CONFERENCE OF EUROPEAN STATISTICIANS
Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 1.2
29-31 October 2019, the Hague, the Netherlands 15 October 2019

Promoting Statistical Disclosure Control for novices: A Handbook

Richard Welpton (The Health Foundation), Arne Wolters (The Health Foundation), Emily Griffiths (University of Manchester), James Scott (University of Essex), Christine Woods (University of Essex) (United Kingdom)
richard.welpton@health.org.uk

Abstract and Paper

There are now many Safe Settings in the UK that provide a secure IT environment to enable access to confidential and detailed microdata. These are data that have been deidentified (direct identifiers have been removed) but retain characteristics which could potentially identify an individual data subject.

As the level of detail in the data are justifiably required for research purposes, little statistical disclosure control is applied to the data themselves. Instead, Safe Settings operate by releasing only statistical results generated from the underlying data. These results are subjected to a statistical disclosure control check to mitigate the risk of identification of data subjects from published statistics, and to ensure confidential data are not released.

Staff recruited to manage Safe Settings frequently request some guidance about how to undertake statistical disclosure control of the statistical results that researchers ask them to release. While statistical disclosure control methods have existed to check the release of aggregate official statistics (tabular outputs) for some time, in a research environment, other types of statistics are produced and requested for release from the Safe Setting.

The Safe Data Access Professionals group have pooled their knowledge and day-to-day experiences of managing these 'output requests' to produce an informative and practical SDC Handbook. This work intends to fill a gap in the knowledge about statistical disclosure control, and facilitate competent and consistent assessments of statistical results that are requested for release from Safe Settings. This paper provides an overview of the Handbook and how Safe Setting staff can now benefit from a set of straightforward guidelines to apply.

Statistical Disclosure Control for outputs: A Handbook

Carlotta Greci**, Emily Griffiths[+], Yannis Kotrotsios[±], Simon Parker[±], James Scott*, Richard Welpton**, Arne Wolters**, Christine Woods*

Corresponding author: richard.welpton@health.org.uk

*UK Data Archive, University of Essex, **The Health Foundation, [+]University of Manchester, [±]Cancer Research UK

Abstract

There are now many Safe Settings in the UK that provide secure IT environments to enable access to confidential and detailed microdata. These are data that have been deidentified (direct identifiers have been removed) but retain characteristics which could potentially be used to identify an individual data subject; or at least, contain characteristics which constitute personal, confidential data.

As the level of detail in the data are justifiably required for analytical purposes, little statistical disclosure control is applied to the data themselves. Instead, Safe Settings operate by allowing access to data and only releasing statistical results generated from the underlying data. These results are subjected to a statistical disclosure control check to mitigate the risk of identification of data subjects from published statistics, and to ensure confidential data are not released.

Staff recruited to manage Safe Settings frequently request some guidance about how to undertake statistical disclosure control of the statistical results that analysts ask them to release. While statistical disclosure control methods have existed to check the release of aggregate official statistics (tabular outputs) for some time, in an analytical environment, other types of statistics are produced and requested for release from the Safe Setting.

The Safe Data Access Professionals group have pooled their knowledge and day-to-day experiences of managing these 'output requests' to produce an informative and practical Statistical Disclosure Control Handbook (now available at https://securedatagroup.org/sdc-handbook/). This work intends to fill a gap in the knowledge about statistical disclosure control, and facilitate competent and consistent assessments of statistical results that are requested for release from Safe Settings. This paper provides an overview of the Handbook and how Safe Setting staff can now benefit from a set of straightforward guidelines to apply.

