0.9892
Gender(F)	1	−0.00815	0.0230	0.1251	0.7236
No. of Placements	1	0.0553	0.0291	0.3425	0.0578

of the 21-25 age group from model 4 to 6 to fit the general trend should be noted, although this is not a significant result. The new independent variable, the number of placements, does show a slight discrepancy between the safe and original data. For model 6 using the original data, the negative effect of an increasing number of placements is just short of the 0.05 significance level; however, using the safe data, the result is considerably further away; in either case the effect is small.

Having established the example research question, it has been demonstrated that this type of sensitivity analysis is important when data are changed or manipulated to satisfy statistical disclosure control conditions. Given the possible variations in analysis, it is difficult to predict the globalised effects of different disclosure control methods without probing the dataset beyond the surface level. In the example, the analyses would lead an analyst to very similar conclusions, although there are possible variations in the relationship between age and attendance, and the potential for a significant result when considering the number of placements. It should be noted that the research question was designed to avoid using variables that had been heavily effected by the disclosure control choices. This was predominately due to a need for actual results for comparison and to demonstrate NIAH's use in a realistic form. We should remember the data utility that has been lost—we suppressed ethnicity and student stage, and heavily recoded national identity and disability. If a research question looking at the effects of ethnicity, disability or national identity had been chosen, an analytical dead-end would have been found rather quickly.

This example analysis reinforces the need for data controllers to be able to work through any number of variations of disclosure control, sensitive to the type of release they intend to make and to the needs of the research community. The tools available in the NIAH tool set allow for this to be carried out as an iterative process, as has been seen above. It provides clear documentation and replicability to aid in building trust with data users and avoid human bottlenecks in the flow of information, as well as to assist in the application of general standards across data releases.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

This research has sought to explore the ideas surrounding statistical disclosure control by applying an interdisciplinary lens in the context of e-Health and administrative data access. Through this lens the duality of managing risk through the combination of scientific and value judgements has been captured and examined in order to provide a theoretical framework for considering SDC. Within this framework the current literature and methods have been critically reviewed and the contemporary arrangements for secure data access have been scrutinised. This scrutiny resulted in the extension of SDC workflows to encompass the data user experience alongside the more common tests for statistical utility. From these foundations, a tool-kit has been designed and implemented to provide a semi-automated approach to SDC and one that can be integrated with existing workflows to increase the efficiency of research output review and therefore improve the overall experience for data users and data controllers.

The theoretical framework for this research was produced by first reviewing the data controller's legal and ethical obligations to estimate and mitigate the risks of unauthorised disclosure. This was established using the existing data protection legislation as well as similar examples from Europe, the US and Canada. The landscape underpinning those rules was then explored in terms of its ethical dimensions and societal attitudes to privacy risk using the perspectives of Garfinkel, Solove and Nissenbaum. Using the work of Ulrich Beck it was possible to describe the fusion of scientific and value judgements that lies at the heart of risk management, and therefore by extension the mitigation of meta-risks like statistical disclosure. Heidegger's writing on the relationship between people, their world and technology and the concepts of disclosure and enframing provide a greater depth to the theoretical framework and enrich it through a discussion of the fear and insecurities that feature in the debate over data access, anonymisation and SDC.

Having established such a narrative it was then challenged from two dimensions. It was noted what impact future legislation could have on the public versus private balance that are central to contemporary SDC debates that the balance is in danger of tipping too far to the private weight. Also the apparent discrepancy between the approach of the public and private sectors was set against the earlier narrative of the relationship between people and technology to demonstrate the complex relationships people have with services that utilise their data.

The existing literature on risk measurement, statistical disclosure control methods, and data utility were discussed and a case study using the Scottish Health Survey was used to illustrate the major elements; risk measurement, disclosure control methods and estimation of data utility. What emerged at this stage was the vast array of innovative statistical methods that have occurred in the literature, however most have failed to find an application in the contemporary data controller workflow and very few seek to marry a sociological conceptualisations of risk and the technical calibration of systems used to estimate and mitigate disclosure risk. Many of the methods discussed cause severe damage to the data's statistical utility which has the potential undermine the trust between data users and data controllers if their subjective experiences are not acknowledged.

From this position, it was possible to elucidate on the potential barriers to research in the data access, analysis and dissemination process within contemporary data access settings, such as SafePods or Virtual Research Environments. Using data collected through qualitative semi-structured interviews, it was shown that the restrictions researchers face in using physical secure-settings in the UK including relocation, lack of access to good meta data or other analytical aids, and the bottleneck and lack of clarity in output checking could be mitigated by the use of virtual research infrastructure. It was also discussed how this virtual infrastructure brings other benefits to increase productivity by allowing users to share data resources within a secure environment. These early models for VREs offered a backdrop for the exploration of SDC workflows within these settings which fed into the e-Science approach proposed, and later tested, in Chapters 6 and 7.

It should be noted that although the case for the user experience was made, it is recognised that operationalising that experience in the development of SDC workflows is not a simple task. Although, aided by the e-Science methodological context, suggestions were made for how this might be approached. For example, framing the user experience in the same way that software development treats it through user centred systems design.

Drawing on these findings, the NIAH tool-set was designed with input from potential users and data controllers through a small, qualitative, pilot study. This tool set makes available an algorithm that implements k-anonymity, a common assessment method for testing data against a predetermined disclosure risk, which should incorporate the value judgements of the preceding chapters in the setting of thresholds for rare combinations of key variables. In addition, the tool-set design was broadened to include tools to service other parts of the data controller's SDC workflow. These included implementing SDC methods for recoding data variables to more coarse-grained values, adding noise to numerical values or randomly miss-classifying categorical values, as well as value suppression. These tools were designed so that data provenance is well documented and SDC decisions can be reproduced. Also these tools were designed to allow for scripting, automation, and integration with existing research infrastructures.

