Proof-of-Concept: Technical Prototype for Data Access, Storage, Sharing and Linkage (DASSL) to support research and innovation in Ireland

Authors

Dr. Orna Fennelly / Dr. Frank Moriarty / Dr. Derek Corrigan / Loretto Grogan / Dr. Simon Wong

Funded by the Health Research Board

ICHEC
Irish Centre for High-end Computing

HRB
Health Research Board

Ollscoil na Gaillimhe
University of Galway

Disclaimer

Any views expressed in this report are those of the authors.

Published by:

Health Research Board

Grattan House

67-72 Lower Mount Street

Dublin 2

DO2 H638

Ireland

t 353 1 234 5000

f 353 1 661 2335

e hrb@hrb.ie w www.hrb.ie

Please cite this publication as Fennelly, O, Moriarty, F, Corrigan, D, Grogan, L and Wong, SS (2022) Proof of concept: *Technical prototype for Data Access, Storage, Sharing and Linkage (DASSL) to support research and innovation in Ireland*, Health Research Board, Dublin.

Project team

Name	Organisation
Dr Orna Fennelly	Irish Centre for High-End Computing (ICHEC)
Dr Simon Wong	ICHEC
Dr Bruno Voisin	ICHEC
Marta Olszewska	ICHEC
Divyajyoti Sarkar	ICHEC
Sherif Nagy	ICHEC
Dr Frank Moriarty	Royal College of Surgeons in Ireland (RCSI)
Dr Derek Corrigan	RCSI
Prof. Tom Fahey	RCSI
Loretto Grogan	Health Service Executive (HSE)
Prof. Colin Doherty	Trinity College Dublin/St James's Hospital
Mary Fitzsimons	RCSI

Stakeholder Committee

Name	Representing/association
Alan Cahill	Department of Health (DOH) – Statistics and Analytics
Ana Terres	HSE – Research and Development
Anne Quirke (Chair)	Beaumont Hospital – Chief Technology Officer
Anthony Macken	Central Statistics Office (CSO)
Barbara Foley	Health Information and Quality Authority (HIQA)
Ciara Carroll	HSE – COVID-19 Contact Management Programme
Deirdre Mac Loughlin	Public Representative
Frank Moriarty	Secondary Data Analysis Project (SDAP) Researcher
Maria Ryan	HIQA
Michael Courtney	DOH – Statistics and Analytics
Nuala Ryan	Patient Representative
Peter Lennon	DOH – Health Information Legislation and Policy
Rosalyn Moran	EKOS Research Associates

Glossary of terms

Term	Definition
anonymisation	Process undertaken to restrict reidentification of individuals (e.g. data aggregation).
centralised model	Pseudonymised data stored centrally within a data hub that are readily made accessible to approved research projects.
clerical review	Manual review of the linked data to assess the data quality.
cloud solution	Uses the Internet to access remote servers.
content data	Data that are of interest and relevance to the researcher/research question.
COVAX	The HSE national immunisation registry for COVID-19 vaccination
data cleansing	Process of detecting, correcting, and replacing incorrect, incomplete, or inconsistent data.
data controller	Person or organisation that determines the purposes for which, and manner in which, the personal data are processed.
data harmonisation	Act of making the fundamental aspects of the data the same.
data hub	Entity that gathers and organises data for distribution and sharing.
data linkage unit	Entity or department responsible for record linkage.
data or record linkage	Matching individuals, family members, or households from two or more discrete datasets.
data processor	Person or organisation that stores or processes data on instruction from the data controller.
data provider	Person or organisation responsible for sharing data either as the data controller or as a data processor.

Term	Definition
deterministic linkage	Compares the matching variables across datasets to find exact matches related to the same person, family, or household.
distributed model	Data gathered on a case-by-case basis from data providers for a specific project/use and removed from the data hub on completion.
European Health Data Space	Proposed model to foster the exchange and sharing of health data within the European Union.
federated analysis	Data remain at the source and analysis is performed on data within their siloes.
free text	Unstructured or narrative data.
governance	Actions and mechanisms that manage and dictate the processing of health data.
hybrid model	A combination of a centralised and distributed model with some datasets stored centrally and others gathered for specific projects only.
individual health identifier	A number that uniquely and safely identifies each person accessing a health or social care service in Ireland.
Information Governance Review Panel	Independent group of individuals from diverse backgrounds who review research project applications and weigh up the risks and benefits on behalf of the public.
national DASSL service	Proposed national solution, agency and resources to support the Data Access, Storage, Sharing and Linkage (DASSL) model.
National health and social care dataset	Routinely collected data to provide a national overview of a particular health or social care service.
output checking	Review of the findings produced from the research analysis, including statistics and publications.
personally identifiable data	Information which can readily identify the individual whom the data belong to.

Term	Definition
population spine	Comprehensive dataset of individuals from a specific population.
proof of concept	Pilot project which demonstrates that a design or concept is feasible.
probabilistic linkage	Compares the matching variables to estimate the probability that two records relate to the same person, family, or household.
pseudonymisation	Technique used which replaces or removes personally identifiable information in a dataset.
Personal Public Service Number	Unique reference number required in Ireland to access social welfare benefits, public services, and information.
R	Programming language for statistical computing.
RStudio	Integrated development environment for the R programming language, which is used to process data.
Research Data Trust	Supports the use and sharing of health data via repeatable mechanisms and approaches to sharing data in a timely, fair, safe, and equitable way.
Research Support Unit	Staff providing the point of access and support to researchers applying to access health data.
routinely collected data	Data collected for clinical, administrative, or research tasks on a regular basis.
safe haven/ secure processing environment	Locked down and secure environment where researchers can access sensitive health data.
standardised terminologies	Common understanding, use, and methods of aggregation, of clinical terms.
statistical disclosure control	Review of research findings to reduce any risks of reidentification of individuals.

Term	Definition
Structured Query Language	Language used in programming and designed for managing data in a relational database management system.
synthetic data	Data that are artificially created to mimic the characteristics of real data but which do not correspond to real people.
trusted third party	An entity that facilitates safe and secure interactions between two or more parties.
two-factor authentication	Applies an additional layer of security for users beyond username and password.
virtual machine	A virtualisation of a computer system within an Internet window that keeps it separate from the computer it is being run on.
virtual patient register	Patient register generated from existing data resources.
virtual private network	Protected network connection that encrypts Internet traffic and disguises identities when using the Internet.