Introduction, background and context

Confidential data from a range of sources (government agencies, the health service etc.) are now accessed in a number of Safe Settings in the UK (and in many other countries too). These Safe Settings provide technical and physical security controls to ensure that the confidentiality of the data remain intact. They are often accredited against information security standards. One key feature of their design is that no data may be removed; only statistical results can be released to the user following a check for statistical disclosure control to ensure that no confidential information, and/or information that could lead to a data subject becoming re-identified, is released. This satisfies the 'Safe Outputs' principle from the Five Safes framework for managing access to data (see Desai, Ritchie and Welpton 2015). Examples include the ONS Virtual Microdata Laboratory (now Secure Research Service), the Secure Lab provided by the UK Data Service, and the network of Trusted Research Environments that form Connected Health Cities. Charities including Cancer

Research UK and The Health Foundation have built their own facilities. These Safe Settings routinely provide secure access to analysts who wish to analyse data as part of projects that will deliver some 'public benefit'.

Statistical disclosure control is a set of methodologies and techniques generally used to ensure that statistical releases from producers of national statistics preserve the confidentiality of the data and the data subjects who have supplied data for which said statistical releases are generated. In addition, the techniques can be applied to data to 'anonymise' them, to varying degrees.

In Safe Settings, where analysts request the release of statistical results, staff operating the secure facility will usually screen the outputs to make sure that the possibility of confidential information being released, and or somebody being re-identified, is minimised. This is also a statistical disclosure control (SDC) process, except it is applied to statistical results. In April 2017, a group of staff responsible for checking statistical results (Safe Outputs) produced from confidential data sources, accessed in Safe Settings, came together to discuss and compare requirements and guidelines about how to undertake statistical disclosure control (SDC) for analytical outputs. These staff represented their organisations at the Working Group for Safe Data Access Professionals (SDAP), an informal group made up of staff working at various Safe Settings, established to share experiences and learn from each other.

At the time of meeting, the group was aware of at least two sources of information that they could turn to for practical guidance about applying SDC to analytical outputs. First, the European Statistical System Network (ESSNet) "Guidelines for Checking of Outputs", produced in 2010 (and updated some time later as part of the Data without Boundaries (DwB) project, provides guidance on how to undertake SDC for a number of analytical outputs. Secondly, staff working to supply access to confidential data from health organisations in England spoke of the requirement to follow the "ISB1523 Anonymisation Standard" (published by the Health and Social Care Information Centre, or NHS Digital). This is a set of guidelines for ensuring that statistical tables are non-disclosive. The advent of the Administrative Data Research Network in the UK in 2013 led to a number of other publications such as Lowthian and Ritchie (2017), which aided discussion about the purpose of SDC for analytical outputs further. In addition, the group were aware of a vast academic literature on statistical disclosure, often technical in nature.

However, the group decided that there was an absence of practical 'how-to' guidelines that could be applied by staff without delving into complex content.

A second aspect considered by the group was the rise in the number of Safe Settings establishing themselves. Staff operating these facilities would surely need to undertake SDC of analytical outputs: the group had heard anecdotal evidence that new staff operating the Safe Settings were not always equipped with the necessary SDC experience and skills to undertake this duty. Some had approached members of the group to ask for advice.

The group decided that a Handbook that easily explained the concepts of statistical disclosure, and SDC, and provided practical advice, should be drafted and produced. A "Handbook on Statistical Disclosure Control for Outputs" was born.

The Handbook was first published as a Beta version in March 2019, and finally published in August 2019. However, the authors consider it to be a 'living document'. Privacy,

confidentiality and data do not stand still. The authors have always recognised that others in the field may have examples and methodologies that they may wish to share with others, and are therefore invited to contribute to the further development of the Handbook.

More recently, a set of training materials have been devised, based on the contents of this Handbook. Another development includes the formal training and accreditation of some staff undertaking SDC in the UK by the Office for National Statistics.

However, the Handbook was produced not only for staff charged with undertaking SDC of statistical outputs; but also for analysts who are using these services. Staff have provided evidence that the outputs they are requested to check and release by some analysts can sometimes fall short of the requirements necessary. The Handbook could be issued to analysts using the facilities to guide them into producing outputs that satisfy SDC requirements.

This paper provides an overview of the Handbook and the resources subsequently developed.