This last element was demonstrated through the integration of NIAH with the STAT-JR statistical analysis environment and a proposal for integration with the VANGUARD system for secure data transfer and linkage. Lastly, the potential advantages of the NIAH tool set over the existing approaches that use *ad hoc* user written scripts for statistical packages were discussed and demonstrated. These included a significant time saving in the reduction of the number of parameters that must be input by hand, a clear log of activity, a lack of reliance on proprietary statistical packages, and the possibility of some automation. Based on these aspects of this research, the tool set is in use by the Scottish Government and their implementation is described in Chapter 7; as such this has the potential to provide a test-bed for future development.

As set out above this research met at least for the most part the aims set out in 1, with perhaps the exception of fully making the case for the operationalisation of the user experience. As an interdisciplinary project, the balance of material across disciplines is difficult to strike and on reflection although the space dedicated to the existing SDC literature seems quite large it was deemed necessary as it was the least intuitive of the Chapters. Beck and Heidegger might be the more unfamiliar to audiences for this research, but the issues they discuss (risk, privacy, uncertainty, fear of technology, etc.) are issues that people can identify with, whereas, statistical methods such as those used to measure and mitigate disclosure risk and data utility can seem quite abstract.

It is important to also acknowledge that the central contribution of this research to the literature is not a new statistical disclosure method (as so many contributions have been) but an attempt to examine the process of statistical disclosure control in practice itself. Through this examination, this thesis grounds disclosure control in the theoretical framework of risks that are an integral part of the modern world and suggests a potential way forward for statistical disclosure control workflows to meet the demands that arise from the data deluge. The challenge for practitioners in this space is to develop rule-sets and tools that operate at scale while demonstrating an understanding of the public's concerns about their privacy and that do not become detached from that understanding during their implementation, while also acknowledging in an active way the needs of the data users who ultimately pursue these data to advance knowledge and perhaps contribute to the perceived public good.

8.2 Future Work

In the process of completing this research a number of avenues for future work have become apparent. In part, these ideas are drawn from the interactions with data users and data controllers, but some are also drawn from other gaps in the literature and the limitations identified in the conclusions above. The first avenue concerns the focus on e-Health and administrative data, in this case limited to data sources in Scotland. It could be assumed that the approach taken here would be applicable to data from across disciplines. However, further research using a broader range of case studies could yield interesting results, especially if the data subject was not an individual. For example, could data about businesses disclose sensitive information about its owners or employees?

Has was noted in the conclusions, the small pilot study on user experiences only goes part of the way to realising the potential for user experience to be considered fully in the development of disclosure control processes and data access more generally. More qualitative work could be undertaken to better understand the user experience, and more research could be dedicated to the problem of operationalisation raised in this thesis.

To shift focus from data to methods, one of the limitations of the current implementation of NIAH is the small range of algorithms for statistical disclosure control. As discussed in this research, few of the more complex methods in the literature have been developed into applications that can perform real-world analyses on different data types at scale without a significant amount of calibration from the user. Given further resources it would be novel to attempt such applications and be able to tie them to the theoretical framework discussed above in order to give data controllers consistency in the transition between value judgements and scientific analyses.

In order to address the potential scaling problem for the NIAH tool set, there could be merit in pursuing future work into how ‘big data’ methods for processing large volumes of data efficiently might be adapted for the purpose of SDC. In the implementation of NIAH presented in this research the data structure used results in a pattern of memory usage that scales linearly at approximately 28 times the input file size. One could distribute the task of sorting the data if is too large to be held in memory. A MapReduce model could map chunks of the dataset to individual machines to provide an in-memory sort of their respective chunks. Then a merge sort engine could be used to reconstruct the dataset. However, it is not the default behaviour of these approaches to enforce a stable sorting algorithm. This would prove problematic for the NIAH algorithm as it relies on records undergoing a stable sort.

Appendix A

Interview Materials: Information Sheet

University of Glasgow | National e-Science Centre

Information Sheet for 'Analysing stakeholder perspectives of statistical disclosure control in the context of cross-sector data-linkages.'

Purpose of the Research:

This data collection form part of my PhD research into statistical disclosure control (SDC) approaches and applications. The purpose is to collect and analyse the views of the various stakeholders in the process of linking personal data from various sources for analysis. These stakeholders include subject experts, data custodians, and members of the public and data analysts. This data will inform the development of statistical disclosure control methods and the implementation of SDC algorithms.

What is involved in Participating?

Key stakeholders are being asked to participate in semi-structured interviews about their experience of disclosure control and their views on how why and how it should be implemented.

Terms for Withdrawal:

All participants have a right to withdraw from the study at any point, and need provide no reason for doing so. Any data collected up to their withdrawal will be securely destroyed and not used in any part of the research.

Usage of the Data:

The data collected will be used in my PhD thesis as a means of providing the context for the implementation of statistical disclosure control algorithms, as well as influencing the development of software tools for such an implementation. From this the data could also be used in other published research outputs connected with this research.

The data will be submitted for archiving to a relevant repository such as the UK Data Archive. This is for the use of future researchers only if they agree to preserve the confidentiality of the information as requested in the accompanying consent form. It is possible to give consent for your data to be used only within this research project and not archived.

Confidentiality:

In the use of survey and interview data, ID numbers will be assigned to participants and all identifying data will be removed from published material. Non-anonymous data will only be used with the explicit permission of the participants involved.

Details of the Research:

This research is funded by the Economic & Social Research Council and is being carried out as part of a PhD research project at the University of Glasgow. More information can be found by contacting the researcher: Michael Comerford, National e-Science Centre, Rm246c Kelvin Building, University of Glasgow, G12 8QQ. Tel No. 0141 330 2598 email: comerm@dcs.gla.ac.uk

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Appendix B

Interview Materials: Consent Form

University of Glasgow | National e-Science Centre

Consent Form for 'Analysing stakeholder perspectives of statistical disclosure control in the context of cross-sector data-linkages.'

