UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

Disclosure control issues in complex medical data

Elizabeth Green[1], Felix Ritchie[1], Jim Smith[1], David Western[1], Paul White[1]
[1]University of the West of England
elizabeth7.green@uwe.ac.uk

Abstract

The covid19 pandemic assisted the acceleration of routine access to medical records for research. In the UK platforms including OpenSafely and NHSDigital, alongside emerging hospital trust based Trusted Research Environments (TREs), demonstrate the utility and need for medical researchers to access and use microdata safely and securely. Whilst many employ traditional principles-based SDC standards to statistical outputs, complexity arises when considering complex medical data which is required to remain highly detailed; for example genome, medical imaging, or fMRI data where the output often includes reference to individual observations. Current imaging libraries and databases have demonstrated awareness and need for metadata standards, but consideration of both input and output protection is less clear. With the need to retain observations with high level of detail this presentation discusses present considerations for potential SDC solutions and also invites conversation from the wider community.

1 Introduction

The use of medical data for research purposes has clear public benefit and direct impact. Medical data by nature is highly detailed and specific to an individual: it is important to include a wide range of observations and background information to allow practitioners to make informed decisions and choices around treatment. Specific medical tests such as genome analysis or an MRI scan, generates large volumes of data which are specific to the individual and is evaluated and examined as a whole entity- not just a one particular fraction of the MRI scan is used, the whole scan is used and retained.

Historically, medical research has long been intertwined with delivery and provision of care to patients, as such research is conducted with direct informed consent and an expectation that the data will be used to further knowledge in the area. The medical data is of course highly detailed and often the number of observations used in a study can be low due to rarity of disease, or the collection of data is limited to particular hospitals/ sites. As such the research outputs can be highly detailed with descriptive tables and survival curves often including singular observations.

In contrast, microdata used in social science is often not directly collected by the researcher (for example census data) so informed consent specific to the research is not obtained. When it comes to accessing and publishing data outputs, social science has established data repositories and access arrangements for research with clear standards for statistical disclosure control (SDC) within both shared datasets and research outputs.

The aim of this paper is: first, to outline some present examples of sharing of medical data and also outputs of medical data; and second, to reflect on the disciplinary differences in disclosure control. In this paper we will illustrate this with some examples and consider whether this is due to lack of awareness or lack of concern. We will illustrate with three commonplace examples of shared data, to illustrate some of the issues and the expectations of the public health world. Finally, we reflection ways forward and where medical science may benefit from the experience of social scientists.

It should be noted that this paper is not intended to embarrass organisations or researchers- examples where potential disclosure and poor practice has been identified by the team are de-identified and described. The team has not directly referenced these examples, and we encourage the community to have an open conversation about how to integrate SDC standards when sharing data.

2 Medical examples

2.1 Genomic data

The devil is in the detail. A genome provides the complete set of all the genetic information in an organism. Genomic analysis (for example, microarray data) allows for the investigation of genes, and provides the necessary insights for developing cures, vaccines, and identification of new diseases and diagnostic tests. Whilst the sharing of individual genome data has facilitated remarkable breakthroughs in fields such as genetics and personalized medicine, it also raises significant privacy concerns.

The current practice of 'anonymization' of genomic data is performed by removing direct identifiers (for example, name, patient ID) and indirect identifiers (hospital, postcode) (Bonomi, Huang and Ohno-Machado, 2020). However other variables such as age of patient, gender, prognosis are not redacted. Below is an example of an 'anonymised' genome array data- available via website in the public domain which does not require sign in. The data is associated with a published research article, a condition of publication with the journal is that the raw data must be made available.

Data collection: The DRAGoN Hospital for Exhausted Researchers

Participant characteristics:

Participant number	Gender	Age	Prognosis
1	Male	48	Bad- chronic insomnia
2	Female	31	Good

Xlsx attachment with participant 1 microarray, participant 2 microarray etc.

The main issue here is not only the level of detail presented in the participant characteristics list, but also the level of detail within the array/ genome dataset. It is effectively the raw output of the individual's entire genetic array. Whilst research has advanced an understanding of the specific roles of different structural points, mutations, and specific markers knowledge, we are still in the process of identifying and discovering the roles of specific which genetic markers. Therefore, when considering SDC we need to be aware that what is considered non-sensitive today may become sensitive in the near future (Ritchie and Smith, 2019; McKay et al. 2022).