Acronyms

2FA	two-factor authentication
AI	artificial intelligence
AWS	Amazon Web Services
CCT	COVID Care Tracker
CHeReL	Centre for Health Record Linkage [Australia]
CIDR	Computerised Infectious Disease Reporting
CPU	central processing unit
CSO	Central Statistics Office
CS	case study
DASSL	Data Access, Storage, Sharing and Linkage
DLU	data linkage unit
DOB	date of birth
DPIA	Data Protection Impact Assessment
DPS	Drugs Payment Scheme
eDRIS	electronic Data Research and Innovation Service [Scotland]
EHDS	European Health Data Space
EHR	electronic health record
EOSC	European Open Science Cloud
EPR	electronic patient record
EU	European Union
GB	gigabyte
GB£	Great British pounds
GDPR	General Data Protection Regulation
GMS	General Medical Services
GP	general practitioner
GUI	Growing Up in Ireland
HDH	Health Data Hub [France]
HIPE	Hospital In-Patient Enquiry
HIQA	Health Information and Quality Authority
HPC	high-performance computing
HRB	Health Research Board
HRCDC	Health Research Consent Declaration Committee
HSE	Health Service Executive

ICD	International Classification of Diseases
ICES	Institute for Clinical Evaluative Sciences [Canada]
ICHEC	Irish Centre for High-End Computing
ICPC	International Classification of Primary Care
IGRP	Information Governance Review Panel
IHFD	Irish Hip Fracture Database
IHI	individual health identifier
IP	internet protocol
ISO	International Organization for Standardization
LTI	long-term illness
MCHP	Manitoba Centre for Health Policy [Canada]
MN-CMS	Maternity and Newborn Clinical Management System
NASS	National Ability Supports System
NCRI	National Cancer Registry Ireland
NDRDI	National Drug-Related Deaths Index
NDTRS	National Drug Treatment Reporting System
NHS	National Health Service [UK]
NIMIS	National Integrated Medical Imaging System
NLP	natural language processing
NOCA	National Office of Clinical Audit
NPIRS	National Psychiatric Inpatient Reporting System
NPRS	National Perinatal Reporting System
NREC	National Research Ethics Committee
NSHRI	National Self-Harm Registry Ireland
NSW	New South Wales
OSS	open-source software
PCRS	Primary Care Reimbursement Service
PoC	proof of concept
PopData	Population Data British Columbia [Canada]
PPI	public and patient involvement
PPRL	privacy preserving record linkage
PPSN	Personal Public Service Number
RAM	random-access memory
RCSI	Royal College of Surgeons in Ireland
REC	Research Ethics Committee
RDT	Research Data Trust

RSU	Research Support Unit
SAIL	Secure Anonymised Information Linkage [Wales]
SA NT DataLink	South Australia and the Northern Territory DataLink
SeRP	Secure eResearch Platform
SQL	Structured Query Language
SNDS	Système National des Données de Santé
SPSS	Statistical Product and Services Solutions
TB	terabyte
TEHDAS	Towards the European Health Data Space
TRE	trusted research environment
TTP	trusted third party
UCS	Unified Compute System
UK	United Kingdom
VDI	virtual desktop infrastructure
VM	virtual machine
VPN	virtual private network
WADLS	Western Australia Data Linkage System [Australia]

Executive summary

Today's healthcare environment is data-intensive, with increasingly complex and voluminous data collected via electronic health records; patient demographics; clinical imaging and laboratory systems; primary care; pharmacy systems; population health surveillance data; quality and patient safety data; systems supporting health administrative functions; financial and workforce data; etc.

The COVID-19 pandemic has highlighted the urgent need for Europe-wide health data sharing and coordination. The use and sharing of health and social care data is essential for providing high-quality direct care and for the planning, management, and evaluation of programmes and services. The secondary use of health data means using health data for purposes other than the primary reason for which they were originally collected, and this includes research, development and innovation, audit and education.

In 2016, the Health Research Board (HRB) published a report outlining what it called the Data Access, Storage, Sharing and Linkage (DASSL) model to facilitate the sharing and linkage of health and related data for research purposes. The analysis in the report described existing and evolving approaches to the access, sharing, and use of health data in different countries. While the results showed differing approaches and stages of maturity in the processes and technologies used to access and share health data, the report identified common approaches and patterns in the initiatives, most notably the concept of 'trusted third party' and 'safe haven' services for accessing, storing, linking, and sharing deidentified data with policy-makers and researchers under controlled conditions.

In 2019, following a competitive process, the HRB awarded funding to the Irish Centre for High-End Computing (ICHEC) – the national high-performance computing centre hosted by the University of Galway – to develop a proof of concept (PoC) of the technical infrastructure to support the DASSL model and to provide recommendations for the roll-out of a DASSL-type infrastructure in Ireland.

This report outlines ICHEC's approach to designing, developing, and testing the DASSL PoC infrastructure (using synthetic data and a selection of case studies). It details the learnings in order to ensure that the infrastructure and related governance, services, and practices are informed by international best practice and can successfully protect the privacy of individuals while supporting use of our health and data resources for research and innovation in the public benefit. An example of a use case for the DASSL model would include the combination of COVID-19 vaccination registry with prescription data in order to determine whether individuals on certain medications are at higher risk of a poorer outcome from COVID-19 compared with individuals not on those medications.

The findings from this PoC study are important and relevant in the context of other developments, both nationally and across Europe. Infrastructure alone will not enable the step change required to facilitate the secondary use of health data for research purposes.

The European Commission's upcoming legislative proposal for the European Health Data Space (EHDS) will provide a legal basis for health data use for primary and secondary purposes and, in particular, will require that member states have one or more 'Health Data Access Bodies' in order to manage sharing, linkage, and reuse of data. In order to support the policy intent of the European Commission, the Joint Action Towards the European Health Data Space (TEHDAS) is working with European Union (EU) member states (including Ireland) and the European Commission. This is with the aim to develop insights about infrastructure, governance, workforce, and other concepts for the secondary use of health data to benefit public health and health research and innovation in Europe. Furthermore, the EU4Health Programme has grants available to support member states in developing or enhancing their infrastructural and related capabilities in readiness to comply with the EHDS legislation.

In early 2022, the Minister for Health, Stephen Donnelly, received Cabinet approval to develop the General Scheme of a Health Information Bill. The proposed Bill will help ensure that Ireland has a fit-for-purpose national health information system that enhances patient care and treatment and supports better planning and delivery of health services. The Bill will also support the introduction of a National Health Information Centre with clearly specified functions and governance rules, including the sharing, linkage, and reuse of data for population health purposes and for research and innovation that leads to better outcomes for patients.

Finally, access to health data outside of direct healthcare delivery raises ethical, political, and social issues, as well as technical issues. Countries need a transparent and appropriate framework for the secondary use of health data, with a robust infrastructure supported by good governance and policies, standards, and best practices. Ensuring that the patient and citizen voice is integrated into the design, governance, and operations of health data access bodies, such as the proposed National Health Information Centre, is critical. Evidence shows that individuals feel relatively uninformed about the use of their health data and want greater transparency and information on the tangible benefits of data use. The better individuals are informed, the more they tend to favour the use and sharing of their health data. The proposed Health Information Bill in Ireland will include provisions in particular areas to strengthen the rights of individuals in relation to their health information.

This report is aimed at informing key decision-makers in Ireland in the context of this rapidly evolving landscape. It provides the learnings from developing and testing the DASSL PoC infrastructure, which are informed by key insights from international models and the perspectives of stakeholders. Subject to caveats and assumptions, it describes the technical and resourcing requirements for establishing and operating a national DASSL service for research purposes in Ireland.

The key learnings and considerations for a national DASSL service from this report are provided below.

International initiatives, programmes, and models

1. Cross-border sharing is essential for some areas of public health (e.g. rare diseases, pandemics, antimicrobial resistance), and governance and processes for data sharing within the island of Ireland, the EU, and worldwide are recommended.
2. A national DASSL service could support many upcoming international initiatives (e.g. the EHDS), and the infrastructure should align with these initiatives in terms of data and infrastructure standards where possible.

