The inadvertently revealing statistic: a systemic gap in statistical training?

Ben Derrick, Elizabeth Green, Felix Ritchie, Jim Smith, Paul White

Data Research Access and Governance Network, University of the West of England Bristol

Corresponding author: felix.ritchie@uwe.ac.uk

Introduction

When confidential data is used to produce statistics, there is always a risk that the statistics inadvertently reveal personal information about the data subjects. For example, a comparison of mean incomes in some small villages might show significant disparities – until one realises that the cause is a famous author living in one village. Some simple assumptions about the distribution of other earnings could then inadvertently reveal the approximate income of the author. As the use of confidential data grows, so does this risk. This is the problem of output 'statistical disclosure control' (OSDC).

For professional researchers using confidential data in secure environments (see below), this is an easily-recognised problem. Novel solutions to make the data available securely have gone hand-in-hand with significant improvements in the theory, practice and teaching of OSDC. To this relatively small group, OSDC and output checking (reviewing statistics for disclosure risk before release) is a familiar and regular part of statistical production.

But for most data users, it is likely that these risk and their solutions have passed them by and there is a need to raise awareness. OSDC is a field with a very small number of practitioners and theorists which has not made inroads into general statistical training. Research methods (RM) courses teach, at best, simple anonymisation techniques when collecting and storing data, but rarely anything about risks in the statistics produced. Neither the American Statistical Society or the Royal Statistical Society, for example, offer courses, guidance or even an acknowledgement of the issue on their websites. A trainee statistician can easily get to PhD without coming across OSDC.

Moreover, not everyone producing statistics is statistically trained. Organisations routinely collect data on their staff and customers, and are encouraged to use this to improve efficiency or services. For example, the HR department of a large organisation may use its staffing database to produce management statistics; or a company could begin producing statistics on its user base. The potential for confidentiality breach rises dramatically if you know who the people in the data are. However, the guidance of the UK Information Commissioner, for example, makes no reference at all to confidentiality risk arising from outputs.

Failing to protect the confidentiality of data subjects is clearly an ethical problem for data users, but failing to check outputs for disclosure risk may well become a significant legal risk as well. Modern data legislation recognises that there are risks all along the data pipeline, and expects responsible organisations to have policies in place to cover all of them. For example, both the UK Digital Economy Act 2017 and the Australian Data Access and Transparency Act 2022 explicitly require data managers to think about how output are produced. Although these two laws only relate to research use, the fact that output checking is specified increases the likelihood that further laws and regulation will follow.

So, this is a potential problem for all users of confidential data; but how serious is it for different groups, and what can be done about it? Before we consider this, we need to be clear what OSDC is.

Output statistical disclosure closure (OSDC)

An earlier article for Significance [1] considered how statistical organisations try to protect outputs. In this case, let us consider how an organisation without those resources may inadvertently release information. Suppose an HR department has carried out a survey of staff and releases the following table of results (taken from [2]).

Male					
	Good health	Fair health	Bad health	Very bad health	Total
White	6	7	3	2	18
Mixed	2	2	3	1	8
Asian	1	0	5	0	6
Black	0	5	0	0	5
Other	0	0	0	1	1
Total	9	14	11	4	38

Figure 1 A problematic output (source: [2])

From this table we can draw several conclusions:

- The single male who doesn't identify with any of the ethnic groups has 'very bad health' (green row)
- All of the individuals who identify as Black have 'fair health' (blue row)
- The one Asian who responded that he enjoys 'good health' knows that his Asian colleagues all report 'bad health' (orange row)

Obviously in real life things are more complicated. Not every small number is problematic; not every big number is safe; multiple tables (and multi-level tables) can create hidden disclosures by differencing; and there are different rules for different statistics (magnitude tables such as means, indexes, estimation results, test statistics and so on). Moreover, in general risky results are also results with low statistical value (very small numbers, outlier values), but sometime the most valuable results are also the most disclosive ("all of the schoolchildren surveyed said they had tried cannabis at some time"; "none of the adult males in the area earned over £17,000 per year", or for small but unusual populations).

However, this simple example raises the question: who would be able to spot this sort of error?

The good: researchers using secure facilities

This century has seen a huge growth in facilities which allow researchers to manipulate very detailed data, whilst remaining under the control of the data holder[3]. These facilities are variously called research data centres, data enclaves, virtual desktop environment or (the current vogue) trusted research environments (TREs). Many statistical agencies now offer a TRE as a way to access their most sensitive data; for example, the Secure Research Service run by the UK Office For National Statistics[1]. TREs operate by allowing researchers to analyse the data under controlled conditions, rather than allowing the researcher to download the data locally. In general, TREs require extensive application processes, formal accreditation of researchers, and virtual environments which block access to other networks. Researchers cannot directly extract their statistical analyses, but must request the TRE for results to be released.

