Ensuring the ethical use of Big Data: lessons from secure data access

Deborah Wiltshire, Seraphim Alvanides

PII: S2405-8440(22)00269-9

DOI: https://doi.org/10.1016/j.heliyon.2022.e08981

Reference: HLY 8981

To appear in: *HELIYON*

Received Date: 30 October 2021

Revised Date: 6 February 2022

Accepted Date: 14 February 2022

Please cite this article as: Wiltshire, D., Alvanides, S., Ensuring the ethical use of Big Data: lessons from secure data access, *HELIYON*, https://doi.org/10.1016/j.heliyon.2022.e08981.

Ensuring The Ethical Use of Big Data: Lessons from Secure Data Access

Ensuring the ethical use of big data: Lessons from secure data access

Authors

Dr. Deborah Wiltshire, GESIS Leibniz Institute for the Social Sciences, Germany
https://orcid.org/0000-0001-6533-2426
Dr. Seraphim Alvanides, GESIS Leibniz Institute for the Social Sciences Germany and Northumbia University, UK https://orcid.org/0000-0003-4905-4109

Abstract

Big data holds great potential for research and for society, large volumes of varied data can be produced and made available to researchers much faster compared to 'traditional' data. Whilst this potential is recognized, there are ethical concerns which users of big data must consider. With the volume and variety of information in big data, comes a greater risk of disclosure. Researchers and data access services working with highly detailed and sensitive, secure data have grappled with this for many years. The sector has developed both ethical frameworks and statistical disclosure control techniques which could be utilized by those working with big data. We discuss the challenges, present some of the frameworks and techniques and conclude with recommendations for secure data access of big data.

Keywords: big data, secure data access, statistical disclosure control

Introduction

There is great potential in big data, the potential for making new discoveries made possible for the first time by vast amounts of data. With the emergence of new forms of data, has come new ways of thinking about and analyzing data, requiring new platforms for analysis. Big data is generated in higher frequencies than other forms of data, such as from social surveys and national censuses that can take months even years to be

made available to researchers. For researchers used to navigating the various, sometimes lengthy, application processes for other forms of data, the scale and speed of production and the ease of access make big data an attractive prospect to those with the computational skills and power to handle it.

There is no single consensus on what makes data 'big', but a common way of thinking about big data is that it consists of multiple data sources that have been combined or explicitly linked to create a data source of significant size. Big data can be thought of as having several key characteristics: volume, variety and speed (Soria-Comas, 2016, Schroeder, 2014). Volume is self-explanatory, it refers to the size of the dataset, formed from multiple sources of data combined through some linking variable. These are much larger than social survey datasets and require significantly more computational power (Sfetu, 2019). The combination of data sources leads to the second characteristic – variety. A big data set will contain information on many different aspects of people's lives. For example, digital trace data might combine information on retail transactions, location histories collected using a mobile phone's GPS, websites visited and so on. The final characteristic, velocity refers to the increased speed of data collection and processing.

That big data offers great potential is certain, but it is not a unanimously positive picture. Like with all forms of data there are challenges and concerns with big data, and these have been widely discussed[1]. Many ethical issues have been raised in relation to big data, around issues of consent and privacy. These ethical issues are complex, and a complete overview is not attempted here. Neither is the goal to provide a complete instruction guide on how to use big data ethically. The focus of this article is on disclosure risk (the risk that individual data subjects will be identified) as a key ethical issue of big data and exploring what lessons can be applied from the experiences of secure data access services.

[1] For example of a broad discussion on the potential and challenges of big data, see Rob Kitchin's book 'The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences, 2014, Sage Publications

Many secure data access services exist across the world. These services specialize in making highly detailed, sensitive data from administrative, official sources as well as from large scale social surveys. Due to the detail and sensitivity of the data, these data sets are potentially disclosive. That is, there is a risk of individual data subjects being reidentified from the use of the data. Over many years secure data access services have developed infrastructures and techniques to ensure the safe use of these data so it is a natural place to turn when considering how to guard against disclosure in big data. This article focuses on secure data access in the UK, although the approaches presented here mirror those applied across Europe and the US.