Please tick the appropriate boxes	**Yes**	**No**
Taking Part		
I have read and understood the project information sheet dated 26/09/12.	□	□
I have been given the opportunity to ask questions about the project.	□	□
I agree to take part in the project. Taking part in the project will include being interviewed and recorded (audio).	□	□
I understand that my taking part is voluntary; I can withdraw from the study at any time and I do not have to give any reasons for why I no longer want to take part.	□	□
Use of the information I provide for this project only		
I understand my personal details such as phone number and email address will not be revealed to people outside the project.	□	□
I understand that my words may be quoted in publications, reports, web pages, and other research outputs.	□	□
*Please choose **one** of the following two options:*		
I would like my real name used in the above	□	
I would **not** like my real name to be used in the above.	□	
Use of the information I provide beyond this project		
I agree for the data I provide to be archived at the UK Data Archive.	□	□
I understand that other genuine researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
I understand that other genuine researchers may use my words in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
So we can use the information you provide legally		
I agree to assign the copyright I hold in any materials related to this project to Michael Comerford.	□	□

________________________ ________________________ __________

Name of participant [printed] Signature Date

________________________ ________________________ __________

Researcher [printed] Signature Date

Project contact details for further information: Michael Comerford, National e-Science Centre, University of Glasgow. Tel. No. 0141 330 2958

Appendix C

Interview Materials: Interview Guide

Open Ended Questions for Researchers/Analysts/Users

- What experiences have you had working with linked data? Or large datasets in general?
- What level of detail would you expect from the data?
- Were there any issues in terms of access or disclosure control (i.e. access to outputs)?
- What expectations do you have of the linked data? What types of analysis do you commonly produce? What output do you regularly publish?
- Have you used a safe haven before?
  - If so what has your experience been like?
  - If not what are your expectations of this type of set up?
- How to do you prefer to develop your analysis? Do you produce SAS/SPSS/STATA code from looking at metadata or do you do much of the development while working with the data?
- Do you use any tools for data management or workflow documentation?
- Would a set of dummy data, be useful for developing early code before booking into a safe haven?
- What statistical disclosure control techniques have you come across in your own work? E.g. in publications, or FoI responses, etc.
- What format do you prefer metadata to be in? Do you have any good examples of metadata format? E.g. Wiki's, html, etc.

Bibliography

B. Adam, U. Beck, and J. v. Loon, editors. *The risk society and beyond: critical issues for social theory*. Sage, London, 2000.

M. Addis, J. Ferris, M. Greenwood, P. Li, D. Marvin, T. Oinn, and A. Wipat. Experiences with e-science workflow specification and enactment in bioinformatics. In *Proceedings of UK e-Science All Hands Meeting 2003*, pages 459–466, 2003.

Administrative Data Liaison Service. ADLS - administrative data liaison service: Safe researcher training. `http://www.adls.ac.uk/safe-researcher-training/`, a.

Administrative Data Liaison Service. ADLS - administrative data liaison service ESRC expression of interest: Prefabricated sensitive data secure room awards. `http://www.adls.ac.uk/adls-resources/esrc-sensitive-data-secure-rooms/`, b. Visited: 21/07/2014.

Administrative Data Taskforce (Technical Group). *Structuring the Administrative Data Research Network*. ESRC, 2013.

R. Agrawal and R. Srikant. Privacy-preserving data mining. In *ACM Sigmod Record*, volume 29, page 439450, 2000.

M. Aitken. Your data and health research: SHIP public workshops. `http://www.scot-ship.ac.uk/sites/default/files/Reports/Your_Data_and_Health_Research.pdf`, April 2012.

M Aitken, S Cunningham-Burley, and C Pagliari. Public responses to the scottish health informatics programme: preferences and concerns around the use of personal medical records in research. *Journal of Epidemiology and Community Health*, 65(Suppl 1):A27, 2011.

R. N. Allan. *Virtual Research Environments: From portals to science gateways*. Elsevier, 2009.

M. R. Andersen and H. H. Storm. Cancer registration, public health and the reform of the european data protection framework: Abandoning or improving european public health research? *European Journal of Cancer*, October 2013.

C. Andersson, A. Holmberg, I. Jansson, K. Lindgren, and P. Werner. Methodological experiences from a register-based census. In *JSM 2013 - Survey Research Methods Section*, 2013.

J. D. Angrist and A. B. Krueger. The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87(418):328–336, June 1992.

Argonne National Laboratory. New institute to tackle ”data tsunami” challenge. `http://www.anl.gov/articles/new-institute-tackle-data-tsunami-challenge`, April 2012. Visited: 25/10/2013.

D. Armstrong. The rise of surveillance medicine. *Sociology of Health and Illness*, 3(17): 393–404, 1995.

P. A. L. Ashfield-Watt, A. A. Welch, N. E. Day, and S. A. Bingham. Is ’five-a-day’ an effective way of increasing fruit and vegetable intakes? *Public Health Nutrition*, 7(02): 257–261, 2004.

M. Askari, R. Safavi-Naini, and K. Barker. An information theoretic privacy and utility measure for data sanitization mechanisms. In *Proceedings of the second ACM conference on Data and Application Security and Privacy*, CODASKY 2012, pages 283–294, New York, NY, USA, 2012. ACM.

S. L. Bartky. Heidegger and the modes of world-disclosure. *Philosophy and Phenomenological Research*, 40(2):212–236, 1979.

BBC. Police force loses memory stick. `http://news.bbc.co.uk/1/hi/scotland/edinburgh_and_east/7932228.stm`, March 2009. Visited:05/05/2014.

A. Beaulieu, A. Scharnhorst, and P. Wouters. Not another case study: A middle-range interrogation of ethnographic case studies in the exploration of e-science. *Science, Technology & Human Values*, 32(6):672–692, 2007.

U. Beck. *Risk society: towards a new modernity*. Sage, London, 1992.

U. Beck, W. Bonss, and C. Lau. The theory of reflexive modernization problematic, hypotheses and research programme. *Theory, Culture & Society*, 20(2):1–33, April 2003.

G. B. Bell and A. Sethi. Matching records in a national medical patient index. *Communications of the ACM*, 44(9):83–88, 2001.