Aims and objectives

The authors recognised early on that the Handbook needed to be produced in a relatively non-technical language. Often, staff who would be using the Handbook may be coming to the role without a statistical background. Given the number of secure data access facilities that have been established, many staff have been recruited with IT and other backgrounds: statistics did not make up a major component of staff experience. The concept of SDC may represent a learning hurdle for these staff; and in addition, they may not have the resources to undertake formal statistical training in their roles. The SDAP Competency Framework (2018) provides guidelines about how new staff can develop their skills, including SDC.

In addition, a quick reference look-up was considered by the authors to be beneficial to staff. The authors also believe that for the first time, analysts accessing the service will have access to the same guidance that staff use when checking outputs. The importance of analysts and staff sharing the same perspective is highlighted in Desai and Ritchie (2009).

Therefore, the aims and objectives of the Handbook are:

1 – to provide easy-to-understand background for non-SDC experts

2 – to provide a straightforward and practicable reference material for staff and users to apply SDC to outputs

3 – to provide staff responsible for the release of safe outputs with some organisational guidelines to help design an efficient SDC system

4 – to enable outputs to be assessed efficiently and consistently

5 – to create a reference that could be adapted and applied in different scenarios, according to different data supplier requirements

6 – to build understanding between analysts who create analytical outputs, and staff who undertake SDC, so that both understand SDC and requirements for releasing outputs.

We note that the SDC Handbook is not aimed at a technical SDC audience; it does not offer a theoretical treatise on how a data intruder scenario could be realised. Where theoretical elements of the SDC literature have been encompassed, we have attempted to explain these concepts in lay terms, assuming that the reader does not necessarily arrive at an SDC role with an SDC background, or even, a statistical background for that matter. We note that various literature exists which already provides a theoretical discourse of SDC.

We also note that organisations that provide access to data (sometimes referred to as 'Data Custodians' because they are ultimately responsibility for the confidentiality of the data) may have different approaches to applying SDC to analytical outputs. We have not tried to be prescriptive about how SDC should be undertaken; therefore, a final aim of the SDC Handbook is to enable different organisations to apply their criteria within the guidance we have provided. As an example, we have published our guidance irrespective if Data Custodians desire to implement a 'threshold' criteria of 3, 5, 10 or 30 in frequency tables.

Approach and structure

The group met to pool together their experiences of undertaking SDC of analytical outputs on a day-to-day basis. As a result, the Handbook has been structure into three sections:

- About SDC (explaining the concepts of confidential data and how a disclosure of information might occur from a analytical output)
- How to undertake SDC (including a number of worked examples based on frequently requested analytical output types)
- Organisational issues to consider when establishing a system for undertaking SDC

The last section is aimed at managers and staff setting up and operating an existing Safe Setting.

Each of the sections are now described in more detail.

About statistical disclosure control

In this section, it was important that we began with the basics; explaining statistical disclosure as a concept, with examples about how this could occur, and the consequences if it were to occur. We drew on experience from both the authors and other analysts who have written on this topic for services including the former Administrative Data Research Network in the UK.

The section includes, for example, an explanation about 'threshold' criteria, with a simple statistical reasoning that underlies the necessity for these criteria. It also includes information about 'principles-based' and 'rules-based' SDC for analytical outputs (following Elliot and Ritchie (2015)).

By the end of this section, a reader previously unfamiliar with the topic should be able to understand what exactly SDC is and why it matters.

How to undertake statistical disclosure control

The group pooled their knowledge based on the most common outputs requested by users of their services. These 'output types' included for example: descriptive statistics, types of

graphs, modelled outputs. For each of the 18 output types presented, the authors considered how they would approach applying statistical disclosure control methods. These methods and the approaches taken have been summarised and carefully presented by each output to describe to the reader our suggested approach for undertaking a statistical disclosure control assessment.

In addition, we applied a 'traffic light' system for making it easier for staff to think about the factors involved in deciding whether to release an output or not:

RED:	releasing this will almost certainly release confidential information, and has a higher than normal probability that an individual could be re-identified.
AMBER:	may be 'safe' to release, but staff may wish to find out more information first.
GREEN:	almost no concerns with confidentiality or identification present.