Disclosure risk as a legal issue

Data protection legislation stipulates that data controllers and those processing data have a legal responsibility to prevent the disclosure of the identity of data subjects. The General Data Protection Regulations (GDPR) which applies across the EU and the UK currently states that:

'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person; (GDPR Article 4, Definitions 13)[2].

It further states in recital 39 that

"Personal data should be processed in a manner that ensures appropriate security and confidentiality of the personal data…"[3]

[2] Art. 4 GDPR – Definitions - General Data Protection Regulation (GDPR) (gdpr-info.eu)
[3] Recital 39 - Principles of Data Processing - General Data Protection Regulation (GDPR) (gdpr-info.eu)

When these stipulations are applied to secure data, it should not be possible to learn anything about a data subject without having direct access to the data itself (Zwitter 2014, Dwork 2006).

For those collecting and handling data, the GDPR triggered a review and update of their practices, especially for large organisations collecting and selling data. For researchers, the new legislation has encouraged more thought about privacy and disclosure which is a positive development in the bid to make data more accessible and to ease the sharing of data across international borders (Meijeringa, et. al., 2020).

Whilst the legislation is extensive and may arguably be sufficient to deal with the issues associated with the ever increasing availability of data, its very length and complexity is problematic. Largely this is because with complexity comes ambiguity and few researchers and data professionals have the legal training to navigate this ambiguity. So the GDPR is open to interpretation across different countries, organisations and researchers.

There remain many areas that are unclear. For example, whilst the legislation gives a description of personal data, a description which centres around identification, how exactly we decide what data is personal and in need of protection is less clear. A legal report published in 2016 offers the following clarification, suggesting that data is considered anonymous if identification is:

“practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant” (Case C-582/14 Breyer v Bundesrepublik Deutschland, 2016, page 9)

But how do we determine whether identification is ‘practically impossible’ and what does that mean? This is still open to interpretation.

Until now, when carrying out statistical disclosure control, assumptions have long been in place that those checking research outputs cannot reasonably be expected to factor in all possible additional data that might be available outside of the research dataset.

The legal report for Breyer v Bundesrepublik Deutschland (Case C-582/14 Breyer v Bundesrepublik Deutschland, 2016, page 8) calls these assumptions into question:

"its wording suggests that, for information to be treated as 'personal data' within the meaning of Article 2(a) of that directive, it is not required that all the information enabling the identification of the data subject must be in the hands of one person."

This is particularly problematic for researchers working with big data which often combines multiple sources of information. Much more education may be required for data professionals and researchers alike.

It would be misleading to conclude that the legislation is simply insufficient to handle forms of data such as big data, and to some degree unfair. There is no way to definitively define key concepts such as level of disclosure risk and 'practically impossible', but some clearer guidance would be beneficial.

How does big data differ compared to "traditional" data?

Concerns about the risk of disclosure are long standing and risk is omnipresent also with "traditional" types of data, resulting from social surveys, population censuses and statutory requirements. However, the majority of administrative and survey data is made available only after careful anonymization and statistical disclosure control to ensure that the level of detail is such that there should not be a risk of re-identification. More detailed, disclosive data is accessible only under strict conditions through secure data access services. In addition, users of administrative data have to accept usage agreements stipulating that identification of individuals (or even speculation of potential identification) will be prosecuted. This ensures that, to date, no breaches leading to re-identification have occurred, resulting in a successful and robust system.

A similar approach is beneficial when working with big data. But where big data differs is in the scale and variety of information available about an individual. This is important as with increasing levels of detail, we see the risk of disclosure increasing and this is

problematic (Soria-Comas 2016). Zwitter (2014) extends this by looking at the way lives become laid out in minute detail through our digital behavior and engagement with social media. With this transparency, the level of effort required for identifying individuals in data is much reduced.