Most TREs operate some form of output checking to ensure that results generated by researchers from this highly sensitive data do not make the sorts of errors shown above (a few TREs allow

[1] https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice

researchers with appropriate training and good track records to clear their own outputs, but this is unusual). In addition, all TREs provide advice to the researchers on OSDC. Sometimes this just takes the form of written guidelines, but it is widely accepted (if not always implemented) that best practice is to train the researchers in good OSDC practices [4]; this makes the whole process more efficient for both the researchers and the staff who do the checking. This is a well-established and uncontroversial process. In the UK, around 1,400 researchers per year go through the national 'Safe Researcher Training'. For the overwhelming majority of researchers, this is their first (and only) introduction to OSDC.

The bad: researchers using distributed data

Many professional researchers collect their own primary data or use secondary data under licence from others, such as the 'end user licence' files downloaded from the UK Data Archive. These present different risks.

In the case of secondary data, the distributors sometimes include guidelines on OSDC (as in the case of Eurostat or the UK Data Archive https://blog.ukdataservice.ac.uk/statistical-disclosure-control-handbook/), but the general presumption is that researchers will pay limited attention to these guidelines. Hence, distributed data has much less detail or sensitivity than that held in TREs, so that a mistake in outputs is less important.

For primary data, there is not this safeguard. Research methods (RM) classes typically give researchers some idea about good data practices (removing identifiers, only collecting necessary information) as well as the ethics of data collection and use, so they will have some awareness of confidentiality issues. However, we are not aware of any RM courses that include output SDC as a matter of course (except the ones we run).

There may be spillovers from individuals who are trained in OSDC; for example, the SRT has introduced around 6,000 UK researchers to OSDC since 2018, and this will be replicated to some extent in other countries that run TREs. In addition, the training material for restricted environments, such as the UK Data Archive guidance, is often available as open access. However, in general for researchers the awareness of OSDC is going to be very variable, and dependent on the knowledge of colleagues.

The ugly: information generators

Organisations generate a lot of digital information about their operations and customers. This could include internal information such as staff details or sales performance by teams; or it could include external information such as customer purchases, contact details, or website activity. This administrative data can have value for operational reasons, as well as internal management, but exploiting that information can generate problems.

The most widely-publicised problems arise when organisations share data (for example, when Netflix released 'anonymised' customer information which could be linked to IMDb to re-identify subscribers[5]), or when they use the data inappropriately. Despite such breaches, organisations are likely to be aware of their legal obligations when using administrative data for research, or when sharing it, as regulators tend to provide guidelines for organisations to follow.

However, there is a massive gap in knowledge when it comes to the risk posed by statistics generated. Using data from human resources (HR) systems seems to pose regular problems: we hear

multiple examples of staff surveys or operational reports inadvertently singling out individuals or groups.

This is not the fault of those we call 'information generators', analysts who are asked to extract information from internal data sources with possibly very little statistical oversight. As we noted above, even professional researchers and statisticians may have only a fleeting acquaintance with OSDC. Consider the case of an HR department wanting to publish a staff survey – why would they be expected to consider the possibility of someone pulling apart their statistics to find out about their colleagues? Similarly, consider a logistics company analysing delivery data – will it have considered whether management information may reveal confidential information about specific drivers? There are resources on the internet which would cover this – but they are almost all written for professional researchers.

The information generators are typically unaware that there is even a risk to be assessed; this is not their fault. They may have little or no statistical training, and are simply using management information systems provided to them. The guidance from regulators is focused on appropriate use of the data and on data sharing. Data protection officers are expected to have expertise on these, but how reasonable is it to expect them to know about things which the regulator does *not* discuss?

Is this a real problem?

We can try to tabulate the characteristics of the three groups, even if for the last group we have little hard information (we have used a question mark to indicate significant uncertainty):

	Secure researcher	Distributed data researcher	Information generator
Data risk	High	Secondary data: Low Primary data: ..?	High?
Training	Yes	Not usually	No
Accessible training materials & guidance	Sent to researcher	Available if you know where to look	Almost nothing designed for this group
Awareness	Very high	Probably low?	Very low?
Double-checks	Yes	No	No

For the secure researchers, this is fully covered, despite the sensitivity of the data. For researchers using distributed data, secondary data use is probably acceptable given that (a) data detail is reduced, (b) OSDC guidelines are often given as part of the application process and (c) some of the researchers will have come across OSDC training in other contexts.

There is a concern over researchers collecting primary data, who are likely to have had some training in handling sensitive data but who are less likely to have come across the risks in statistical outputs. Moreover, the simpler the statistics (in general), the more problematic they are likely to be.

The big concern is for those who are not in researcher/analysis roles but are still required to produce statistics from sensitive data, possibly with minimal statistical oversight. With little or no guidance, or even awareness of the issues, the scope for serious error is large.