Health and related social datasets

3. Many valuable health and related social datasets exist in Ireland, and a catalogue of all health and related datasets with standardised metadata should be developed and maintained in order to drive use of data resources for public benefit.

4. Health-related social datasets (e.g. education, housing, employment, social deprivation, criminal justice) provide critically important insights when linked with health datasets, and this ability to link datasets should be embedded within a national DASSL solution.
5. Usability of a dataset to answer a specific research question depends on the characteristics of the dataset, completeness of the relevant fields, accuracy of data collection, and time/population coverage.
6. Standardised terminologies, national data dictionaries, and common data formats provide higher-quality data and support aggregation of data and interpretation of the findings, but changes to these terminologies and formats over time create challenges.
7. Some datasets may require additional consideration of ethics and governance over the data (e.g. consented datasets, genomics, imaging).
8. Access to primary care data for linked data research is critical and will require more digital transformation in primary care centres and the pulling (and mapping, if possible) of data from different general practitioner (GP) information systems.

Record linkage

9. Every national dataset needs to consider collecting personal identifiers to support record linkage and drive the discussed benefits of national data resources.
10. A population spine will be critical to enable record linkage in Ireland.
11. Application of unique identifiers across all health and related datasets would improve linkage quality and reduce the resources required for record linkage.
12. Use of names, addresses, dates of birth, etc. for linkage is unavoidable in Ireland at present, and probabilistic matching, data cleansing, and clerical review may be required by a securely separated data linkage unit (DLU).

Content data management, processing, and preparation

13. A centralised model that stores pseudonymised data on an ongoing basis is more efficient than a distributed model, but it has additional resourcing requirements and data protection concerns.
14. Due to the scalability and flexibility required during the development of a national DASSL solution, a cloud-based solution is recommended, and the procurement process should consider both public and private cloud options.

15. A data management platform that can receive incoming data from data providers should be developed with secure and user-friendly methods for inputting data.
16. A DLU and Research Support Unit (RSU) would require secure processing environments which are completely separate from each other and which contain the required software (e.g. linkage package, relational database).
17. Safe havens should have Internet access, copy-and-paste functions, etc. disabled, with researchers' specific requirements (e.g. statistical packages) made available where possible, and with a mechanism for outputs being vetted by the RSU for statistical disclosure control before exportation from the DASSL system.

Organisational security and data protection measures

18. Policies and processes that align with the Five Safes framework – such as the separation principle (i.e. data providers should split content data from personal identifiers), clear approvals and accreditation processes, training, security and risk management policies, and data sharing and secrecy agreements – are recommended.
19. The entire DASSL infrastructure should be protected by firewall software/hardware solutions that restrict network traffic only between authorised machines and user roles (e.g. access controls, audit logs, two-factor authentication (2FA), virtual private network (VPN)) and that reduce the risk of malicious attacks (e.g. antivirus/antimalware scanners, rootkit hunters).

Governance and approvals process

20. A clear, lawful basis and approvals process for linking health and related social data by different types of users is required, with consideration given to the ability to reidentify individuals for public interest, use of artificial intelligence (AI), and particularly sensitive data such as genomics.
21. Sharing, linking, and preparing datasets, as well as the associated technical infrastructure and software packages, usually incur a charge.

22. The application process should be streamlined as much as possible with ideally a single online application form that would collect all relevant information for a project proposal in one place. This form will integrate information such as details of the proposed research including the specific data that are being requested and linked, feasibility assessments, a Data Protection Impact Assessment (DPIA) as well as information that may be required by different approvals boards, e.g. a National Research Ethics Committee (NREC), the Health Research Consent Declaration Committee (HRCDC), and an Information Governance Review Panel.

Stakeholder involvement and engagement

23. Public trust and support is critical to the success of a national DASSL solution, and a campaign to raise public awareness and education regarding the benefits, as well as an ongoing public advisory panel, public and patient involvement (PPI) in individual projects, and research on public attitudes, is recommended.
24. Other key stakeholders should be engaged via the governance board, advisory committees, surveys, and interviews.

Resourcing

25. Appropriate personnel will need to be hired, including system administrators, data scientists, statisticians, and administrative and communications staff. Expertise will also be required, via secondments or working closely with data providers, to develop in-depth practical knowledge of frequently-accessed and/or complex datasets.
26. The cost of the overall infrastructure, software packages, and staffing will largely depend on the type of service being offered (e.g. centralised versus distributed, internal analysis versus only external researchers, research versus other secondary uses) and the number of projects undertaken.

Table of Contents

List of Figures

List of Tables

1 Introduction to the Data Access, Storage, Sharing and Linkage proof-of-concept project

1.1 Project background

A vast amount of health and related data are routinely collected in Ireland for patient management, service planning and monitoring, and research. These national-, regional-, and organisational-level health datasets are usually more comprehensive and inclusive of individuals compared with data collected as part of a specific research study (1). Using these routinely collected data for research can result in more valid and accurate data, which facilitates data-driven decisions and reduces time demands on the participants as well as the researchers. However, despite the large effort and resources invested in the establishment and maintenance of these datasets, their full potential is not currently reached, and they remain underutilised. The ability to merge datasets by linking or matching individuals from two or more discrete data collections further expands the uses and benefits of these national datasets, resulting in even larger discoveries (2–5).

Currently, there is limited infrastructure to enable researchers to securely access health and social care data for research purposes, and access procedures differ across data providers. The Central Statistics Office (CSO) provides a secure processing environment for the sharing and linking of data for statistical purposes but there is no specific secure processing environment to enhance the sharing and linkage of health datasets, and no mechanism to provide sharing and linkage services for research purposes. This lack of a formal and secure infrastructure to integrate, link, and support remote access to data for secondary purposes, including for research, has led to valuable projects being inordinately delayed or, in some cases, abandoned.

To help address the challenge of how researchers and policy-makers can avail of one of our most valuable national assets – existing data – and use such data in a safe, secure manner while protecting the privacy and confidentiality of the data subjects, and in accordance with existing legislation, the Health Research Board (HRB) proposed the Data Access, Storage, Sharing and Linkage (DASSL) model in 2016. Initially developed based on international experiences and stakeholder input, the model was presented in the report titled *Proposals for an Enabling Data Environment for Health and Related Research in Ireland* (6), hereafter referred to as the HRB Report.

Similar to what has been implemented internationally to maximise the benefits from health data collections, this model aims to provide a single point of access to, and facilitate linking of, health data in a safe and trusted manner, with patient anonymity secured at all times.

Following the publication of the HRB Report and the positive response and debate that ensued among the research community and other stakeholders, the HRB issued a call for applications in 2019 to develop a proof of concept (PoC) technical infrastructure to generate learning that will inform the future development of a DASSL model in Ireland. While the scope of this HRB-funded study was for research and innovation purposes, it was clear at the outset that the learning from this study would be relevant to other secondary purposes (e.g. policy and planning, audit, and education and training).

Following a competitive peer-review process, the Irish Centre for High-End Computing (ICHEC), along with collaborators from the Royal College of Surgeons in Ireland (RSCI), the Health Service Executive (HSE), St James's Hospital, and Trinity College Dublin, were awarded HRB funding. ICHEC, which is part of the University of Galway, delivers complex computer solutions to Irish higher education institutions, enterprises, and the public sector on behalf of the State and manages the national high-performance computing (HPC) infrastructure.