In an example discussed by Duhigg (2012) a large retailer used big data to predict pregnancies for a target advertising campaign. They identified a college student as being pregnant, sending her coupons, disclosing her pregnancy to her family who had been unaware. The same problem can occur with the identification of groups. There are many examples where retailers and political parties have used big data to target specific groups and steer people's behaviors (Zwitter, 2014). To this end, Hand (2018) clarifies that it is not the data itself that is problematic, but how they are used and to what end. In this way, disclosure can be framed not just as a legal issue but also as an ethical one.

First, we need to consider what we mean by disclosure. Similarly, to attempts to define big data, disclosure may be defined or categorized in a number of ways.

Duncan & Lambert (1989) identified four types of disclosure[4]:

1. Identity disclosure
2. Attribute disclosure
3. Inferential disclosure
4. Population disclosure

When people are asked what they think disclosure means, often the first response will be that disclosure means the identity (i.e., name) of an individual becomes known[5]. This is identity disclosure, the unmasking of a data subject's identity which is of course, legally problematic. Direct access to the data is not necessarily required for identity disclosure to occur. For example, suppose analytic results from a social survey in the

[4] Identity and attribute disclosure are also discussed in Lambert, 1993

[5] Source: During training on data protection and statistical disclosure delivered by the author, attendees are asked to consider that they think disclosure means. To reveal the identity of an individual is the most common response.

form of tabular data are published that show a single female black dentist. If knowledge is obtained that such an individual took part in the survey, then she can easily be identified (Lambert 1993, 315). Attribute identity moves beyond the identification of a data subject, it is considered that attribute disclosure has occurred when information has been revealed about that person that was previously unknown. For example, an individual's income might be disclosed, or it might be disclosed that they have been diagnosed with a particular cancer. Disclosure is not limited to the disclosure of specific information in Duncan and Lambert's (1989) disclosure categories. Inferential disclosure may also occur where information is inferred about an individual based on analytic results, such as tabular data, that are published. This can occur even if specific information about the individual is not released, the inference also does not need to be entirely accurate. The final type of disclosure discussed by Duncan and Lambert (1989) is population disclosure where information is disclosed about a population or group rather than an individual. They explain this as considering the relationship between employee characteristics and salary as confidential rather than just the salary of an individual.

Whilst disclosure is a legal issue and guarding against disclosure is the legal responsibility of those working with data, it extends beyond the law. There are clear ethical concerns around disclosure, specifically centered around the harm potentially caused by disclosing something about an individual. In practice when using data and publishing analytical results, identity disclosure and attribute disclosure often go hand in hand. Lambert (1993) argues that the disclosure is the re-identification of an individual while the harm of disclosure depends on which attribute is disclosed. The information attributed to the data subject does not need to be accurate to cause harm to that individual, although the consequences can differ between true and false attribution.

There is considerable overlap between law and ethics. Whilst disclosure is often discussed (correctly) in terms of legal and data security considerations, it should also be

framed as the responsibility of researchers to ensure their use of data is ethical and does not cause harm[6]. This also applies to researchers who use big data.

The ethical use of big data

For the purpose of this article, ethical use is defined broadly as ensuring that no harm is caused by the use of big data for research. Whilst harm can be caused to many actors in the research process, such as the data source and wider research community, the focus here is on the potential for harm to the data subject.

Many organizations and authors have produced ethics criteria for research[7]. Whilst these vary depending on the academic discipline and data type, they have commonalities:

1. The project or research should have an element of public good or benefit
2. Data security and confidentiality must be ensured
3. Data subjects should not be identified or harmed as a result of the research
4. The research community should demonstrate trustworthiness
5. Research methodology must be robust and produce statistically valid findings

Existing ethics frameworks can demonstrate consideration of these five areas. For example, researchers defining the usefulness or merit that comes from their research (Kassner 2017) and data must be kept secure and results must be robust (Drew, 2016). Metcalf (2014) states that vulnerable people should not be harmed by the research project, and that researchers should demonstrate their trustworthiness so as not to need strict controls.

