Evidence-based, default-open, risk-managed, user-centred data access

Felix Ritchie1, Hans-Peter Hafner2, Rainer Lenz3, Richard Welpton4

1 Bristol Business School, University of the West of England

2 right.basedonscience

3 University of Dortmund

4 Cancer Research UK

Corresponding author: Felix Ritchie (felix.ritchie@uwe.ac.uk)

**Keywords:** data access, confidentiality, statistical disclosure control, EDRU

# Introduction

In recent years, there has been an increasing demand from the academic community for more access to confidential data for research purposes, particularly that data collected by government departments. This has happened for three reasons. First, both governments and users have become conscious of the research value in data resources and the financial pressure to re-use data. Second, the growing use of administrative data has greatly increased the range of questions that can be answered. Third, over the last decade or so there has been growing evidence that secure academic research use of the most sensitive data can be managed without placing excessive burdens on users or data holders. In addition, the wider user community is increasingly demanding more granular statistics, such as user-generated tables.

However, the supply of secure efficient data access solutions is still a minority sport: data access is dominated by defensive decision-making which has changed little in decades. A major influence is the dominance of downside risk in the literature on statistical disclosure control (SDC). Over fifty years this literature has successfully developed a coherent approach to analysing problems and proposing solutions, to the extent that anyone facing an anonymisation problem can pick an effective, uncontroversial off-the-shelf solution. However, the SDC literature is uniformly defensive, encouraging users of SDC analyses to take the same attitude.

The net result of defensive decision-making in government and defensive SDC literature is to generate a ‘policing’ model of data security, where right, wrong and responsibilities are clearly defined, and the aim of the data owner shifts from user needs to ‘due diligence’. Moreover, defensive decision-making encourages a focus on theoretical worst cases: the evidence base that has been built up on how researchers actually use data plays almost no part in the literature.

Such a view of the world is likely to lead to higher costs and fewer benefits, both for society and the data owners. Moreover, this holds hidden dangers: it gives a false sense of security, and it is particularly badly suited to future developments in information technology and public opinion.

This paper argues that a change in attitudes can lead to outcomes which are cost effective, more secure, more sustainable, more resilient, and encourage good relationships with stakeholders. We refer to this as the evidence-based, default-open, risk-managed, user-centred (‘EDRU’) model, and it reflects insights from economics, psychology, criminology, and cybersecurity.

This paper summarises the case for this evidence-based holistic approach to data access management. The common themes are use of evidence, integration of statistical and non-statistical approaches, and the effective use of limited resources. While this approach is no longer novel in some communities, it is still unfamiliar enough to cause concerns amongst those implementing data access strategies. This paper aims to address such concerns and demonstrate the importance of grounding strategy in realistic expectations of risk, uncertainty, cost and incentives.

The ideas in this paper were first presented coherently in a 2015 conference paper[1]; since then, there have been substantial developments in both institutional and statutory frameworks, including the GDPR which is much more closely aligned to the EDRU perspective. This paper reviews and updates the 2015 paper, and considers whether the data landscape is now ready for a fundamental shift in perceptions.

# Characterisation of the dominant data access model

The traditional model of data access and SDC is default-closed. That is, it assumes that data must not be released unless it can be demonstrated to be ‘safe’. Safety is determined by the application of statistical techniques to reach a pre-determined definition of ‘safety’. The data are being protected against an ‘intruder’: a person who has access to the data who deliberately attempts to breach protection measures, in order to glean information about one or more data subjects. This intruder is assumed to have access to some external information which can be combined with the protected information, and attention is focused on ‘worst-case’ scenarios.

The traditional model has a number of advantages for methodological development: familiarity; reproducibility; accumulated expertise; a level playing field for comparative analysis: and envelopment of alternative models. Clearly, there is immense value in the body of knowledge built up on SDC techniques.

# Problems with the traditional model, and an alternative

Whilst the traditional model is a very effective methodological tool, [1] lists multiple problems with using this approach in practice:

* the scenarios used in theoretical discussions have little basis on reality, limiting their applicability
* the use of mathematical language to express and analyse problems provides a false sense of objectivity and measurement
* almost all discussions assume a measurable risk, whereas Knightian uncertainty is the relevant concept
* the data-centred approach leads data managers to overestimate the risks and underestimate the benefits from non-statistical approaches

In contrast the EDRU approach

* uses the fifty-odd years of accumulated evidence about how confidential data is accessed and used to determine risk scenarios and appropriate responses
* takes a default-open approach: access to data is given; the only question is *how* this is to be done most effectively while maintaining confidentiality
* uses a cost-benefit framework (which includes wider public benefit, not just the interests of the data holder) to assess the all risks; these includes the risks to the public of not making data available
* is user-centred, focusing on the purpose of data access with confidentiality as a constraint, rather than focusing on confidentiality with usable data as a desirable outcome

This model was originally developed to support controlled-access research data centres, where it has proved to be both more cost-effective and more secure. This is done by building a partnership/shared risk-reward model between data holders and data users, as opposed to the antagonistic defensive/policing model.

# Barriers and opportunities to wider implementation

The EDRU approach requires data holders to actively acknowledge the risks in data access and share responsibility with a third party; the defensive approach needs no such commitment. The defensive approach is not absolving the data holder of responsibility, it is merely hiding that responsibility. This is becoming increasingly clear with changes to the legal framework, with new laws such as the European General Data Protection Regulation (GDPR) making clear that organisation need to be actively managing risk, rather than devolving responsibility. Simultaneously, the GDPR actively encourages organisations to look at non-statistical approaches to risk management – for example, user training or ethical approval processes can both contribute to confidentiality management.

However, while the legal framework is evolving to support multi-dimensional approaches to data access, the most significant barrier to adoption of more user-centred working remains the institutional context: in particular, the explicit acceptance of risk. As a number of papers have noted (see [2] for references), government decision-makers in particular tend to be more conservative because they are incentivised to avoid negative outcomes rather than achieve positive outcomes. A decision-maker adopting the EDRU approach makes himself or herself open to criticism; in contrast, the defensive decision-maker can always protect himself or herself by asking for more guarantees. If “no-one gets fired for buying IBM”, then certainly “no-one gets fired for stating that they feel the case for confidentiality protection has not yet been made”.

# Conclusion

The EDRU approach is being adopted by organisations, for two reasons. First, and perhaps most importantly, it tends to lead to lower-cost outcomes; second, the EDRU model encourages positive engagement and trust-building between actors, which speaks to fundamental human psychology. This adoption is however slow: EDRU requires much more engagement by decision-makers, which usually means face-to-face interaction between parties. As adoption is slow, it limits further adoption: government organisations tend to follow trends rather than lead them. Nevertheless, it seems likely that in the next few years we may pass a tipping point in attitudes to data access as the result of a combination of modern legal frameworks, increasing numbers of exemplars, and a data landscape which requires system thinking to resolve confidentiality issues.

# References

[1] Hafner H-P., Lenz R., Ritchie F., and Welpton R. (2015) "Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use", in UNECE/Eurostat Worksession on Statistical Data Confidentiality 2015, Helsinki.

[2] Ritchie F. (2014) "Resistance to change in government: risk, inertia and incentives". Working papers in Economics no. 1412, University of the West of England, Bristol. December