J. Betham. *Panopticon or the inspection-house*. Dodo Press, 2008.

R. Bhopal, C. Fischbacher, C. Povey, J. Chalmers, G. Mueller, M. Steiner, H. Brown, D. H. Brewster, and N. Bansal. Cohort profile: Scottish health and ethnicity linkage study of 4.65 million people exploring ethnic variations in disease in scotland. *International Journal of Epidemiology*, 40(5):1168–1175, October 2011.

Big Brother Watch. Time for action on google's privacy policy. `http://www.bigbrotherwatch.org.uk/home/2013/04/time-for-action-on-googles-privacy-policy.html`, 2014. Visited: 02/05/2014.

J.A. Blaya, H. S. Fraser, and B. Holt. E-health technologies show promise in developing countries. *Health Affairs*, 29(2):244–251, 2010.

S. Bohle. What is e-science and how should it be managed? `http://www.scilogs.com/scientific_and_medical_libraries/what-is-e-science-and-how-should-it-be-managed/`, December 2013. Visited: 06/08/2014.

R. Boreham and R. Constantine. Understanding society, innovation panel wave 1. Technical report, National Centre for Social Research, December 2008.

J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 70–78, 2008.

D. A. Brown, P. R. Brady, A. Dietz, J. Cao, B. Johnson, and J. McNabb. A case study on the use of workflow technologies for scientific analysis: Gravitational wave data analysis. In *Workflows for e-Science*, pages 39–59. Springer, 2007.

R. Cameron. A sequential mixed model research design: Design, analytical and display issues. *International Journal of Multiple Research Approaches*, 3(2):140–152, 2009.

M. Carlson. Assessing microdata disclosure risk using the poisson-inverse gaussian distribution. *Statistics in Transition*, 5:901–925, 2002.

C. M. J. Charlton, D. T. Michaelides, B. Cameron, C. Szmaragd, R. M. A. Parker, H. Yang, Z. Zhang, and William J. Browne. Stat-JR software. Technical report, NCRM EPrints, 2012.

C. M. J. Charlton, D. T. Michaelides, B. Cameron, C. Szmaragd, R. M. A. Parker, H. Yang, Z. Zhang, William J. Browne, A.J. Frazer, H. Goldstein, K. Jones, G Leckie, and L. Moreau. Stat-JR, 2013. URL `http://www.bristol.ac.uk/cmm/software/statjr/`.

L. Cleveland, R. McCaa, S. Ruggles, and M. Sobek. When excessive perturbation goes wrong and why IPUMS-international relies instead on sampling, suppression, swapping, and other minimally harmful methods to protect privacy of census microdata. In *Privacy in Statistical Databases*, pages 179–187, 2012.

Committee for the Coordination of Statistical Activities. Microdata dissemination best practices. Technical report, United Nations Statistics Division, 2014.

Council for Science and Technology. Better use of personal information: opportunities and risks, 2005.

J. W. Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2013.

T. Dalenius and S. P. Reiss. Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6(1):73–85, January 1982.

F.K. Dankar, K.E. Emam, A. Neisa, and T. Roffey. Estimating the re-identification risk of clinical data sets. *BMC Medical Informatics and Decision Making*, 12(1):66, July 2012.

D. De Roure and C. Goble. myExperiment a web 2.0 virtual research environment. Technical report, University of Southampton, 2007.

D. De Roure, C. Goble, J. Bhagat, D. Cruickshank, A. Goderis, D. Michaelides, and D. Newman. myExperiment: Defining the social virtual research environment. In *IEEE Fourth International Conference on eScience, 2008*, pages 182–189, 2008.

P. P De Wolf, J. M Gouweleeuw, P. Kooiman, and L. Willenborg. *Reflections on PRAM*. Citeseer, 1999.

B. Debatin, J. P. Lovejoy, A-K. Horn, and B. N. Hughes. Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication*, 15(1):83–108, October 2009.

E. Deelman and Y. Gil. Managing large-scale scientific workflows in distributed environments: Experiences and challenges. In *Second IEEE International Conference on e-Science and Grid Computing 2006. e-Science'06.*, pages 144–144. IEEE, 2006.

Department of Health. Report on the review of patient-identifiable information. http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4068403, 1997.

T. Doherty, S. McCafferty, and J. Watt. Sociological classifications: The GESDE services for classifications involving occupations, educational qualifications and ethnicity (practical session). http://www.dames.org.uk/workshops/leeds2010/gesde_practical_8jun2010.pdf, June 2010. Visited: 31/07/2014.

J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14 (1):189–201, 2002.

J. Domingo-Ferrer and V. Torra. A critique of k-anonymity and some of its enhancements. In *ARES 08. Third International Conference on Availability, Reliability and Security*, pages 990–993, 2008.

P.S. Donaldson. The shakespeare interactive archive: New directions in electronic scholarship on text and performance. *Contextual Media: Multimedia and Interpretation*, page 103, 1997.

J. Drechsler, A. Dundler, S. Bender, S. Rassler, and T. Zwick. A new approach for disclosure control in the IAB establishment panel: Multiple imputation for a better data access. *AStA Advances in Statistical Analysis*, 92(4):439–458, 2008.

G.T. Duncan, M. Elliot, and J.-J. Salazar-Gonzalez. *Statistical confidentiality principles and practice*. Springer, New York, 2011.

The Economist. Technology: The data deluge. http://www.economist.com/node/15579717, February 2010.

K. El Emam and F. K. Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association JAMIA*, 15(5):627–637, 2008.

M. Elliot and A. Dale. Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics*, 14(Spring):6–10, 1999.

M. Elliot, A. Hundepool, E. S. Nordholt, J. L. Tambay, and T. Wende. Glossary on statistical disclosure control. In *Joint UNECE Eurostat Work Session on Statistical Data Confidentiality*, 2005a.

M. Elliot, A. Manning, K. Mayes, J. Gurd, and M. Bane. SUDA: A program for detecting special uniques. In *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, pages 353–362, 2005b.

M. Elliot, S. Lomax, E. Mackey, and K. Purdam. Data environment analysis and the key variable mapping system. In *Privacy in Statistical Databases*, pages 138–147, 2011.

M.J. Elliot, C.J. Skinner, and A. Dale. Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk. *Research in Official Statistics*, 1(2):53–67, 1998.