[6] See Code of Conduct | Data Science Association (datascienceassn.org) for an example of a Big Data code of conduct, https://www.datascienceassn.org/code-of-conduct.html

[7] Janeja discusses this is her blog Do No Harm: An Ethical Data Life Cycle | S&T Policy FellowsCentral (aaaspolicyfellowships.org), April 2019, https://www.aaaspolicyfellowships.org/blog/do-no-harm-ethical-data-life-cycle

The big data world is not idle in this area and much discussion occurs around ethics and ethical data use (Bishop 2017). In the pursuit of an ethical framework, big data could benefit from learning from the social and administrative data world who have several well-established ethical frameworks. More recently the UK Statistics Authority have made considerable investment in the ethical use of secure or legally controlled data. In the quest for consistent, ethical research practice, the UK Statistics Authority have developed a comprehensive framework: a self-assessment tool for researchers to use to conduct a thorough ethics assessment of their research projects. It is a mandatory part of the application process for accessing secure data from the Office for National Statistics but is recommended for all projects that use secondary data sources.

It covers 6 main principles, which include 21 items for researchers to assess. The 6 principles cover:

1. The use of data has clear benefits for users and serves the public good
2. The data subject's identity (whether person or organization) is protected, information is kept confidential and secure, and the issue of consent is considered appropriately
3. The risks and limits of new methods and/or technologies are considered and there is sufficient human oversight so that methods employed are consistent with recognized standards of integrity and quality.
4. Data used and methods employed are consistent with legal requirements such as Data Protection Legislation, the Human Rights Act 1998, the Statistics and Registration Service Act 2007 and the common law duty of confidence
5. The views of the public are considered in light of the data used and the perceived benefits of the research
6. The access, use and sharing of data is transparent, and is communicated clearly and accessibly to the public

Whilst this framework has been developed with survey and administrative data use in mind, these principles are designed to be embedded into good research practice and

would serve as a strong foundation for building a framework for big data research. The remaining article focuses on the first two principles of the UK Statistics Authority ethics framework as these are most directly relevant to discussions around disclosure risk and harm. The discussion will focus on the procedures in UK secure data services and personal experiences from delivering those services.

Principle 1: Public good and avoiding causing harm

To meet principle 1, researchers are asked to consider whether their research is in the public good and whether there is potential for harm to respondents from their use of the data and the publication of their results. In the UK greater emphasis is on researchers to demonstrate that not only is their proposed research feasible, but it will provide some public good. In practice this is addressed through the assessment of the project proposal that researchers submit during the application for data access. For secure data access in the UK researchers have to submit applications outlining their research, demonstrating both the public good and how their proposed methodology will enable that public good to be realized. These applications undergo robust review by data access committees made up of experienced researchers and data access professionals with expertise in the disciplinary field who consider the feasibility, the potential for public good and the potential harm of the proposed research.

Researchers generally state that their research is in the public good, although Ritchie and Welpton (2011) suggest that most projects are primarily for the researchers' benefit as few projects are policy commissioned. However, they point out that this does not mean that the public good criteria is not met. All research adds to the pool of knowledge which can be very much in the public good and one way to ensure that a research project is in the public good is through the publication of research findings (Ritchie & Welpton 2011). The publication of results is another area of potential harm which must be considered: disclosure risk which will be discussed later and harm through distress, discrimination or stigmatization of the data subjects. Therefore, consideration is needed about whether the publication of their could cause harm to the data subjects and the

wider research community[8]. What then are the potential harms that could be identified in project proposals?

In terms of the harms to the research community, whilst it is in the interest of society to allow research to go-ahead, poor-quality research can damage the reputation of the data owner, the survey or data source and indirectly the public. For example, if the methodology is not robust, conclusions may be drawn from results that are not statistically valid leading to poor policy decisions (Desai, Ritchie & Welpton 2016). In terms of direct harms to the data subjects, this can be demonstrated by looking at the criteria used by METADAC (Managing Ethico-social, Technical and Administrative issues in Data Access)[9], the data access committee that considered applications for genomic and phenotype data and biological samples collected from the UK cohort studies, to assess and approve biomedical and biosocial research projects. The relevant criteria to this discussion are listed here with key points highlighted:

1. The application **does not risk producing information that may allow individual study participants to be identified**.
2. There is **no significant risk that the application might upset or alienate study members or of reducing their willingness to continue as active participants**.
3. There is **no significant risk that the application might harm individuals in the study, or the study as a whole**.