For medical research it is difficult to define what information is disclosive and what is not. For example, it is possible to extract information about the individual such as eye colour, hair colour, hair texture (curly), baldness, physical traits etc from array data. Previous studies demonstrated the possibility of generating 3D face maps based on genomic data which could be used to reidentify individuals (Lippert et al. 2017, Crouch et al. 2018, Venkatesaramani and Vorobeychik, 2021). From a social science perspective we would be considering whether a form of input disclosure control could be employed; alternatively, could we safeguard who is accessing the data, and what might the consequences be if we did introduce such practices?

Input SDC on the sharing of genomic data is only one part of the puzzle. There are also disclosure issues in research outputs. As previously explained the data is uploaded to a shared platform- available for anyone to download, this sharing is often a mandatory requirement from both funders and journals. Below (figure 1) is an example of a published survival analysis which outlines the probability of survival for patients with a particular disease overtime. With small number of values it is easy to identify when individuals die at specific time points- accompanying the survival curve is a table detailing the change in numbers across time.

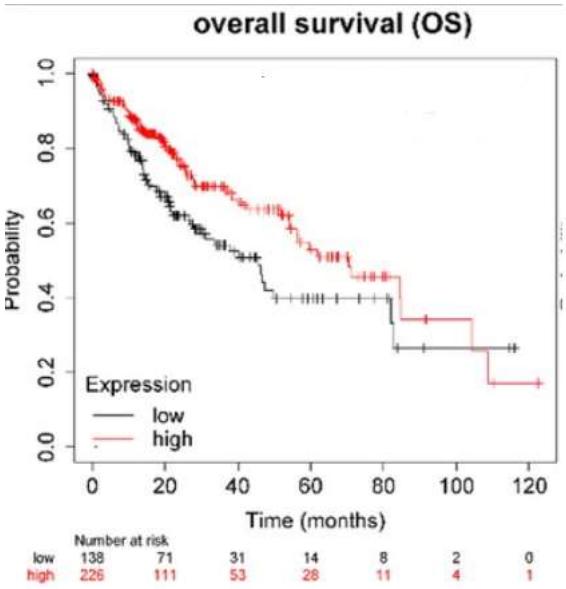
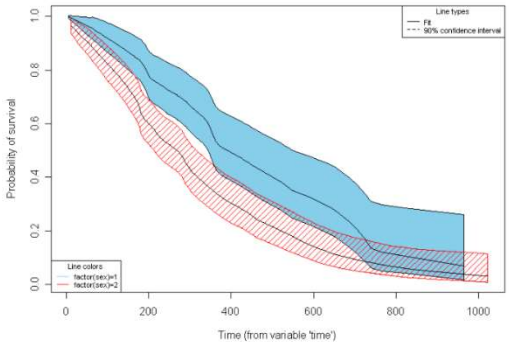


Figure 1 Example Kaplan-Meier curve with low numbers

Survival analysis is commonly used in medical research to demonstrate the relationship between diagnosis (or treatment) and death. Concerns around disclosure relates to number of observations between each step down in the curve, with detailed graphs often detailing a step down with less than 3 observations. O'Keefe et al. (2012) suggests smoothing and incorporating confidence intervals, while SDAP (2019) proposes checking to ensure thresholds are met within each step change.



(a) Confidentialised Output

Figure 2 Demonstration of a confidentialised output taken from O'Keefe et al. (2012) p134

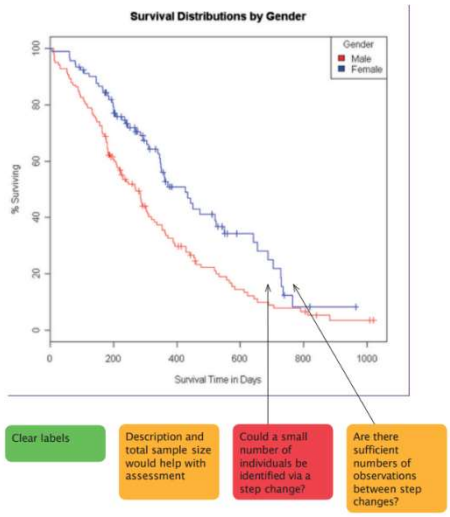


Figure 2 Guidance for SDC in Kaplan Meier graphs by Welpton et al 2019

Interestingly a tool which specifically generates Kaplan-Meier plots for genomic research is being used within the medical community- https://kmplot.com/ (Gyorffy, 2023). This open-access, free for use website allows researchers to perform survival analysis on different gene expressions from database of over 30k different samples. The user can select below the cancer subtype they wish to research and then the level of analysis (see below). By default the website is set to censor at the threshold for the plot, but the user is able to turn off this function.