In many of the listed outputs, alternative methods of presenting the statistics are provided, such that concerns about statistical disclosure are addressed, without affecting the overall results that are requested for presentation. For example, a map presented with points, could, in circumstances, be presented as a heat map instead. A scatter plot could be transformed so that each point no longer represents an individual data subject and potentially two pieces of confidential information.

Key considerations that should be thought about when assessing the statistical outputs are listed. For example, are the statistics based on a sufficient number of observations? Is there sufficient information provided about how the statistics have been constructed? Are clear labels and explanations/interpretations of the results provided? We provided these considerations to enable consistency in decision-making by output checking staff, and also to ensure that analysts using the Handbook would understand the thought process by which an output checker assesses an output to be clear of what information is required.

An example set of guidelines for a particular type of output, a box plot, is reproduced here. In this case, an analyst has produced a box plot of income by gender. The Handbook presents the analytical output with green, amber and red labels to indicate to somebody assessing the output for statistical disclosure, what aspects of the output to consider.

The green label indicates where the analyst has produced an output which is easy to understand and therefore assess for statistical disclosure (in

this case, the analyst has neatly labelled the axis). Many staff undertaking SDC receive requests for outputs which are difficult to interpret, often due to poor labelling and information.

The amber labels indicate where further investigation may be required. One amber label indicates that the median value could be suitable for release, providing it is based on a sufficient number of observations. The person checking the output could raise this with the analyst who produced the output. Another amber box provides a suggestion about how some data points could be perturbed to reduce any risk of disclosure, if this was deemed necessary.

The red labels indicate issues with the output that could result in a disclosure. For example, some outliers are indicated in the output, which could be attributed to a single data subject. Another label indicates that some information that would help the person assessing the output for disclosure is absent, and this information could be requested from the analyst.

Organisation issues

In addition to providing guidance about assessing individual types of statistical outputs, a further section of the Handbook delivers general advice about establishing systems within secure data access facilities.

A particular concern has been addressed, which is about how best to manage analysts' expectations about SDC, and follows the work of Ritchie and Welpton (2015). During a workshop on SDC organised by SDAP in August 2018, participants were invited to describe, and provide examples of, good and bad output requests.

The Good…..	The Bad and the Ugly……..
Box plot requested with explanation about why outliers were safe to release	Output requested with no explanation of results
Only requested what was required	Output was a huge log file of the day's analysis and exploration, with no explanation about what was requested
Made request in plenty of time before presentation due to be delivered	Analyst wanted to release a dataset to share with others

This feedback captured from this exercise has been included in the Handbook to encourage staff setting up and operating a Safe Setting to think about how they might encourage analysts to request analytical outputs that are straightforward to assess, and meet SDC criteria. This material could be shared with analysts using the Safe Setting in training courses and through other forms of communication to encourage efficient SDC systems.

Additional materials

The group never intended that the Handbook would remain a static output. Following the release of the Handbook, it has already been distributed around networks of analysts within government agencies in the UK. The UK Data Service intends to provide copies to every analyst who accesses its Secure Lab, following completion of the mandatory training course.

It is intended that the Handbook will periodically be updated. In addition:

- A set of web resources will be created whereby the material from the Handbook can more easily be accessed
- Training resources have been developed collaboratively by staff at The Health Foundation and Cancer Research UK. These will include classroom-based materials and online resources

Contribution to the secure data access landscape

As mentioned previously, while there is a significant and valuable literature investigating various aspects of SDC, as authors we remain convinced that there is little practical material that can assist staff and/or users who need to apply SDC to statistical outputs before they release these results to the wider world.

In this sense, we believe we have developed material that can quickly get people up to speed with the concepts and importance of SDC; and can then apply guiding principles to undertake SDC assessments of statistical outputs before they are released from their Safe Settings.