Like in many ethical criteria, the risk of disclosure and identification are included and this will be discussed in more detail later. Criteria 5 and 6 focus on harm to the data subjects due to distress or stigmatization and these formed an important part of the discussions of the METADAC committee. Consider this possible example, a researcher wants to use genotypic data to investigate the relationship between IQ and criminal

[8] Source: based on the authors experience in assessing the UK Statistics Authority ethics self-assessment forms

[9] METADAC (https://www.metadac.ac.uk/) was formed in 2015 to assess all research projects using genotype and phenotype data and biological samples from UK longitudinal and cohort studies. The functions of the committee were moved to an alternative committee, but the committee made a huge contribution to the data access community and information about their work can still be found online.

behavior. During their analyses they find an incidental finding that a particular group are more likely to have a propensity for criminal behavior. The potential for harm would be high here. Members of that group could have particular characteristics falsely attributed to them and could suffer discrimination and harassment as a result. Such findings may be picked up by the media who may present them in a sensationalist way. Certain topics, such as criminal behavior, mental health, single parenthood, are often highly emotive and controversial topics. In METADAC applications, researchers were required to consider the possibility of controversial findings and to provide a statement on their publication and media strategy.

Principle 2, the risk of re-identification and attribute disclosure

Statistical disclosure control (SDC) is a key part of this principle, and many secure data access facilities will ensure that researchers have received relevant training before they access the data. To ensure that there is no residual risk of disclosure in the research outputs, the analytical results are provided as “safe outputs”. Disclosure risk is generally considered to be low in such outputs, but it is not zero. When data is accessed through a secure data facility it is common practice for all analytical results to undergo SDC before they can be published. This is a process by which all outputs are accessed individually and in combination to ensure that the final output could not be used to identify a data subject, or attribute information to them. The standard approach is to apply a set of rules or guidelines to each piece of outputs to determine whether there is any risk to publishing the results. These rules vary between secure data access facilities but will generally center around 3 guidelines: the threshold rule, the dominance rule and group disclosure[10].

The threshold rule requires that a minimal cell count or number of observations (a threshold) must be reached in all analytic results. For example, in a table of frequencies,

[10] The dominance rule and group disclosure will not be discussed here, but information about these can be found in Statistical Disclosure Control guides such as the Secure data Access Professionals (SDAP) SDC Handbook (THF_DataReport_AW_PRINT_art_2019_10_14 (wordpress.com))

all cells of the tables must include at least *n* observations (where *n* is the threshold). Each secure data access facility will set their own, but generally a 'safe' threshold is considered to be 3 or more. Imagine that a data source contains a variable for gross income for two individuals (person A and person B), with a mean income of €47,000. Person A would be able to use the mean to calculate the income of Person B. So counts of 1 and 2 are disclosive. If a third person, Person C is added to the dataset, then even knowing the mean income and their own income, Person A would no longer be able to accurately calculate the income of Person B or C without some collusion. Thus 3 is considered to be the minimum safe threshold.

Thresholds are applied to all types of outputs, from frequency or magnitude tables to regression models and correlation coefficients. But the problem of low counts can be best demonstrated with frequency tables. The following table uses synthetic data for a cross-tabulation between marital status and a diagnosis of Bipolar disorder for 16 year old males in City X.