N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook friends: Social capital and college students use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

P.M. Eng, I. Kawachi, G. Fitzmaurice, and E. B. Rimm. Effects of marital transitions on changes in dietary and other health behaviours in US male health professionals. *Journal of Epidemiology and Community Health*, 59(1):56–62, January 2005.

England & Wales High Court. Department of health, r (on the application of) v information commissioner [2011] EWHC 1430 (admin). `http://www.bailii.org/ew/cases/EWHC/Admin/2011/1430.html`, April 2011.

ESRC. Secondary data analysis initiative. `http://www.esrc.ac.uk/research/skills-training-development/sdai/`, 2012. Visited: 05/05/2014.

EU Committee on Civil Liberties, Justice and Home Affairs. Draft report: on the proposal for a regulation of the european parliament and of the council on the protection of individual with regard to the processing of personal data and on the free movement of such data (general data protection regulation), December 2012.

European Commission. Proposal: On the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation), November 2011.

Eurostat. *Manual on disclosure control methods.* Office for Official Publications of the European Communities, Luxembourg, 1996.

Facebook. Data use policy. `https://en-gb.facebook.com/about/privacy/`, November 2013. Visited: 06/05/2014.

I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

S. E Fienberg and L. Willenborg. Introduction to the special issue: disclosure limitation methods for protecting the confidentiality of statistical data. *Journal of Official Statistics*, 14(4):337–345, 1998.

D. V. Ford, K. H. Jones, J.P. Verplancke, R. A. Lyons, G. John, G. Brown, C. J. Brooks, S. Thompson, O. Bodger, and T. Couch. The SAIL databank: building a national architecture for e-health research and evaluation. *BMC Health Services Research*, 9(1):157, 2009.

S. Garfinkel. *Database nation: the death of privacy in the 21st century*. O'Reilly Media, Inc., 2000.

A. Giddens. Risk and responsibility. *The modern law review*, 62(1):1–10, 1999.

A. Goderis, P. Li, and C. Goble. Workflow discovery: the problem, a case study from e-science and a graph-based solution. In *Web Services, 2006. ICWS'06. International Conference on*, pages 312–319. IEEE, 2006.

Google. Privacy policy & terms. `http://www.google.com/policies/privacy/`, March 2014. Visited: 06/05/2014.

J. M. Gouweleeuw, P. Kooiman, L. Willenborg, and P. P. De Wolf. Post randomisation for statistical disclosure control: Theory and implementation. *Journal of official statistics-stockholm*, 14:463–478, 1998.

T. Govani and H. Pashley. Student awareness of the privacy implications when using facebook. Last Accessed:2014-05-06, 2005.

B. Greenberg. Disclosure avoidance research at the census bureau. In *Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC*, pages 144–166, 1990.

B. Greenberg and L. Voshell. Relating risk of disclosure for microdata and geographic area size. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 450–455, 1990.

B.V. Greenberg and L.V. Zayatz. Strategies for measuring risk in public use microdata files. *Statistica Neerlandica*, 46(1):33–48, 1992.

T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

J. Gulliksen, B. Göransson, I. Boivie, S. Blomkvist, J. Persson, and Å. Cajander. Key principles for user-centred systems design. *Behaviour and Information Technology*, 22(6): 397–409, 2003.

P. Halfpenny and R. Procter. The e-social science research agenda. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1925):3761–3778, 2010.

M. E. Halstuk and B. F. Chamberlin. The freedom of information act 1966-2006: A retrospective on the rise of privacy protection over the public interest in knowing what the government's up to. *Communication Law and Policy*, 11(4):511–564, 2006.

M. Hammersley. Case study. In M. S. Lewis-Beck, A. Bryman, and T. Futing Liao, editors, *Encyclopedia of Social Science Research Methods*, pages 93–95. SAGE Publications, 2004.

Health Informatics Centre (Dundee). TASC safe haven at HIC. `http://medicine.dundee.ac.uk/health-informatics-centre/information-governance/safe-haven`. Visited: 05/12/2013.

M. Heidegger. *Being and time*. The library of philosophy and theology. SCM Press, London, 1962.

M. Heidegger. *Discourse on thinking: a translation of Gelassenheit*. Harper & Row, New York, 1969.

M. Heidegger. *Basic writings from 'Being and time' (1927) to 'The task of thinking' (1964)*. Routledge, London, 1978.

J. Heldal and S. Badina. Confidentiality protection of large frequency data cubes. In *Joint UNECE/Eurostat work session on statistical data confidentiality*, 2013.

A. J. G. Hey and A E Trefethen. The UK e-Science Core Programme and the Grid. *Future Generation Computer Systems*, 18(8):1017–1031, September 2002.

A. Hundepool and L. Willenborg. μ and τ-argus: software for statistical disclosure control. In *Third International Seminar on Statistical Confidentiality*, 1996.

A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P. P De Wolf. *Statistical disclosure control*. John Wiley & Sons, 2012.

Information Commissioner's Office. Determining what is personal data. Technical report, ICO, December 2012.

V. S. Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288, 2002.

M. Janssen, Y. Charalabidis, and A. Zuiderwijk. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4):258–268, 2012.

I. Jansson. Issues and plans for the disclosure control of the swedish census 2011. In *Workshop on Statistical Disclosure Control of Census Data, Luxembourg*, 2012.

M. A Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14 (57):491–498, March 1995.

T. D. Jick. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, pages 602–611, 1979.

H. Jones and J. H. Soltren. Facebook: Threats to privacy. Technical report, Project MAC: MIT Project on Mathematics and Computing, 2005.

K. P. Keraminiyage, R. P. Haigh, and R. D. G. Amaratunga. Achieving success in collaborative research: the role of virtual research environments. *Journal of Information Technology in Construction (ITCon)*, 14:59–69, 2009.

I. Kerr and J. Earle. Prediction, preemption, presumption: How big data threatens big picture privacy. *Stanford Law Review Online*, 66:65, 2013.

J. J. Kim. A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the section on survey research methods*, pages 303–308, 1986.