1.2 Project scope, aims, and objectives

The aim of the PoC project was to develop and test a technical infrastructure and provide a demonstrator to support the DASSL model and inform the roll-out of a national service based on the proposed DASSL model.

The scope of the project was limited to linking health and related data for research purposes; however, other secondary uses of linked data are mentioned where relevant. Defining the governance and legislation was also out of scope for this project, which largely focused on the technical components, but governance is discussed, where relevant, based on the findings of this PoC project. Finally, no personal data were processed for this project; synthetic data were generated to test and demonstrate the potential of the technical DASSL prototype.

For the purpose of this report, the proposed national service for accessing, sharing, storing, and linking national health and related data is referred to as the national DASSL service. The HRB Report outlining the DASSL model reviewed and reported on the international and national health data sharing and linkage landscapes when it was written in 2016. To support the development of the prototype technical infrastructure (and related processes) in this project, the team conducted a landscape analysis to collect information on key developments (nationally or internationally) since the 2016 publication of the original HRB Report. A comprehensive review was undertaken, which included published and unpublished literature, reports, website information, and conferences. Additionally, information was gathered from stakeholders and experts in the field. One-on-one and group interviews were conducted with international experts, researchers, academics, members of the public, patients, data controllers (e.g. hospitals, national datasets, general practitioners (GPs)), data processors (private companies), the Health Information and Quality Authority (HIQA), the HSE, and the CSO.

Additionally, a DASSL PoC stakeholder group was formed, comprising representatives from the HSE, the HRB, the Department of Health (DOH), HIQA, members of the public, patients, other data controllers (Voluntary Hospitals, Office of the Clinical Information Officer), and researchers. This group provided invaluable input during the landscape analysis as well as on broader aspects of the PoC project, including the selection of use cases and consideration of implications for the final report.

This report describes the methods, approaches, principles, and assumptions that informed the delivery of the PoC infrastructure, the lessons learned from the case studies, and considerations for a national roll-out of a DASSL-type model.

2 Overview of the DASSL model

The HRB Report proposed that a new entity, the Research Data Trust (RDT), would be established to create the institutional and technical environment where the operationalisation of the DASSL model would take place. Data trusts usually provide repeatable mechanisms and approaches to sharing data in a timely, fair, safe, and equitable way (7).

An RDT encompasses each of the components of the data linkage model rather than necessarily being a specific legal entity or institution itself (8). In this instance, the proposed RDT (Figure 1) included seven components: governance; a health research data hub; a trusted third party data linkage unit (DLU); a safe haven; a research support unit (RSU); output checking and disclosure control; and public involvement and engagement.

The HRB Report described a process whereby a researcher would submit an access request to the DASSL RDT, along with the subsequent steps and safeguards to provide secure access to deidentified data for a research project. This is summarised in Figure 2, and the steps are outlined as follows:

1. **The researcher contacts the RSU and completes the necessary requirements, including submitting applications to governance boards.**
2. **On receiving the necessary approvals, data controllers provide personally identifiable data to the trusted third party (TTP) who links and pseudonymises/anonymises them.**
3. **Variables of interest with pseudo-identifiers are provided to the health research data hub, where they are stored.**
4. **The RSU supervises researcher access to the data within a secure operating environment (safe haven).**
5. **Once analysis is complete, the outputs are checked in order to mitigate any risks of disclosure.**
6. **Public engagement occurs throughout.**

The DASSL Model

Data to Public Benefits Committee

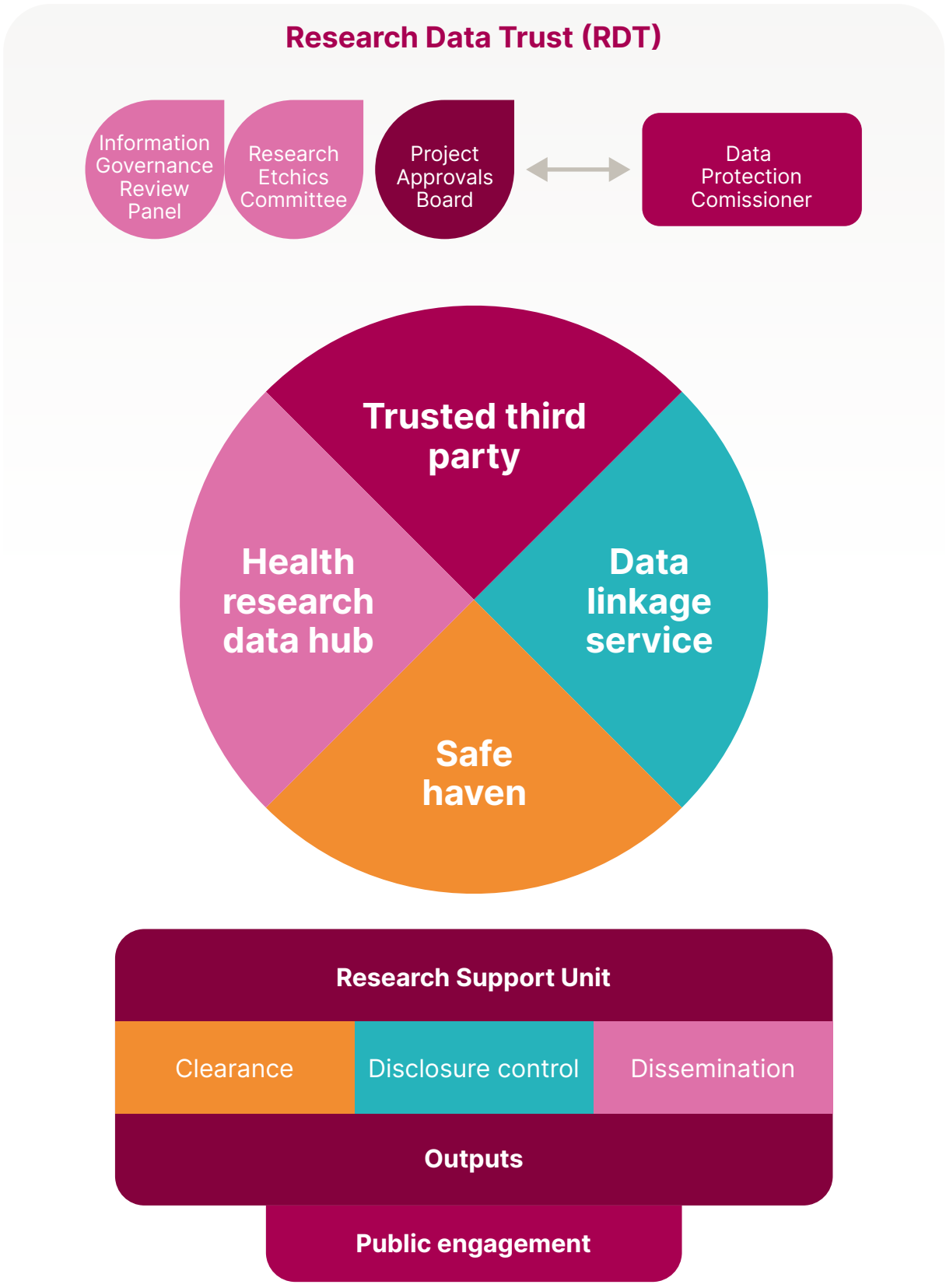


Figure 1 DASSL model
Source: Moran, 2016 (6)