Table 1: Cross-tabulation of marital status and the presence of a bipolar disorder diagnosis among 16 year old males in City X

Marital Status	**Diagnosed with Bipolar Disorder**	
	Yes	No
Single	4886	32,498
Married	21	647
Divorced	13	309
Widowed	1	0

There is just one 16 year old man who is both widowed and has received a diagnosis of bipolar disorder. In SDC training sessions it is not uncommon for researchers to point out that due to the lack of detailed information about the individual (i.e. ethnic group, occupation) and the large geographical area (e.g. a city with a million plus inhabitants), reidentification is unlikely. But is this the case and how much information would be required here to enable reidentification to occur? In our example, very little additional information would be required. Consider the known characteristics – the individual is widowed at 16 years. It is relatively unusual to be married at 16, and even rarer to have lost one's spouse so this is a highly unusual combination of characteristics and there would be very few men in that category in City X. The circumstances of his spouses death are not known, but might be due to an event that would be likely to have been reported in the local news or the individual may have posted on social media platforms about his bereavement. Thus relatively little effort would be required for disclosure to occur. The potential for harm is high in this example, because in addition to reidentification, it would also be disclosed that the individual had been diagnosed with bipolar disorder.

Consider the implications for such a disclosure. Firstly at the macro level, such disclosures harm the public's trust in data access services and in the research community. Data owners may also stop trusting researchers and choose to withdraw their data. At the micro level, there can be considerable harm to the data subject. Through the disclosure of information that's sensitive, individuals can suffer discrimination or harassment. In the example above, information about the male's mental health was disclosed. There is still unfortunately a great propensity for stigma and discrimination to be experienced by people with mental illness, thus the above output could cause considerable harm. Stigma theory possets that discriminatory behaviours or microaggressions experienced as a result of the disclosure of his mental health status could impact on his everyday interactions as typically those around him might seek to reduce their interactions with him (Johnson, et al., 2020).

Unfortunately, similar forms of disclosure do occur. In a 2021 BBC news report in the UK, HIV Scotland were fined £10,000 for a data breach which led to a number of email

addresses revealed, many of which included names[11]. HIV status is highly sensitive data and as with bipolar disorder, great stigma still exists around HIV so the inference of HIV status from this disclosure will have caused harm. There is an additional point to consider here. Some argue that false or incorrect attribution of information can be more harmful [Lambert 1993]. Ladd (1989) discusses the harm caused due to the incorrect attribution following a disclosure of criminal records, and in the HIV Scotland breach, the list of emails disclosed included some advocates so there is a potential also for false attribute disclosure.

SDC for big data? Implications and recommendations

The case for carrying out SDC on secure data outputs is clear, but what of non-secure data and big data? For non-secure data or open data (data outside of a secure data facility), the assumption is often that there is no risk of disclosure or reidentification because the data is 'safe' or anonymised. The use of secondary data sources separates the researcher from the process of data preparation and anonymisation, and the temptation is to rely on the assumption that all disclosure is taken care of by the time the data is made available. This is of course not an unreasonable assumption. The very fact that data is available outside of a secure access facility, with minimal or no access conditions is a strong indicator of 'safe data', a term used in everyday parlance among data professionals. However, the risk of disclosure is rarely absolutely zero, even in openly available sources where such unique combinations of characteristics are not found. Therefore a good understanding of disclosure risk and SDC is an essential part of ensuring that all data is used safely and ethically; it is proposed here that this is a skill that all researchers should acquire. In the defining characteristics of big data, we find strong justification for this approach. By its very nature, big data collects a large variety of information about an individual, and although direct identifiers have been removed, through piecing together multiple pieces of information (variables) - sometimes referred

[11] "HIV Scotland fined £10,000 for email data breach - BBC News" October 21 2021 https://www.bbc.com/news/uk-scotland-59008366

to as *jigsaw identification*- a complete picture of an individual can be put together which would allow for re-identification[12].

Sometimes surprisingly little information is required for disclosure to occur. Consider the example of data on new COVID infections and hospitalisations. These data are not considered secure data and so would not usually be subjected to SDC. However, the research team adopted best practice and carried out a SDC review of a table showing the number of new infections and new hospitalisations at the population level for a two week period[13]. At first glance, this would appear to be unproblematic even though for one time period and one population there was a single hospitalisation – no additional information is available such as gender or age of the individual. However consider the context – the data is for a number of small island communities, the single hospital admission was from a small island with a population of less than 2000. The data collectors confirmed that the community is active on social media so identification could potentially be achieved with minimal effort.