N. Kompridis. *Critique and disclosure: Critical theory between past and future*. MIT Press, 2011.

N. Kroes and A. Vassiliou. Promoting coding skills in europe is part of the solution to youth unemployment. `ec.europa.eu//digital-agenda/en/news/promoting-coding-skills-europe-part-solution-youth-unemployment`, July 2014.

D. Lambert. Measures of disclosure risk and harm. *journal of official statistics stockholm*, 9:313–313, 1993.

P. Lambert. Dealing with data on ethnicity: Principles and practice. `http://www.dames.org.uk/gemde/workshops/28jan2010/presentations/1_data_intro.pdf`, 2010. Visited: 23/10/2013.

P. Lambert and V. Gayle. Data management and standardisation: A methodological comment on using results from the UK research assessment exercise 2008. Technical Paper 3, University of Stirling, 2008.

G. Laurie and N. Sethi. Information governance of use of health-related data in medical research in scotland: Towards a good governance framework. Technical Report 2012/13, Edinburgh School of Law, 2012.

N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE 23rd International Conference on Data Engineering, ICDE*, pages 106–115, 2007.

T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526, 2009.

Liberty. Liberty's submission to the joint committee on the draft communcations data bill. `http://www.liberty-human-rights.org.uk/pdfs/policy12/liberty-submission-to-the-draft-communications-data-bill-committe` `pdf`, August 2012. Visited: 17/12/2012.

C. Lynch. Big data: How do your data grow? *Nature*, 455(7209):28–29, 2008.

D. Lyon. *Surveillance as Social Sorting: Privacy, risk, and digital discrimination*. Psychology Press, 2003.

R. A. Lyons, K. H. Jones, G. John, C. J. Brooks, J.P. Verplancke, D. V. Ford, G Brown, and K. Leake. The SAIL databank: linking multiple health and social care datasets. *BMC Medical Informatics and Decision Making*, 9(1):3, 2009.

A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1 (1):3, 2007.

E. Mackey and M. Elliot. Understanding the data environment. *XRDS: Crossroads, The ACM Magazine for Students*, 20(1):36–39, 2013.

P. Mansell. New e-health research centres to optimise UK patient data resources. `http://www.pharmatimes.com/article/12-08-07/New_e-health_research_centres_to_optimise_UK_patient_data_resources.aspx`, August 2012.

J. Mares and V. Torra. Clustering-based categorical data protection. In Josep Domingo-Ferrer and Ilenia Tinnirello, editors, *Privacy in Statistical Databases*, number 7556 in Lecture Notes in Computer Science, pages 78–89. Springer, January 2012.

C. Marsh, C. J. Skinner, S. Arber, B. Penhale, S. Openshaw, J. Hobcraft, D. Lievesley, and N. Walford. The case for samples of anonymized records from the 1991 census. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(2):305–340, 1991.

D. Mascalzoni, B. M. Knoppers, S. Aym, M. Macilotti, H. Dawkins, S. Woods, and M. G. Hansson. Rare diseases and now rare data? *Nature Reviews Genetics*, 14(6):372–372, June 2013.

D. A. McAllister, J. R. Morling, C. M. Fischbacher, W. MacNee, and S. H. Wild. Socioeconomic deprivation increases the effect of winter on admissions to hospital with COPD: retrospective analysis of 10 years of national hospitalisation data. *Primary Care Respiratory Journal*, 22(3):296, 2013.

R. McCaa, K. Muralidhar, R. Sarathy, M. Comerford, and A. Esteve. Analytical tests of controlled shuffling to protect statistical confidentiality and privacy of a ten per cent household sample of the 2011 census of ireland for the IPUMS-international database. In *Privacy in Statistical Data 2014, LNCS*, October 2013.

S. McCafferty, T. Doherty, RO Sinnott, and J. Watt. e-infrastructures supporting research into depression, self-harm and suicide. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1925):38–45, 2010.

R. H. McGuckin and S. V. Nguyen. Use of'surrogate files' to conduct economic studies with longitudinal microdata. In *Proceedings of the Fourth Annual Research Conference, Bureau of the Census, Washington, DC: US Department of Commerce*, pages 193–209, March 1988.

G. T. McKee. Virtual robotics laboratory for research. In *Proceeding of The International Society of Optics and Photonics (SPIE)*, volume 2589, pages 162–171, 1995.

S. Menard. Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1):17, February 2000.

Microsoft. The big bang: How the big data explosion is changing the world, Feburary 2013. URL `http://www.microsoft.com/en-us/news/features/2013/feb13/02-11bigdata.aspx`. Visited: 25/10/2013.

K. Muralidhar and R. Sarathy. Data shuffling: A new masking approach for numerical data. *Management Science*, 52(5):658–670, May 2006.

K. Muralidhar, R. Sarathy, and R. Dandekar. Why swap when you can shuffle? a comparison of the proximity swap and data shuffle for numeric data. In *Privacy in Statistical Databases*, pages 164–176, 2006.

A. Naska, V. Vasdekis, A. Trichopoulou, S. Friel, I. U. Leonhauser, O. Moreiras, M. Nelson, A. M. Remaut, A. Schmitt, and W. Sekula. Fruit and vegetable availability among ten european countries: how does it compare with thefive-a-day'recommendation? *British journal of nutrition*, 84(4):549–556, 2000.

National Cancer Institute. SEER research data agreement. http://seer.cancer.gov/data/sample-dua.html. URL http://seer.cancer.gov/data/sample-dua.html.

National Cancer Institute. Accessing the data - SEER datasets. http://seer.cancer.gov/data/access.html, July 2014.

National Records of Scotland Census Division. Census statistical disclosure control - the use of record-swapping to protect data confidentiality. In *The Population and Migration Statistics (PAMS) Conference*, 2013.

D. Nelson. *Pursuing Privacy in Cold War America*. Columbia University Press, 2002.

NHS England. The care.data programme better information means better care. http://www.england.nhs.uk/ourwork/tsd/care-data/. Visited: 20/08/2014.

