Institution: University of Edinburgh (UoE)

Unit of Assessment: 7

Title of case study: Transforming access to sensitive personal data - Enhancing data infrastructure, access and capacity for policy development

Period when the underpinning research was undertaken: 2013: 2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Chris Dibben	Director of Administrative Data Research Centre-Scotland, Director of Scottish Longitudinal Study, Director of SafePOD network and Chair in Geography (Health and Environment)	11/2013- present
Beata Nowok	Research Fellow Geography, School of Geosciences	12/2013- present
Gillian Raab	Research Fellow, ADRC-S	12/2013- present

Period when the claimed impact occurred: 2014 – December 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact (indicative maximum 100 words)

Integration of administrative and population data sources (eg medical records, household composition) is pivotal for evidence-based policy making. However, access to, and use of, linked data types is hampered by data protection regulation.

The Administrative Data Research Centre Scotland (ADRC-S) has developed a set of measures enabling enhanced, legally compliant and secure access to sensitive personal data for professional training purposes and evidence-based policy making.

These ADRC-S innovations have enabled training of the next generation of professional information and data scientists, enabled the provision of novel data analysis forums. Their generation of experimental data has informed evidence-based policy by UK government agencies, police forces and the National Health Service, and their work on grouping populations into ‘households’, notable care homes, proved critical in influencing the Scottish Government’s response to the COVID pandemic.

2. Underpinning research (indicative maximum 500 words)

Authorised use of linked whole population data, though potentially extremely powerful, has been limited in many settings, nationally and internationally, by informational gaps, data protection regulations and privacy laws. The ADRC-S has established a series of foundational research programmes enabling the introduction of novel data access and linkage approaches for the fuller integrated use of routinely collected health and social administrative data. The key underpinning research by ADRC-S, led by Professor Chris Dibben, is summarised below.

The legal concept - Functional anonymisation.
ADRC-S has developed the novel legal concept of **functional anonymisation [3.1,3.2]** for a dataset containing personal information, which enables its treatment as legally anonymous if it is ‘not reasonably likely’ rather than ‘not possible’ for a person's identity to be deduced from the data**.** Hence very detailed information can be released to analysts legally.

The digital security concept of ‘embassy safe spaces’ – SafePODS.
ADRC-S have developed a novel concept of ‘Embassy’ micro-safe infrastructures (SafePODS) **[3.1]** which provide a fully controlled and consistent environment for data analysis remote from the main data centres but controlled by those centres – allowing the data to be treated as functionally anonymous because of the controlled environment.

A statistical method for jointly analysing datasets held on offline servers – eDATASHIELD.
A capacity to combine datasets is crucial. This increases the size of research datasets and enables comparative research, but legal restrictions can prevent agencies sharing data across borders. A statistical process and software programme in R has been developed to allow remote and non-disclosive analyses of sensitive data to be carried out via the eDATASHIELD protocol **[3.3].** The protocol exchanges non-disclosive summaries of statistics between agencies, making it possible for exact statistical results to be calculated.

Software and statistical methods for producing synthetic population data – ‘Synthpop’.
Synthetic data allows the widespread release of otherwise sensitive data. It mimics the real data and preserves the relationships between variables but is safe to release because the data is ‘artificial’. Whilst a developing area in the literature, there were no software packages that could be used easily to implement these methods. We resolved a number of significant methodological issues and developed a new software package in R - ‘Synthpop’. This involved novel methods for making inference and for estimating the utility and privacy of the synthetic data **[3.4]**.

Standards and acceptability – GUILD (GUidance for Information about Linking Datasets).
ADRC-S has developed standardised guidance specifying each step of the linkage pathway to improve the transparency, reproducibility, and accuracy of linkage processes, and the validity of analyses and interpretation of results. This procedure has become foundational for data linkage research practice and protocol **[3.5]**.