We can therefore consider that the level of risk depends on how much is known about the data subject after an output is released. In a digital era, this is highly problematic as large amounts of information may be easily obtainable and used in combination with the output to aid identification. More consideration of the potential impact on levels of disclosure risk as a result of increasing volumes of openly available data is required. However it is clear that SDC should no longer be the reserve of the secure data world, but should be embedded in good research practice for researchers using all data types.

Lambert 1993 suggests that the researcher plays an important part in determining the level of risk. Who accesses the data and how they behave matters, so how do we mitigate that? Currently in the UK, researchers requesting access to secure data must undergo mandatory training, and this is also the case in other European data services. This training covers the key aspects of data access, legislation and data protection as

[12] This term is used by the Medical Research Council Microsoft Word - GDPR GN 5_Ident, anon, pseudon_v2019-09-12 (ukri.org)

[13] Source: this example comes from a meeting between the author and data stewards seeking advice on the potential disclosure risk of the outputs.

well as statistical disclosure control[14]. This training is vital as researchers should play an active part in any data security model aimed at keeping data use safe (Desai & Ritchie 2010). Whilst evidence does not show any incidents of malintent on the part of researchers, with poor knowledge or frustration with SDC rules, disclosure can occur as a result of poor practice or mistakes. Thus the aim of the training is to ensure that researchers can be considered 'safe people' (Desai, Ritchie & Welpton 2016).

Such training courses play a vital role, covering data protection, how disclosure occurs and SDC. The current training does not specifically discuss these as ethical issues but it can be framed in terms of ethical use as well as safe use. Where researchers show reluctance to engage with the rules, this approach can be helpful. Currently this training is targeted at a very niche group – those applying to access secure data via certain secure data centres. However, we argue that all researchers would benefit from attending such training and learning the principles of disclosure risk and statistical disclosure control.

Concluding remarks

It is argued here that it is the responsibility of researchers to use data both legally and ethically. Guarding against the risk of disclosure is a key part of meeting that responsibility. Disclosing identity or attributing information to individuals risk causing harm, perhaps through discrimination or stigmatisation. These are concerns faced by all researchers using data, both small and big. However, the emergence of big data and the trend towards recording more of our lives on virtual platforms is compounding these concerns and makes it more imperative that sound ethical research practice is adopted. Hand, 2018 summarises the issue faced by researchers:

[14] These training materials are generally not made available outside of the training sessions, however the Safe Data Access Professionals group have produced a set of training materials that are available for use. These can be found on their website: Training (securedatagroup.org)

"A fundamental aspect of this is that one does not know, indeed cannot know, how data will be used in the future, or what other data they will be linked with. This means we cannot usefully characterize data sets as public (vs. not public) or by potential use (since these are unlimited and unforeseeable), and that the intrinsic nature of the data cannot be used as an argument that they are not risky. It is not the data per se that raise ethical issues, but the use to which they are put and the analysis to which they are subjected."

The introduction of additional processes such as SDC and mandatory researcher training may seem antithetical to open science, indeed many researchers who have accessed secure data will have experienced a sometimes lengthy process requiring much more thought and administrative effort. Restrictions are also often placed on who can access what data and for what research. If we define open science as allowing open data access to all, then these measures are contradictory to that aim. However, if we define open science as making a wider range of data available for research purposes, then the additional measures taken to guard against disclosure risk and/or inadvertent inappropriate use of sensitive, detailed data sources, support the endeavours of open science. This is because data producers are encouraged to share data with the research community that would previously have been unavailable to researchers.

This is one of the great challenges faced in terms of disclosure risk – we cannot guarantee that additional information could be found and combined to affect re-identification, likewise we cannot predict what additional information will be available in the future. As more and more data becomes available, the risk of disclosure will continue to increase and we will need to consider how we deal with that. Key areas such as how we define and measure disclosure risk and how we adapt our statistical disclosure control techniques to new forms of data, will all require re-examining over the next few years. In this article we have argued and demonstrated that in developing policies on ensuring the ethical use of big data, researchers should seriously consider the frameworks and disclosure control procedures adopted by secure data access services.