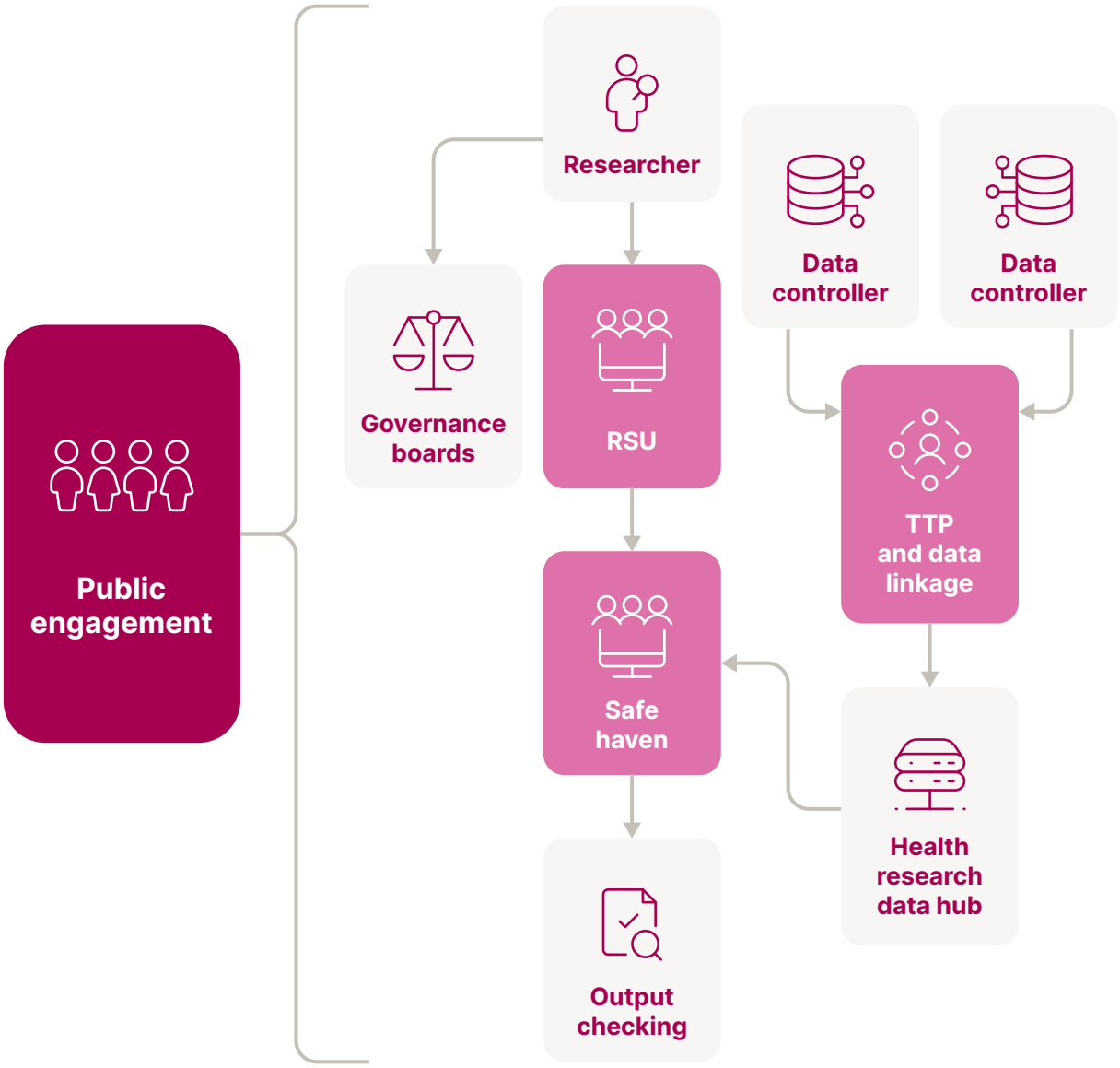


Figure 2 Proposed DASSL model process

A brief overview of each of the seven elements of the DASSL RDT, as described in the HRB Report, is set out below.

2.1 Governance

The HRB Report did not specify where the DASSL RDT should be established, as models differed across the countries examined; however, it acknowledged that further discussion on the necessity and utility of introducing special legislation to underpin the DASSL infrastructure and services would be required, along with a principled, proportionate, risk-based governance approach. This included researcher training, project approvals via authorising entities (i.e. a project approvals board, Information Governance Review Panel (IGRP), Research Ethics Committee (REC)), and data sharing agreements with data controllers. While defining the governance of the DASSL model is out of scope for this PoC study, a discussion on the aspects of governance required is covered where learnings from the PoC arose.

2.2 TTP data linkage unit

The DLU matches individuals across datasets based on their personal identifiers. The DLU can exist either as an independent organisation from the rest of the RDT, or as a distinct unit with its own infrastructure within the RDT. Therefore, those operating the DLU are often referred to as a TTP. The personal identifiers are separated from the content data by the data provider, and then shared with the DLU, which performs the linkage. The DLU then shares only the linkage key with the health research data hub, rather than sharing any personally identifiable information.

2.3 Health research data hub

The corresponding content data separated from the personal identifiers are sent to the health research data hub by the data providers, and the linkage key is sent by the DLU. These pseudonymised datasets are then either stored centrally within the health research data hub and regularly updated and maintained, or they may be gathered for a specific project and subsequently destroyed once the project is complete.

2.4 Research Support Unit

The RSU comprises the statisticians and other staff who combine the required datasets received by the health research data hub and provisions the combined datasets to researchers. The RSU also has a number of other roles in the process, including assessing and reviewing applications, assessing project feasibility, supporting the researcher and data providers in the process, providing access to and supervising the safe haven, and checking project outputs for statistical disclosure control. It may also manage publications and communications, put data sharing and licence agreements in place, manage delivery of safe researcher training and certification, and perform its own analysis when required and allowed.

2.5 Safe haven

The safe haven is a 'locked-down' and 'leak-proof' environment, containing statistical and analytical packages designed to allow researchers to safely conduct analysis on the data provided by the RSU. External network access, USB ports, CD drives, printing, and the taking of screenshots are disabled and the RSU checks any outputs from the analysis for disclosure control before they are released. Access to the safe haven could be on-site under direct supervision by the RSU or via virtual access, depending on the data provider permissions. Similarly, if the researcher would like to import their own data, code, or software packages into the safe haven, this would need to be assessed and imported by the RSU.

2.6 Output checking and disclosure control

Thorough checking of data being exported from the safe haven by highly trained statisticians in the RSU with expertise in disclosure control aims to ensure that individuals or entities cannot be identified, e.g. ensuring that cell counts in tables have a minimum frequency. On completion, an RSU team member would transfer the checked data files to the researchers. Data providers could also stipulate that the RSU check any final outputs, such as publications and presentations, in order to ensure that the data are correctly described and that the approved acknowledgement has been used. The RSU may also require notification of presentations/publications using the data for dissemination purposes, e.g. on the DASSL website, and to promote transparency.