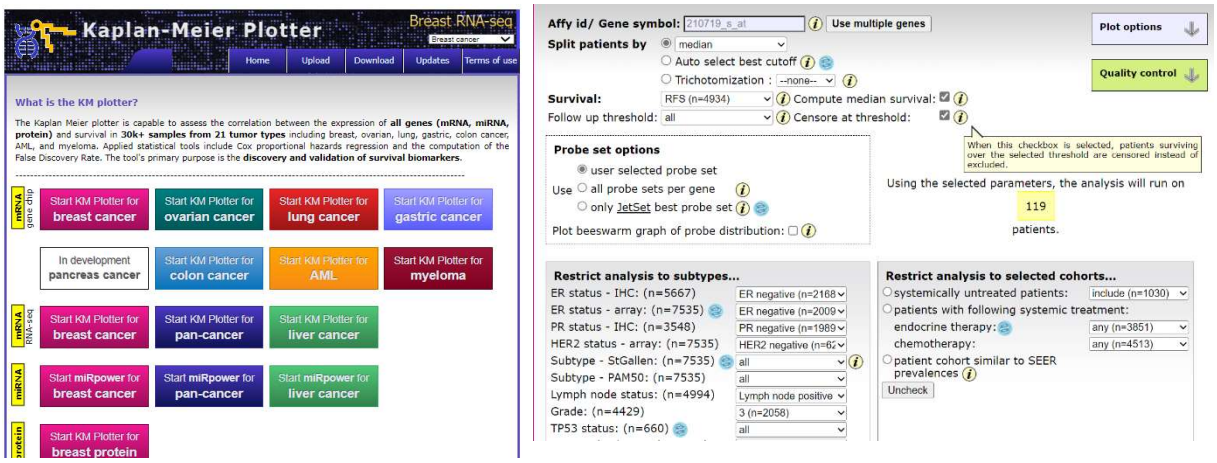


Figure 3 Website Kaplan- Meier plotter

While this is an extremely useful resource for researchers, it is also of potential concern. It seems likely that very small subsets of the data could be selected and associated with personal characteristics – these would not produce meaningful graphs, but they could be used to challenge the anonymisation of the data.

2.2 Inappropriate use of medical dermal images

In dermatology, photographic capture of clinical findings is routine, with digital images providing support and awareness in both practice, research, training, and education. One publicly available tool is the DermAtlas (available http://www.dermatlas.net/reference/index.cfm) which stores a wide array of clinical images demonstrating the presentation of different dermological conditions. Anyone can access this tool and explore the wide range of photos it holds. In terms of impact this tool can help aid health professionals in identifying and evaluating their own patients, it can also be used by the general public to help them feel empowered or understand their own conditions/ potential diagnosis.

As the skin is the largest organ of our bodies, some dermatological conditions are localised to specific personal areas, this coupled with also an array of different clinical photos providing insights across the age range, the dermatology archives found it had become susceptible to misuse. Lehman, Cohen and Kim (2006) described the journey of discovery, ongoing detection, and management of misuse of DermAtlas content across a period of 4 years. A shocking 14.3% of all referrals originated from pornography / fetish sites (Lehman, Cohen and Kim, 2006).

This leads to concerns surrounding how to share safely medical information from what is undoubtedly a valuable medical resource. Any referral from a pornography/ fetish site resulted in the user being presented with a denial page (Lehman, Cohen and Kim, 2006). The DermAtlas implemented filters through user query patterns, with IP addresses of frequent queries for genital images being restricted. Restricted IP addresses were still able to use DermAtlas, but were presented with thumbnail sized images and unable to retrieve full images of genital sites. However, this approach was not straightforward: for example, the NHS in the UK and US military services were then inappropriately restricted.