NHS National Services Scotland. eDRIS user agreement. http://www.isdscotland.org/Products-and-Services/eDRIS/Docs/eDRIS-User-Agreement-V7%2820140618%29.pdf, 2013a. Visited: 20/05/2013.

NHS National Services Scotland. National PAC for NHS scotland - consultation document. http://www.nhsnss.org/uploads/publications/130320_Consultation_Document_version_1.0.pdf, March 2013b. Visited: 05/05/2014.

NHS Scotland ISD. Statistical disclosure control protocol. Technical Report 2.2, NHS Scotland ISD, January 2012a.

NHS Scotland ISD. Statistical disclosure control protocol. Technical Report 2.2, NHS Scotland ISD, January 2012b.

R Nikolov and I. Nikolova. A virtual environment for distance education and training. In *IFIP WG3.6 Conference*, 1996.

H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1), 2004.

E. Nordholt. ”statistical disclosure control from a users point of view.”. In *ETK/NTTS 2001 conference in Crete*, 1999.

Northern Ireland Longitudinal Study Research Support Unit. NILS policies, 2013.

Office for National Statistics. Statistical disclosure control for 2011 census. Technical report, ONS, 2011.

Office for National Statistics. The census and future provision of population statistics in england and wales: Recommendation from the national statistician and chief executive of the UK statistics authority. `http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-report-on-autumn-2013-consultation--and-recommendatio national-statisticians-recommendation.pdf`, March 2014. Last Accessed:2014-05-05.

Ohio State University. Cross-national equivalent file. `http://cnef.ehe.osu.edu/`. Visited: 29/11/2014.

F. A. Olafson. *Heidegger and the ground of ethics: a study of Mitsein*. Modern European philosophy. Cambridge University Press, Cambridge, 1998.

G. Paass. Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics*, 6(4):487–500, October 1988.

C. Pagliari, D. Detmer, and P. Singleton. Potential of electronic personal health records. *BMJ: British Medical Journal*, 335(7615):330–333, 2007.

M. C. Ploem, M. L. Essink-Bot, and K. Stronks. Proposed EU data protection regulation is a threat to medical research. *British Medical Journal*, 346(4):f3534–f3534, July 2013.

M. Plummer. JAGS - just another gibbs sampler. `http://mcmc-jags.sourceforge.net/`. Visited: 11/07/2014.

R. Polt. *Heidegger: an introduction*. Routledge, London, 2003.

J. Pomerleau, K. Lock, C. Knai, and M. McKee. Interventions designed to increase adult fruit and vegetable intake can be effective: a systematic review of the literature. *The Journal of nutrition*, 135(10):2486–2495, 2005.

K. Purdam and M. Elliot. A case study of the impact of statistical disclosure control on data quality in the individual UK samples of anonymised records. *Environment and Planning A*, 39(5):1101–1118, 2007.

M. Ranganathan and R. Bhopal. Exclusion and inclusion of nonwhite ethnic minority groups in 72 north american and european cardiovascular cohort studies. *PLoS medicine*, 3(3): 44, 2006.

J. Rasbash, C. Charlton, W.J. Browne, M. Healy, and B. Cameron. MLwiN. `http://www.bristol.ac.uk/cmm/software/mlwin/`, 2009.

V. Rastogi, D. Suciu, and S. Hong. The boundary between privacy and utility in data publishing. In *Proceedings of the 33rd international conference on Very large data bases*, pages 531–542, 2007.

J. P Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics Stockholm*, 18(4):531–544, 2002.

J. P Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003.

J. P. Reiter. Using CART to generate partially synthetic public use microdata. *journal of official statistics stockholm*, 21(3):441, 2005.

J. P. Reiter and R. Mitra. Estimating risks of identification disclosure in partially synthetic data. Working Paper M08/07, University of Southampton, 2008.

J.P. Reiter, A. Oganian, and A.F. Karr. Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis*, 53 (4):1475–1482, February 2009.

D. B. Rubin. Multiple imputation for nonresponse in surveys. *Wiley Series in Probability and Statistics*, 1987.

L. Salayandia, P. P. Da Silva, A. Q. Gates, and F. Salcedo. Workflow-driven ontologies: An earth sciences case study. In *e-Science and Grid Computing, 2006. e-Science'06. Second IEEE International Conference on*, pages 17–17. IEEE, 2006.

P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *PODS*, volume 98, page 188, 1998a.

P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, Technical report, SRI International, 1998b. URL `http://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf`.

M. S. Sarwar, T. Doherty, J. Watt, and R. O. Sinnott. Towards a virtual research environment for language and literature researchers. *Future Generation Computer Systems*, 29(2):549–559, 2013.

J. L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, 1997.

J. L Schafer and J. W Graham. Missing data: Our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

J. L. Schafer and M. K. Olsen. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571, 1998.

Scottish Government. Home care census, April 2003. URL http://www.scotland.gov.uk/Topics/Statistics/Browse/Health/HomeCareCensus.

Scottish Government. Statistical disclosure risk and statistical disclosure control. Technical report, Scottish Government, 2007.

Scottish Government. Health, social care & housing - background. http://www.scotland.gov.uk/Topics/Statistics/Browse/Health/Datalinking/HealthSocialCareandHousin, September 2011a. Visited: 06/04/2012.

Scottish Government. Linking social care, housing and health data: Social care clients' and patients' views. http://www.scotland.gov.uk/Publications/2011/09/20085846/0, September 2011b.

Scottish Government. Variations in the experiences of inpatients in scotland: Analysis of the 2010 scottish inpatient survey, August 2011c. URL http://www.scotland.gov.uk/Publications/2011/08/29131615/0.

Scottish Government. Children looked after statistics 2009-10. http://www.scotland.gov.uk/Publications/2011/02/18105352/0, February 2011d.

Scottish Government. Children's social work statistics, 2011-12. http://www.scotland.gov.uk/Publications/2013/03/5229, March 2013.

Scottish Health Informatics Programme (SHIP). Homepage. http://www.scot-ship.ac.uk/. Visited: 15/06/2012.