2.7 Public and patient involvement and engagement

Education, consultation, and engagement with the public regarding the development and operations of the RDT are vital components of the DASSL model. Typical public engagement activities proposed included awareness raising and public education, and transparent information provision, whereas public involvement could include representation on advisory panels, RECs, and other governance structures; research on public attitudes; discussion forums; and a public engagement policy.

3 International RDTs

3.1 Overview of other models

While the DASSL model was largely modelled on the operations in Scotland and Wales (6), other notable RDTs that resemble the DASSL model have been established in Finland, Australia, and Canada (9–11). An overview of these national- or regional-level models (shown in Figure 3) and approaches elsewhere is provided in Sections 3.1.1–3.1.10, and learning from these models informed the development of the DASSL PoC infrastructure prototype and considerations outlined in this report.

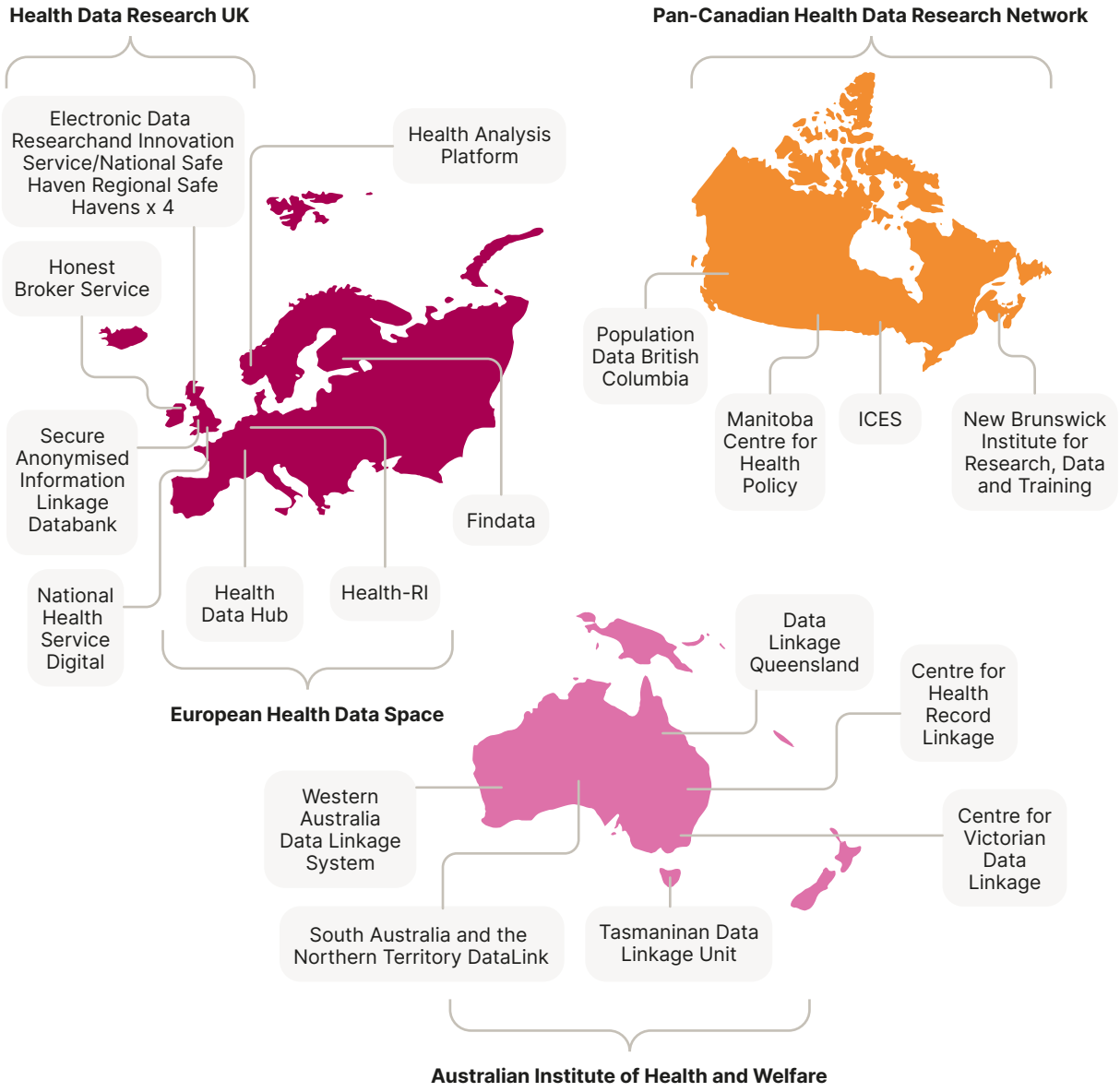


Figure 3 National data linkage models

3.1.1 Scotland

A Charter for Safe Havens in Scotland defines a safe haven as a specialised secure environment supported by trained, specialist staff, and where health data are processed and linked (12). The National Safe Haven facilitates access to national-level data from the National Health Service (NHS) and is hosted by EPCC (formerly the Edinburgh Parallel Computing Centre) at the University of Edinburgh. Whereas researchers requiring access to regional-level data can apply directly to one of the four Regional Safe Havens: Grampian Data Safe Haven (DaSH) at the University of Aberdeen; Lothian Research Safe Haven at EPCC, University of Edinburgh; Tayside and Fife Safe Haven at the Health Informatics Centre, University of Dundee; and Glasgow Safe Haven at the University of Glasgow. The blueprint for the National and Regional Safe Havens was originally proposed by the ScottisH Informatics Programme (SHIP) and was later funded by the Farr Institute, which has since been replaced by Health Data Research UK (HDR UK) with the aim of increasing collaboration in health data sharing and usage across the United Kingdom (UK) (13).

Access to the National Safe Haven is the responsibility of the electronic Data Research and Innovation Service (eDRIS) in Public Health Scotland (14). eDRIS accreditation for researchers includes completing one of the data protection or safe researcher training sessions, having a proven track record, and endorsement by an approved public institution (14). It also supports researchers in completing their project application, assessing the feasibility of their research project, and gathering data on a project-by-project basis. The project application is then reviewed by the independent Public Benefits and Privacy Panel for Health and Social Care (HSC-PBPP), equivalent to the Health Research Consent Declaration Committee in Ireland, for information governance and public interest assessment (15). Ethical approval is sought if required (14).

On approval, the matching variables from each dataset are directed to the National Records of Scotland/NHS National Services Scotland to perform the data linkage. Different pseudo-identifiers are sent back to each data controller to replace the personally identifiable data and a linkage key is sent to the National Safe Haven to facilitate the combination of datasets. The data view is made available to the research team via an on-site or virtual safe haven, and virtual access may be restricted to those with a strong track record in research (16) and with approval from the data controllers (14).

eDRIS assesses any outputs for statistical disclosure control (17). Unlike the National Safe Haven, the Tayside and Fife Safe Haven stores pseudonymised copies of datasets on behalf of NHS Tayside and Fife (18) and has developed its own open-source research data management platform (RDMP) in order to facilitate both data management and linkage (19).