Developing new data assemblages to characterise populations.
ADRC-S has developed novel methods for assembling data so it can be used to explore key characteristics of populations. For example, to form an understanding of housing, families and households (which are not recorded in UK administrative data), we developed the CURL tool which links Community Health Index and Unique Property Reference numbers. The entire Scottish population was probabilistically linked to their exact residence enabling people to be grouped into ‘households’ and the nature of these to be understood, such as whether it is a care home **[3.6].** Work using this proved critical in influencing the Scottish Government’s response to the COVID pandemic.

3. References to the research (indicative maximum of six references)

[3.1] Dibben, C., Elliot, M., Gowans, H. and Lightfoot, D. (2015). The data linkage environment. *Methodological Developments in Data Linkage*, pp.36-62. In: Harron K, **Dibben** C, Goldstein H, editors. ***Methodological Developments in Data Linkage***. London: Wiley. doi: 10.1002/9781119072454 [3 citations]

[3.2] Elliot, M., O'hara, K., Raab, C., O'Keefe, C.M., Mackey, E., **Dibben, C.,** Gowans, H., Purdam, K. and McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. ***Computer Law & Security Review***, *34*(2), pp.204-221. doi:10.1016/j.clsr.2018.02.001 [4 citations]

[3.3] Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., **Dibben, C**... & Wilson, R. (2014). DataSHIELD: taking the analysis to the data, not the data to the analysis. ***International journal of epidemiology***, *43*(6), pp.1929-1944. doi:10.1093/ije/dyu188 [71 citations]

[3.4] Raab, GM, Nowok, B & Dibben, C. (2018). Practical data synthesis for large samples, ***Journal of Privacy and Confidentiality***, *7(3)*, pp.67-97. doi:10.29012/jpc.v7i3.407

[3.5] Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.C., Smith, P., **Dibben, C**. and Goldstein, H. (2017). GUILD: guidance for information about linking data sets. ***Journal of Public Health***, *40*(1), pp.191-198. doi:10.1093/pubmed/fdx037 [33 citations]

[3.6] Akgün, Ö., Dearle, A., Kirby, G., Garrett, E., Dalton, T., Christen, P., **Dibben**, C. and Williamson, L. (2019). Linking Scottish vital event records using family groups. *Historical Methods:* ***A Journal of Quantitative and Interdisciplinary History***, *53*(2), pp.130-146. doi:10.1080/01615440.2019.1571466 [1 citation]

The underpinning research listed was published in highly ranked academic journals (Scopus citations as of December 2020 shown above), and supported by peer-reviewed grants. Examples include:

Co-I, Dibben, C. (2013-2018). Administrative Data Research Centre – Scotland. *ESRC* [ES/L0007487/1] GBP6,877,894

PI, Dibben, C. (2018-2022). Administrative Data Research Centres – 2018. *ESRC* [ES/S007407/1] GBP4,185,000

4. Details of the impact (indicative maximum 750 words)

Administrative Data Enhancements

ADRC-S has developed state-of-the-art 'administrative data enhancements' that have led to a wider and more varied range of linked personal administrative data being made more widely available. Secure and private by design, these approaches have made this enhanced provision legally compliant. These innovations have been critical to the development of increased national-level provision of data in Scotland: *"To realise the benefit from data requires us to be able to safely and quickly share data, and link different streams of data up. Having a clear legal basis for that is vital, so having [ADRC-S] help shaping a new approach to how we share and link data has been very influential, including for the development of Research Data Scotland, the new major data infrastructure in Scotland."* Chief Statistician, Scottish Government **[5.1]**

These 'administrative data enhancements' have had impact in the UK and internationally, being adopted by national statistical and government agencies, eg the UK Office for National Statistics (ONS) and Stats Canada. Many of the approaches have been widely accepted across the UK, such as SafePODs (since 2015) and eDATASHIELD (since 2014), as well as outside the UK in Europe and America, notably the Synthpop package (since 2014). Collectively these processes have enhanced the capability and capacity for research within and for Government at multi-national level.