Y. Shafranovich. Common format and MIME type for comma-separated values (CSV) files. http://tools.ietf.org/html/rfc4180, 2005. Visited: 11/07/2014.

N. Shlomo and C. J. Skinner. Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *The Annals of Applied Statistics*, 4 (3):1291–1310, 2010.

D. Silber. *The case for eHealth*. European Institute of Public Administration, 2003.

L. Silva. PLoS new data policy: Public access to data. http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/, February 2014. Visited: 05/05/2014.

R. O. Sinnott, A. J. Stell, and O. Ajayi. Initial experiences in developing e-health solutions across scotland. In *Workshop on Integrated Health Records: Practice and Technology*, March 2006.

R. O. Sinnott, A. J. Stell, and O. Ajayi. Supporting grid-based clinical trials in scotland. *Health Informatics Journal*, 14(2):79–93, June 2008a.

RO Sinnott, O. Ajayi, A. Stell, and A. Young. Towards a virtual anonymisation grid for unified access to remote clinical data. *Global healthgrid: e-Science meets biomedical informatics: proceedings of Healthgrid 2008*, 138:90, 2008b.

C. J. Skinner and M. Elliot. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):855–867, 2002.

C. J. Skinner and D. J. Holmes. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14(4):361–372, 1998.

C. J. Skinner and N. Shlomo. Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association*, 103(483):989–1001, 2008.

C. J. Skinner, C. Marsh, S. Openshaw, and C. Wymer. Disclosure avoidance for census microdata in great britain. In *Proceedings of US Bureau of the Census Annual Research Conference, US Bureau of the Census, Washington*, pages 131–143, 1990.

D. J. Solove. *Understanding privacy.* Harvard University Press, 2008.

StataCorp. Stata statistical software. `http://www.stata.com/`, 2013.

P. Steel and J. Sperling. The impact of multiple geographies and geographic detail on disclosure risk: Interactions between census tract and ZIP code tabulation geography. Technical report, Bureau of Census, 2001.

R. Steinbrook. Personally controlled online health data-the next big thing in medical care? *New England Journal of Medicine*, 358(16):1653, 2008.

A. Stell, R. Sinnott, O. Ajayi, and J. Jiang. Designing privacy for scalable electronic healthcare linkage. In *International Conference on Computational Science and Engineering, 2009. CSE '09*, volume 3, pages 330–336. IEEE, August 2009.

K. Stokes and V. Torra. n-confusion: a generalization of k-anonymity. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 211–215. ACM, 2012.

K. J. Strandburg. *Privacy and Technologies of Identity: A cross-disciplinary conversation.* Springer Science & Business Media, 2006.

L. Sweeney. *Computational disclosure control: A Primer on Data Privacy Protection*. PhD thesis, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science., 2001.

C Szongott, B. Henne, G. von Voigt, et al. Big data privacy issues in public social media. In *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on*, pages 1–6. IEEE, 2012.

L. Tan, P. Lambert, K. Turner, J. Blum, V. Gayle, S. B. Jones, R. O. Sinnott, and G. Warner. Enabling quantitative data analysis through e-infrastructure. *Social Science Computer Review*, 2009.

O. Tene and J. Polonetsky. Privacy in the age of big data: a time for big decisions. *Stanford Law Review Online*, 64:63, 2012.

The MethodBox Project. Methodbox. `https://www.methodbox.org/`, 2011.

R. Thomas and M. Walport. *Data sharing review*. Ministry of Justice, London, 2008.

J. Townsend, P. Roderick, and J. Cooper. Cigarette smoking by socioeconomic group, sex, and age: effects of price, income, and health publicity. *BMJ*, 309(6959):923–927, October 1994.

T. M Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In *Proceedings of the 22nd International Conference on Data Engineering*, pages 94–94, 2006.

UK Clinical Research Collaboration. *UKCRC R&D Advisory Group to Connecting for Health Report of Research Stimulations*. UKCRC, London, 2007.

UK Data Archive. Terms and conditions of access. `http://ukdataservice.ac.uk/get-data/how-to-access/conditions.aspx`. Visited: 15/07/2014.

UK Data Archive. Annual report 2001-2002: Serving and preserving digital resources for research and learning. `http://www.data-archive.ac.uk/media/54749/ukda-annualreport20012002.pdf`, 2002.

UK Data Archive. Secure lab breaches penalties policy. `http://ukdataservice.ac.uk/media/176861/UKDA142_SDS_SecurityBreaches_public.pdf`, 2014.

US Federal Committee on Statistical Methodology. Report on statistical disclosure and disclosure-avoidance techniques. Technical report, Office of Federal Statistical Policy and Standards, 1978.

A. J. Webster. Risk and innovative health technologies: calculation, interpretation and regulation. *Health, Risk & Society*, 4(3):221–226, 2002.

Wellcome Trust. Qualitative research into public attitudes to personal data and linking personal data. `http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtp053205.pdf`, July 2013. Visited: 05/05/2014.

L. Willenborg and T. De Waal. *Statistical disclosure control in practice*. Number 111 in Lecture Notes in Statistics. Springer, 1996.

L. Willenborg and T. De Waal. *Elements of statistical disclosure control*. Springer, 2001.

W. E Winkler. Masking and re-identification methods for public-use microdata: Overview and research problems. In *Privacy in Statistical Databases*, pages 519–519, 2004.

K. M. Witkowski. Disclosure risk of geography attributes: The role of spatial scale, identified geography, and measurement detail in public-use files. Technical Report No. 2, ICPSR Working Papers Series, 2008.

W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata protection. In *Inference control in statistical databases*, pages 135–152. Springer, 2002.

A. M Zaslavsky and N. J Horton. Balancing disclosure risk against the loss of nonpublication. *Journal of Official Statistics Stockholm*, 14:411–420, 1998.

L. Zayatz, P. Massell, and P. Steel. Disclosure limitation practices and research at the US census bureau. *Netherlands Official Statistics*, 14(Spring):26–29, 1999.

M. E. Zimmerman. *Heidegger's confrontation with modernity: technology, politics, and art*. Indiana University Press, Bloomington, 1990.