3.1.2 Wales

Established in 2007, the Secure Anonymised Information Linkage (SAIL) Databank in Wales is hosted by the Health Informatics team at Swansea University and is part of HDR UK. Strategic direction is provided by the SAIL Management Team, with guidance from an international scientific advisory committee. Researchers from public institutions may apply to access data by contacting a data analyst at SAIL to assess the feasibility of the project. The researchers must then submit a project application to an independent IGRP, comprising of representatives from professional and regulatory bodies along with members from the public, to review the public interest and sensitivity risk (10). Ethical approval is not required for the use of anonymised data in the UK, but there is an ethics committee representative on the IGRP, and researchers requesting to link their own datasets may require ethical approval and consent (10, 20). Additionally, the researchers must complete the approved data protection or safe researcher training (20, 21). Privacy notices are provided on behalf of data controllers; SAIL is not a data controller as it reportedly does not have access to, or control over, the personally identifiable data.

The SAIL Databank holds pseudonymised national datasets, including clinical data abstracted directly from GP systems using the Audit+ system (22); other types of data, such as data on education, housing, and employment; and emerging health data types, such as genomic, free text, and imaging data (10). The data controller uploads File 1 (i.e. matching variables) and File 2 (i.e. variables of interest) to the secure upload websites of the NHS Wales Informatics Service (NWIS) and SAIL Databank via their unique account. They may reserve the right to review the research project proposals related to their dataset(s) alongside the IGRP.

After the data controller splits their dataset into two files, the matching variables are directed to NWIS, which matches them to the Welsh Demographic Service Dataset (WDSD) using a Structured Query Language- (SQL-) based matching algorithm, Matching Algorithm for Consistent Results in Anonymised Linkage (MACRAL) (10) and an automated 'black box' system where no one sees the identifiable data being processed (10). Lexicon matching and Soundex matching are used, which match Welsh-specific variants in registered names and use variant phonetic spellings of the forename or surname, respectively (10). Likelihood ratios are calculated using a Bayesian approach of prior and posterior odds, by taking into account the distributions of the set of variables on the WDSD. The NHS number is then encrypted and becomes an Anonymous Linking Field (ALF), which replaces the personally identifiable data and is shared with the SAIL Databank along with minimal demographic data (including sex, week of birth, and area of residence) and the thresholds of match accuracy. SAIL then re-encrypts the ALFs and mask practitioner codes, and uses Residential ALFs (RALFs) to allow researchers to associate individuals within the same home and geographic region (10).

Preparation of the data view for the researcher sees further data minimisation (e.g. aggregation and suppression) and encryption of the ALF so that researchers running more than one project cannot use this key to link across discrete studies (10). These data are then made available to researchers via the UK Secure eResearch Platform (SeRP), also referred to as the SAIL Gateway, which uses a VMware Horizon infrastructure. To support specific project needs, other UK SeRP components can be made available – such as the HPC cluster or Kubernetes cluster to support processing pipelines, or a graphics processing unit and artificial intelligence (AI) cluster for training computing models – as well as collaboration through data space, file storage, wiki pages, Git (version-controlled repository system), and wider support and help materials. The proposed outputs from the UK SeRP are scrutinised for disclosure risk by a SAIL senior analyst. Occasionally, data are released via other secure environments with the necessary approvals.

3.1.3 England

Recent developments in England include the publication of *Better, broader, safer: using health data for research and analysis* (the "Goldacre review") in April 2022 [23], which informed the national data strategy for England, *Data Saves Lives: Reshaping Health and Social Care with Data,* published in June 2022. The Goldacre review evaluates how best use can be made of NHS data for research and analysis, and provides a range of recommendations, centred on data platforms, security, open working methods, data curation, data analysts, governance and strategy [23]. Key recommendations focus on actions to increase trust, privacy, and transparency and on creating shared 'Trusted Research Environments' (TRE), secure analytical platforms that would be used for all analysis of NHS patient records data (unless patients have consented for further transfer of their data). OpenSAFELY [24] is an example of a TRE currently operating, which facilitates federated data analysis across 58 million patients' GP records and other datasets, underpinned by privacy, transparency, and open science. Established during the COVID-19 pandemic, it has been used to generate evidence on how the pandemic impacted delivery of care, the course of COVID-19 illness and effectiveness of treatments and vaccines. Rather than being a single TRE, OpenSAFELY is portable software that can be deployed where health data is already stored. It has been deployed in multiple NHS settings, allowing federated analysis across data centres.

3.1.4 Northern Ireland

In Northern Ireland, the Health and Social Care (Control of Data Processing) Act (Northern Ireland) 2016 supports the secondary use of health data. Health research projects proceed under the Honest Broker Service (HBS), while non-health projects fall within the scope of Administrative Data Research Northern Ireland (ADR NI).

Researchers from approved public organisations or internal staff members (for audit, public health monitoring, etc.) can request access to health data via the HBS, which is the Trusted Research Environment for Health and Social Care (HSC) Northern Ireland and is hosted within the HSC Regional Business Services Organisation (25).

Projects are reviewed by the HBS Governance Board. However, this does not support requests for access to identifiable data for consented studies or trials. Available data include those from the Regional Data Warehouse and Family Practitioner Services, and metadata are also available. For ethically approved research requiring linkage to data not in the Regional Data Warehouse, individual data access agreements will be required and measures will need to be put in place to protect confidentiality. Ethical approval can be sought via the NHS or the HSC REC. There is also a cost recovery service in place. The Health Data Research Northern Ireland UK Secure e-Research Platform (HDRNI UK SeRP) is the safe haven that supports researcher access to the pseudonymised data and has a range of analytical and statistical tools, as well as facilities to share code and findings with their approved team members. Remote access to this platform was also planned at the time of report writing.

3.1.5 Finland

Findata began operating in 2020, with the introduction of the Act on the Secondary Use of Health and Social Data (26). Consumers, including those from private companies, can request access to pseudonymised personal data via a data permit, or to statistical data (i.e. aggregated and analysed data) via a data request (27) for one of the seven purposes outlined in the Act on the Secondary Use of Health and Social Data.

Staff at Findata who review research project applications for combined datasets may contact the data custodian regarding feasibility and costs. In 2020, 318 applications were received and close to 100 data controllers were approached. However, if only one dataset is required, the researcher submits their request directly to the data custodian and Findata will still anonymise the data on behalf of the data custodian.

The cost recovery model is determined based on the type and location of the researcher (e.g. students within the European Economic Area (EEA)), the level of pre-processing required, data controller costs for data extraction, and use of the secure environment. These data are then gathered on a project-by-project basis from public and private social and healthcare providers (including data from ePrescriptions) and, at the time of conducting the landscape analysis for this report, plans were underway to facilitate access to data from the national Electronic Health Record (EHR) Kanta (27). In addition, subject to legislation being enacted, Findata may also facilitate access to biobank data.