Functional anonymization and safe research spaces

The concept of functional anonymisation **[3.1, 3.2]** has been used as the core legal justification for releasing data for analysis in Wales, England and Scotland. It is a cornerstone concept for advisory services such as the UK Anonymisation Network **[5.2]** and the EU 'datapitch' **[5.3].** It is now a familiar concept for professional bodies: the International Association of Privacy Professionals (IAPP) have produced a lawyer's guide to functional anonymization and pseudonymisation more generally **[5.4].**

'Embassy safe research spaces' **[3.1]** have been developed for national organisations such as the ONS and the Welsh Secure Anonymised Information Databank. The ESRC's UK Data Service has significantly extended the locations where sensitive data can be accessed, from initially 1 or 2 sites to over 25 locations across the UK (eg Oxford, Exeter, University of Highland and Islands) as SafePODS are rolled out across the country, significantly reducing travel times and increasing accessibility and capacity for a wide range of policy-related research – meaning much more analysis is now possible by the national organisations cited. **[5.5]**.

Making useful sensitive personal datasets more widely available*.*

The 'Synthpop' package considerably simplified the process of producing safe and high utility synthetic versions of otherwise sensitive private data **[3.4]**. It was made available to practioners in 2014 and has been downloaded 23,259 times across 129 different countries. Synthpop is judged "*Easy to use, fast and high quality*" **[5.6]**. It has influenced professional development internationally, eg for regular short courses delivered to at least 250 participants for the Institute for Employment Research (IAB), Germany. The Distinguised Research at IAB argues "*For many years, one of the major obstacles that prevented many statistical agencies and organisations from adopting this innovative data protection method [synthetic data], was the fact that no software tool was available that could help practitioners*" and "*This changed with the advent of the synthpop package*", "*It lowers the barrier for statistical agencies …in exploring whether the synthetic data approach would be a feasible strategy*" **[5.7]**. Users of Synthpop include: Institute for Employment Research, Germany; Labor Dynamic Institute, Cornell University; and the Open Source Policy Center, American Enterprise Institute, USA. Synthpop has also enabled creative engagement with data, eg Statistics Canada used it to produce an analytically-rich synthetic data file that was used during external 'codefest' events. One such 'codefest', run in collaboration with IBM, tested cloud-based tools and international teams produced new visualisation suites for Statistics Canada linked data **[5.8]**. It is also used in the private sector, the Head of Data Science at synthetic data company Hazy explains "*One of the projects we were inspired by and use in our daily activities is Synthpop developed at University of Edinburgh…The simplicity of Synthpop and the quality of the data generated make it a great resource for industry*" **[5.9]**.

Allowing cross country research

The eDATASHIELD approach **[3.3]** has been used across the UK to allow comparative research previously proscribed by law. This has allowed, for the first time, research across all three of the UK's Census Longitudinal Studies, for example allowing the Scottish Government's Glasgow Centre for Population Health in 2016 to better understand ill health in Glasgow through comparison to similar de-industrialising towns in England **[5.10]**.

Building confidence with data controllers and developing a new policy evidence base.

Through our careful legal and technological underpinning research, the centre has built trust and acceptability with the public and Government departments and data controllers. This led to increased data sharing, changing practice, attitudes and standards in the Scottish Government and the data landscape more broadly. The overall impact has fundamentally shifted access and approaches to administrative data, unlocking its significant value and building capacity in this area. The Scottish Government's Chief Statistician also explains how ADRC-S leadership "*has meant a number of high-profile evidence gaps, such as benefit claimants, veterans or people in the justice system have been or are being filled, but even more so I think the way you have gone about this work is starting to change the relationship between Government policy officials and analysts, and academic experts to be more collaborative*" **[5.1]**.

Providing an evidence base for COVID-19 policy.