DermAtlas presents an interesting example of the complexities when hosting data in a public domain which is aimed for a universal audience. The benefits of the tool for both public and health professionals are clear, but the tool is also being used for other purposes not intended by the designers. When considering potential solutions for de-identification or anonymisation of medical photos, current practice in social research where direct informed consent has not been obtained (such as photographing a busy city) is often to use object and

face detection software to automatically mask individuals (Fitwi et al. 2021). When considering the clinical dermatological case photos, the current simplistic approach is to redaction is to mask the eyes and mouth, but for many case photos there is no form of redaction, and sometimes it is not possible to redact the eyes and mouth. We therefore assume, as is common practice within medical research, that the emphasis is on obtaining direct informed consent- and the patient consents to data being held within the public domain. However, can such consent be truly informed when unanticipated uses are made of the data? The DermAtlas and indeed other similar tools face a impossible triad: how can we retain detailed photographs *and* provide an open access tool *and* ensure no misuse?

2.3 fMRI scans

Functional magnetic resonance imaging or functional MRI (fMRI) provides a highly detailed image of the blood flow and structure of an item/ body part, these scans are being used to assist in treatment of the patient (diagnosis) but also medical research. Due to the large volume of high data produced by these scans sharing this information has proven to be invaluable for medical research. Current examples of sharing fMRI includes the Brain Imaging Data Structure (BIDS) website https://bids.neuroimaging.io/ . Here users can contribute, access, and download de-identified fMRI data.

When considering the input disclosure control BIDS requires contributors to remove all direct identifiers alongside 'defacing' the scan images (which can be achieved using a module https://raamana.github.io/visualqc/gallery_defacing.html). Interestingly facial reconstruction based on detailed medical scans (such as CT, fMRI) has been achieved. Schwarz et al (2019) found that the software achieved an impressive re-identification rate of 83% (70 of the 84 participants) when comparing their MRI scan to photos.

BIDS ensures that the data entering the service is de-identified by providing excellent support to depositors- ensure that data uploaded to their service is stripped of direct identifiers and defaced. However uploading and publishing/ sharing data in tandem is common practice so the sticky issue of secondary disclosure is more apparent in this example. To highlight this a recent published journal article, cites that they have deposited the data used in publication in BIDS, but within the journal article the participants' demographic characteristics are highly detailed with low numbers in particular cells and distinctive characteristics. If the identity of the depositor is known, then it increases the chance of knowing where the sample comes from (i.e. which hospital/patient group), dramatically increasing the chances of re-identification. Finally, with more researchers using data depositories such as BIDS to deposit datasets used in publications/ research, information already in the public domain about the dataset may be crucial for re-identification, but it is not necessarily considered by the individual depositor. Now the problem here is not within the data depository input side, but a lack of statistical disclosure control awareness from the authors- demonstrating the need for training and standards amongst the medical community.

3 Discussion

We are not stating that the above examples are necessarily disclosive or provide direct identification- a number of steps would be required to reidentify the individual and the value to an intruder would be questionable. For example, safe to assume that social media profile pictures in the public domain are not going to be viable for identification/ reconstruction of an fMRI scan. Venkatesaramani and Vorobeychik, (2021) found that the overall effectiveness of re-identification (when using social media photos) was substantially lower than previously suggested- as literature often uses high-quality data (both genomic and photographic) which is not consistent with real life scenarios. Conceptualisation of what is a reasonable threat is beyond the scope of this paper.

Nevertheless, the three examples have highlighted a number of issues and challenges within disclosure control from both an input and output side along with how to share. Many of these challenges are unique to the data, and traditional methods used to aid disclosure control in social research may be inappropriate. There are also some very unexpected factors; for example DermAtlas and the actual use. Going forward what mitigations and recommendations might social scientists offer the medical community?

On microdata access we must always accept a level of risk, risk needs to be conceptualised as to the realism of risk (i.e. what is the true likelihood of an intruder performing this for nefarious gain? And can we ever meaningfully and more importantly reliably measure this risk?). It is also essential that whilst discussing risk we must also discuss benefit, we are all to familiar of the invaluable findings and applications of health research and to potentially halt or delay findings is harm within itself. So, whilst we highlight areas of weakness and vulnerability we must objectively generate new paths going forward.