In theory, disclosing information about an individual due to inappropriate statistics can be as serious a breach of data protection law as sharing the source data. In practice, given that it is much harder to extract personal data from a statistical output and requires some technical skill, regulators are

likely to look more favourably on an output breach (and, of course, regulators seem largely unaware themselves that this problem exists, going by their websites).

Nevertheless, this is still a potential breach. Secure facilities take a cradle-to-grave approach to the production of statistics from sensitive data, largely because of their parentage in National Statistics Institutes; these organisations have also shown that OSDC management can be carried out effectively without restricting necessary outputs. For other organisations, including perhaps the ICO, they are simply not sighted on the potential problem – and that means that, when a problem arises, there is no clear way to deal with it.

Plugging the gaps

The gap for the researchers is the easiest to fill. Very few RM courses, at undergraduate or postgraduate level, give guidance on OSDC. This includes courses on data science or data management. Finding space on the timetable for OSDC might not be easy, but as awareness of the issue grows, the demand for such materials should also grow. This should be feasible to accommodate: we have developed OSDC training for a range of academic audiences over the years, and we can state with some confidence that an hour is more than adequate to provide researchers with sufficient practical knowledge. The practical problem is more to do with capacity. Trainers in OSDC for researchers are few and far between, and are almost exclusively associated with TREs. Trainers in output checking are even rarer; as far as we are aware, the authors and Statistics Netherlands (on behalf of Eurostat) are the only organisations offering courses for training output checkers.

There is an additional opportunity to intervene – the institutional (or ethical) review board, or IRB. IRBs could raise awareness of OSDC by requiring researchers to state their plans for managing outputs and directing them to useful resources. However, this assumes IRB members are aware of the problem, and this appears unlikely; in fact, the authors have provided training to their internal IRBs precisely because of this gap in their knowledge (two of the authors sit on IRBs).

For the information generators, the simplest route might be to use company data protection officers, who have an interest in avoiding breaches in their organisations. This does not need to be complicated, and can be embedded into staff professional development for anyone producing statistics or analysing confidential data. We have produced a five-minute video for our own University management teams; this discusses the example in Figure 1 above, and then provides some simple guidelines for staff to consider when producing tables. We are working on simple guidelines for businesses more generally. The aim is not to make the management team experts on the subject, but to raise awareness so that they know future statistical products produced by them should be subject to an OSDC review.

Ironically, one inappropriate development might be raising awareness quite effectively with the information generators. ‘Differential privacy’ (DP) purports to offer mathematical guarantees of privacy by adding noise to statistics [6]. Large companies such as Apple tout their use of DP, and many consultancy firms sell off-the-shelf DP solutions for businesses. Although in reality DP is a technique with strictly limited application [7], the fact of raising these issues nevertheless may increase awareness usefully. On the downside, moving from no awareness of OSDC to seeing it as a technical problem to be solved by software may be counter-productive in the long term.

Conclusion

Checking for confidentiality risks in published statistics can be an important gap in the education of people producing those statistics, but simple low-cost training programmes and materials can rectify this. Output checking can be fiendishly difficult and technical, especially for those dealing with highly sensitive data to produce data with a great public good. However, it usually doesn't need to be. The basics are easily taught, even to non-statisticians, and a well-planned awareness-raising programme can make a substantial difference to the risks of using sensitive data.

References

1 McKay Bowen C. (2022) The art of data privacy. *Significance*, January. https://doi.org/10.1111/1740-9713.01608

2 Lowthian P. and Ritchie F. (2017) *Ensuring the confidentiality of statistical outputs from the ADRN*. ADRN Technical paper. Available from https://uwe-repository.worktribe.com/output/888435

3 Ritchie, F. (2021). Microdata access and privacy: What have we learned over twenty years?. *Journal of Privacy and Confidentiality*, 11(1), 1-8. https://doi.org/10.29012/jpc.766

4 Green, E., Ritchie, F., Tava, F., Ashford, W., & Ferrer Breda, P. (2021, July). *The present and future of confidential microdata access: Post-workshop report.* https://uwe-repository.worktribe.com/output/8175728

5 Narayanan A. & Shmatikov V. (2008). *Robust De-anonymization of Large Sparse Datasets*. University of Austin, Texas. https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

6 Dwork C. (2011) A Firm Foundation for Private Data Analysis Communications of the ACM 54(1) 86-95 https://doi.org/10.1145/1866739.1866758

7 Bambauer J.R., Muralidhar K. & Sarathy R. (2013) Fool's Gold: an Illustrated Critique of Differential Privacy. *Vanderbilt Journal of Entertainment & Technology Law* 16(4) https://scholarship.law.vanderbilt.edu/jetlaw/vol16/iss4/1/