During the COVID-19 pandemic, it has been paramount for government and health agencies to have detailed information quickly. In order to understand transmission, it was especially important to understand who was living together. The ADRC-S measure of households, enabled by the CHI-UPRN Residential Linkage (CURL) tool meant, for example, Public Health Scotland could provide Scottish Ministers with information on transmission from hospitals to care homes. The Chief Data Officer for the Scottish Government comments that "*The work you have led in developing tools to enable administrative data research, such as your work on linking medical data to property locations (Unique Property Reference Number) … enables a better understanding of settings, such as care homes or households, in turn enabling vital COVID-19 research and understanding for Government.*" **[5.11]**

The **CURL** tool was used to inform the key Scottish government report 'Discharges from NHS Scotland Hospitals to Care Homes' **[5.12a].** In October 2020 the Cabinet Secretary for Health in Scotland, outlined: "*I commissioned this report because it is right that residents, families, staff, and Parliament, have accurate data and independent analysis on the transfer of patients to care* *homes and the impact that had in those care homes ...The data from this report gives us a better understanding of the impact of discharges on outbreaks in care homes. We will be taking forward the recommendations that Public Health Scotland make in their report, and we will continue to adapt our guidance and the steps we are taking to protect care home staff and residents in line with the latest data, scientific evidence and clinical advice*" **[5.12b]**.

5. Sources to corroborate the impact (indicative maximum of 10 references)

[5.1] Chief Statistician, *Statistics Scotland*, *Scottish Government* (testimonial letter 12/11/2020)

[5.2] UK Anonymisation Network's *Anonymisation Decision Making Framework,* p.10 https://msrbcel.files.wordpress.com/2020/11/adf-2nd-edition-1.pdf

[5.3] EU datapitch.eu - Legal and Privacy Toolkit p.27 https://datapitch.eu/wp-content/uploads/2018/08/DataPitch-D3.5- Legal-and-Privacy-Toolkit-v2.pdf

[5.4] International Association Privacy Professionals (IAPP) https://iapp.org/news/a/de-identification-201-a-lawyers-guide-to-pseudonymization-and-anonymization/

[5.5] Director, *The UK Data Service*, (testimonial email, 05/11/2020)

[5.6] Synthpop User Survey.

[5.7] Distinguished Researcher, *Institute for Employment Research (IAB), Federal Employment Agency, Germany*, (testimonial letter, 26/10/2020)

[5.8] a) Stats Canada Use of Synthpop: https://www150.statcan.gc.ca/n1/pub/12-206-x/2018001/02-eng.htm - Section 2.3;

b) (R)evolution of generalized systems and statistical tools at Statistics Canada http://rproject.ro/conference2018/presentations/Suzie_Fortier_(R)evolution.pdf - p.13;

c) UNECE - Successes and Challenges in Increasing Accessibility at Stats Canada

http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S5_CAN_THOMAS_AD.pdf - Section 4.1 p.6;

d) The LIDIC hackathon: LInked Data Innovation Challenge https://ipdln.org/2018-conference-info/preconference/student-hackathon - Hackathon Information for Participants, p.3

[5.9] Head of Data Science, *Hazy Limited* (testimonial letter, 12/11/2020)

[5.10] David Walsh *et al. (2016)* History, politics and vulnerability: explaining excess mortality. Report of the: Glasgow Centre for Population Health. https://www.gcph.co.uk/assets/0000/5988/Excess_mortality_final_report_with_appendices.pdf

[5.11] Chief Data Officer, *Scottish Government,* (testimonial letter, 12/11/2020)

[5.12] a) Public Health Scotland (2020) Discharges from NHS Scotland Hospitals to Care Homes between 1 March and 31 March 2020 - https://beta.isdscotland.org/find-publications-and-data/population-health/covid-19/discharges-from-nhsscotland-hospitals-to-care-homes/
b) Hospital to Care Home Discharge Data, Scottish Government news article - https://news.gov.scot/news/hospital-to-care-home-discharge-data-published