In addition, we believe the Handbook complements other work being undertaken to provide training and assess the skills of staff who are responsible for this activity. The Handbook can accompany such work as a reference.

This activity shows there is a need for practical guidance about how to apply SDC; and as mentioned previously, many new facilities for providing secure access to confidential data are opening; and more staff who may not necessary have this type of expertise are becoming responsible for this type of work.

The contribution of this work therefore is to ensure data confidentiality is protected by providing staff and users with practical guidance about how to release Safe Outputs.

Finally, and most importantly, we believe this Handbook can bring together staff managing secure data services and their users: it is essential that there is alignment in how both users and staff think about secure data access, and output checking is an aspect in which there are many gains to be realised from joint-understanding (Desai and Ritchie 2009), not least efficiency in how statistical outputs are requested.

Feedback and future development

A 'beta version' of the Handbook was launched in March 2019. This was circulated widely amongst staff working at SDAP-represented organisations. Initial feedback that the authors have received has been encouraging.

One statistical support officer at the ONS explained that the Handbook was being provided to new members of staff as part of their induction and training.

New secure data access facilities have also reported using the Handbook: staff are routinely referring to its pages for guidance when releasing statistical outputs. The team of Secure Data Environment Marshalls, at The Health Foundation began using copies of the Handbook immediately; ditto staff in the Cancer Intelligence team at Cancer Research UK.

However, the authors acknowledge that the publication of this Handbook will not be the last word. As analysts develop new and innovative ways of presenting their statistical findings,

so too will SDC techniques evolve. The authors anticipate updating the Handbook with new examples and methods in the future, and are eager to hear from experts in the field who can contribute.

References

Brandt, M., Franconi, L., Guerke, C., Hundepool, A., Lucarelli, M., Mol, C., Ritchie, F., Seri, G., Welpton, R. (2010) *"Guidelines for the checking of output based on microdata research",* available at http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf

Desai, T., and Ritchie, F. (2009) *"Effective Researcher Management"* http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.15.e.pdf

Desai, T., Ritchie, F., Welpton, R. (2016) *"Five Safes: Designing data access for research",* available at https://uwe-repository.worktribe.com/output/914745

Elliot, M., and Ritchie, F. (2015) "Principles- versus rules-based output statistical disclosure control in remote access environments," Working Papers 20151501, available at https://ideas.repec.org/p/uwe/wpaper/20151501.html

Greci, C., Welpton, R., and Woods, C. (2018) *"The Safe Data Access Professionals Competency Framework",* available at https://securedatagroup.org/guides-and-resources/

Health and Social Care Information Centre (HSCIC) (2013) *"ISB1523: Anonymisation Standard for Publishing Health and Social Care Data",* available at https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care-data

Lowthian, P., and Ritchie, F. (2017) *"Ensuring the confidentiality of statistical outputs from the ADRN: Technical Report"*, available at http://eprints.uwe.ac.uk/31986/1/SDC_Guide_1.pdf

Ritchie, F., and Welpton,. R (2015) "*Operationalising principles-based output SDC*", available upon request (please contact the author of this paper).

Information about the **Working Group for Safe Data Access Professionals (SDAP)** can be found at their website: https://securedatagroup.org/. A copy of the Handbook on Statistical Disclosure Control for Outputs can be found in the Guides and Resources section.

Acknowledgements

The authors wish to express their thanks and gratitude to the following individuals for their guidance and support as the Handbook was developed: **Anthea Springbett**, Professor **Felix Ritchie** (University of the West of England), Professor **Mark Elliot** (University of Manchester), Professor **Matthew Woollard** (UK Data Archive, University of Essex), and for the **participants of the SDAP Workshop on SDC** (including staff from HMRC Datalab,

ONS Secure Research Service, UK Data Service, Cancer Research UK, NHS National Services Scotland, The Health Foundation) held at The Health Foundation in August 2018.

Finally, the authors are grateful for the generosity of The Health Foundation for sponsoring the design and publication of the Handbook.