References

BBC (2021) “HIV Scotland fined £10,000 for email data breach - BBC News” October 21 2021 https://www.bbc.com/news/uk-scotland-59008366

Bishop L. (2017).Big data and data sharing: Ethical issues. UK Data Service, UK Data Archive https://dam.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf

Case C-582/14 Breyer v Bundesrepublik Deutschland. (2016). ECLI:EU:2016:779 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62014CJ0582&from=EN

Desai T. and Ritchie F. (2010) "Effective researcher management", in Work session on statistical data confidentiality 2009; Eurostat

Desai, T., Ritchie, F. and Welpton R. 2016 Five Safes: Designing data access for research. Economics Working Paper Series 1601.

Drew C. (2016) Data science ethics in government. Phil. Trans. R. Soc. A 374: 20160119. http://dx.doi.org/10.1098/rsta.2016.0119 (pg 4)

Duhigg C (2012), How companies learn your secrets. New York Time Magazine. February 16

Duncan GT and Lambert D (1986), Disclosure-Limited Data Dissemination, Journal of the American Statistical Association, 81, 10-28

Duncan, G and Lambert D, (1989) The Risk of Disclosure for Microdata, American Statistical Association, 7(2) 207-217

Dwork C. (2006) Differential Privacy. In: Bugliesi M., Preneel B., Sassone V., Wegener I. (eds) Automata, Languages and Programming. ICALP 2006. Lecture Notes in Computer Science, vol 4052. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11787006_1

Ethics Self-Assessment Tool – UK Statistics Authority, https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/#pid-potential-harm

General Data Protection Regulations Art. 4 GDPR – Definitions - General Data Protection Regulation (GDPR) (gdpr-info.eu) https://gdpr-info.eu/art-4-gdpr/ [accessed 21/10/2021]

Hand. DJ. (2018) Aspects of Data Ethics in a Changing World: Where Are We Now?, Big Data. 6(3) 176-190. http://doi.org/10.1089/big.2018.0083

Johnson TD, Joshi A, Hogan T. (2020) On the front lines of disclosure: A conceptual framework of disclosure events. *Organizational Psychology Review*. 10(3-4):201-222. doi:10.1177/2041386620919785

Kassner, M. (2017) '5 ethics principles big data analysts must follow', 5 ethics principles big data analysts must follow - TechRepublic [accessed 26/10/2021]

Ladd J. (1989) Computers and Moral Responsibility: A Framework for an Ethical Analysis. In Gould C, ed.The Information Web: Ethical and Social Implications of Computer Networking. Westview Press, Boulder.

Lambert, D. 1993 Measures of Disclosure and Harm, Journal of Official Statistics, 9(2) 313-331

Meijeringa, L., Osborne, T. Hoornb, E. & Montagnerc, C. (2020) How the GDPR can contribute to improving geographical research, GeoForum, 117 291-295. https://doi.org/10.1016/j.geoforum.2020.05.013

Metcalf J. (2014). Ethics codes: History, context, and challenges. Council for Big Data, Ethics, and Society. Available online at https://bdes.datasociety.net/wp-content/uploads/2016/10/EthicsCodes.pdf page 5

Ritchie, F and Welpton, R. (2011). Sharing risks, sharing benefits: Data as a public good. In Work session on statistical data confidentiality October 2011, UNECE/Eurostat

Schroeder, R. (2014) 'Big Data and the brave new world of social media research', Big Data & Society. http://doi:10.1177/2053951714563194

Sfetcu, N. (2019) Big Data Ethics in Research, SetThings, http://doi:10.13140/RG.2.2.11054.46401

Soria-Comas, J and Domingo-Ferrer, J. (2016). Big Data Privacy: Challenges to Privacy Principles ad Models. Data Science Engineering, 21-28. http://doi:10.1007/s41019-015-0001-x

Zwitter A. (2014) Big Data ethics. Big Data & Society
https://doi.org/10.1177/2053951714559253