Our primary concern is the lack of standards, guidance and continuity- this is not being checked or reviewed or updated to current practices known within the SDC community (for example thresholds). Perhaps this demonstrates a lack in training and awareness around SDC, as in the examples there are demonstration of de-identification. This also could potentially be an area in which re-identification back to the individual is important for example if the research generates incidental findings on an individual and it's necessary the receive intervention. Consent for data to be shared is often obtain directly with individuals being more inclined to trust the research and a presumption that they had "agreed to use this for research and we said we would anonymise it...".

What about outputs? Sharing the data seems to happen in tandem with the outputs so output SDC not as relevant, however is this an output or input issue? Should we consider the attached journal participants characteristics tables as secondary disclosure or is this an example of input? What is clear however is a want to de-identify and a concern around ethics and consent in the medical community. Derrick et al 2022, highlights that training in OSDC is mostly limited to TRE users and lots of medical research on very sensitive data is not traditionally held in TREs (compare do social science), so moving forward training appears to be long-hanging fruit in supporting disclosure control in this area.

4 Future considerations

Identification of problem/risk – at first glance appears poor practice when compared to standards in social science but is it a genuine risk? How do we balance genuine risk vs perceived risk vs utility of data?

Training – what is done and to what level (again as social scientists not great but perhaps have experience and also conceptual understandings of thresholds, rounding etc).

Standards- what is done and to what level- can we support a harmonised approach?

Is open sharing good? In social science the move has been to open access not open data i.e. anyone with genuine reason has access to the data but not everyone gets access – need to review data sharing models and also pressures from funders and journals.

We especially welcome views from medical research community dragon@uwe.ac.uk

5 References

Bonomi, L., Huang, Y., & Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. Nature genetics, 52(7), 646-654.

Crouch, D. J., Winney, B., Koppen, W. P., Christmas, W. J., Hutnik, K., Day, T., ... & Bodmer, W. F. (2018). Genetics of the human face: Identification of large-effect single gene variants. Proceedings of the National Academy of Sciences, 115(4), E676-E685.

Derrick, B., Green, E., Ritchie, F., & White, P. (2022, September). The Risk of Disclosure When Reporting Commonly Used Univariate Statistics. In International Conference on Privacy in Statistical Databases (pp. 119-129). Cham: Springer International Publishing.

Fitwi, A., Chen, Y., Zhu, S., Blasch, E., & Chen, G. (2021). Privacy-preserving surveillance as an edge service based on lightweight video protection schemes using face de-identification and window masking. Electronics, 10(3), 236.

Gyorffy B: Discovery and ranking of the most robust prognostic biomarkers in serous ovarian cancer, Geroscience, 2023, doi: 10.1007/s11357-023-00742-4.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., ... & Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS genetics, 4(8), e1000167.

Lehmann, C. U., Cohen, B. A., & Kim, G. R. (2006). Detection and management of pornography-seeking in an online clinical dermatology atlas. Journal of the American Academy of Dermatology, 54(4), 633-637.

Lippert, C., Sabatini, R., Maher, M. C., Kang, E. Y., Lee, S., Arikan, O., ... & Venter, J. C. (2017). Identification of individuals by trait prediction using whole-genome sequencing data. Proceedings of the National Academy of Sciences, 114(38), 10166-10171.

McKay, F., Williams, B. J., Prestwich, G., Bansal, D., Hallowell, N., & Treanor, D. (2022). The ethical challenges of artificial intelligence-driven digital pathology. The Journal of Pathology: Clinical Research, 8(3), 209-216.

O'Keefe, C. M., Sparks, R. S., McAullay, D., & Loong, B. (2012). Confidentialising survival analysis output in a remote data access system. Journal of Privacy and Confidentiality, 4(1).

Schwarz CG, Kremers WK, Therneau TM, et al. (2019) Identification of anonymous MRI research participants with face-recognition software. N Engl J Med; 381:1684-6.

Venkatesaramani, R., Malin, B. A., & Vorobeychik, Y. (2021). Re-identification of individuals in genomic datasets using public face images. Science advances, 7(47), eabg3296.

Welpton, Richard (2019). SDC Handbook. figshare. Book. https://doi.org/10.6084/m9.figshare.9958520.v1