The data custodian encrypts the data and uploads them to Findata via a Nix cloud interface, which is restricted by strong authentication and Internet protocol (IP) addresses. However, this process can reportedly be difficult for non-technical researchers, and data controllers would prefer an application programming interface (API) that interfaces directly with their dataset, rather than a third-party site. The unique personal identifiers. These pseudo-identifiers are stored to enable re-identification used to combine the datasets and are then replaced with pseudo-identifiers and stored in order to support the data subject's right to object under the General Data Protection Regulation (GDPR) or to notify the data subject of an important discovery. No specific software is used for handling the data. The pseudonymised data are then made available to the researcher in a secure remote user environment via an Internet browser. All virtual machines include statistical software packages such as Statistical Product and Service Solutions (SPSS) Statistics, Stata, SAS, and R (27), and the computer processing power is customisable based on the research requirements (28). Any additional code or packages required by the consumers can be sent to Findata, which are made available in the secure remote user environment once checked. Data may be transferred to another environment if necessary and only if that environment meets the security requirements. The outputs from the secure environment are sent to Findata to ensure that they are anonymised, and are then released. Data are then archived for a maximum of 5 years (27).

3.1.6 France

The French Health Data Hub (HDH) was set up by a legislative decree in November 2019 with the aim of facilitating data sharing and exploitation in high-level security conditions (29). This was in response to the challenges of using AI in healthcare, which were highlighted in a 2018 report by Villani *et al.* (30). The HDH was put into service in April 2020 to manage the COVID-19 pandemic.

The Ethics and Scientific Committee for Research, Studies and Evaluations in the Field of Health (*Comité Éthique et Scientifique pour les Recherches, les Études et les Évaluations dans le domaine de la santé:* CESREES) was also established to assess project applications to use personal health data without implicit consent, and is made up of a network of external experts (31). Following approval by CESREES, authorisation from the French Data Protection Authority (*Commission Nationale de l'Informatique et des Libertés*; CNIL) may be requested.

A catalogue of datasets available to both public and private users (e.g. researchers, start-up companies, healthcare professionals) (32) has been built in a progressive and iterative manner in partnership with data controllers, and it is planned that copies of the most relevant datasets (with personally identifiable information removed) will be stored and updated regularly on the HDH technological platform (31). The catalogue will comprise 18 databases from the National Health Data System (Système National des Données de Santé: SNDS), including genomics, laboratory data, mobile health (mHealth) data, and patient-reported outcome measures (33).

This platform was initially a partnership with Microsoft Azure but, following concerns with data sharing outside of the European Union (EU), a solution provided by a European or French company was being explored at the time of conducting the landscape analysis for this report. Until this transition, the French HDH signed a contract with Microsoft's Irish affiliate in order to ensure that data are hosted in data centres within the EU (34). A virtual secure project space with analytical tools is provided by the HDH when required (31), and data linkage is facilitated by the French national identifier (35).

Another project, Digital Health Space (*Espace Numérique de Santé:* ENS), aims to provide an eHealth personal space to allow citizens to access their health records from birth and to provide informed consent regarding data collection, use, and dissemination. In conjunction with the HDH, this may also support patient enrolment in clinical trials and predictive services that send personalised advice to patients.

3.1.7 Sweden

Registerforskning.se, operated by the Swedish Research Council, provides researchers with information on existing registers, as well as support during the process of conducting register-based research (36). It has also developed the Register Utiliser Tool (RUT), which enables efficient searching and matching of standardised metadata in health registers, biobanks, government registries, and research databases. The RUT directs the user to the data controller that they must contact for access to the dataset, as opposed to providing a central access point to Swedish public authority data. Ethical approval is also required for register-based research in Sweden, and this research is conducted under the GDPR.

3.1.8 Norway

In 2018, the Norwegian Government implemented the Health Analysis Platform (HAP), which was proposed by the Directorate for e-Health and which will be supported by a legal framework that was under review at the time of conducting the landscape analysis for this report. Two components are planned in order for this national infrastructure to facilitate access to health data, including a permit authority and a platform for users.

It proposes to include four ecosystems: analytical tools, actors working with health analytics, data providers, and interaction with other national and international data ecosystems. It is planned that this data platform will have copies of data from the health registers, health examiners, and biobanks, and a common analysis infrastructure to facilitate the public and commercial ecosystem.

There will also be support for production of synthetic data and exploratory analysis services (37). Data users will then be able to access these data via the HAP, another secure environment, or their own local environment. Patients will also be able to see what their data are being used for, and may opt-out. However, development of the HAP had been put on hold at the end of 2021 due to legal challenges as a result of the Schrems II judgement.

3.1.9 The Netherlands

The Health Research Infrastructure initiative (Health-RI) in the Netherlands is building a national health research infrastructure to support access to knowledge, tools, facilities, health data, and samples. Its workplan includes national alignment on the ethical, legal, policy, and governance frameworks on data collection and reuse, and on supporting Findable, Accessible, Interoperable, and Reusable (FAIR) data and distributed access via regional nodes and a central hub (38). Another initiative is the Personal Health Train, which supports citizen control over their own data and which involves the researcher travelling to the data source rather than data from various sources having to be transported to the researcher (39).

3.1.10 Australia

Several centres conduct data linkage in Australia (40) including:

- **National or Commonwealth data:** Australian Institute of Health and Welfare
- **Western Australia:** Western Australia Data Linkage System (WADLS)
- **Northern Territory and South Australia:** South Australia and the Northern Territory DataLink (SA NT DataLink)
- **Queensland:** Data Linkage Queensland
- **New South Wales (NSW) and Australian Capital Territory (ACT):** Centre for Health Record Linkage (CHeReL)
- **Tasmania:** Tasmanian Data Linkage Unit, and
- **Victoria:** Centre for Victorian Data Linkage.

Additionally, the Centre for Data Linkage (CDL) at Curtin University was tasked with "establishing a secure and efficient data linkage system to facilitate linkage between jurisdictional datasets, and between these datasets and research datasets using demographic data" (11) p2 under the Population Health Research Network (PHRN) Initiative (11). Curtin University has also developed privacy preserving record linkage (PPRL) techniques that are employed internationally using LinXmart. These are included within SeRP and facilitates both clear data linkage and PPRL (41).

Linked data research in Australia is allowed under The Privacy Act 1988 on the basis of ethical approval (42), which is usually obtained from a local REC (43); however, the Australian Institute of Health and Welfare has its own REC for access to national-level data (44).

For the purpose of the landscape analysis for this report, we largely focused on the practices of WADLS and CHeReL (which are managed by the relevant Ministries of Health) due to their experience and international recognition (45, 46).

Researchers accessing data via WADLS or CHeReL specify how the data will be stored and managed (45, 46) and may employ the Secure Unified Research Environment (SURE) (47) or E-Research Institutional Cloud Architecture (ERICA) (48). There are instances of ERICA operating at University of New South Wales (UNSW) Sydney, the Australian Institute of Health and Welfare, and the NSW Data Analytics Centre (NSW Secure Analytics Lab), with further use of ERICA planned for the University of Melbourne, SA NT DataLink, and the University of Western Australia (49).

WADLS first began operating in 1995 and is managed by the Western Australian Department of Health (46). WADLS was originally established to broker access to different datasets under a distributed model (50), but it now stores partial copies of datasets in its Custodian Administered Research Extract Server (CARES) (51, 52). The Linkage Team, which is physically separate from the Research Data Services team, conducts the routine linkage of core datasets and performs geocoding and genealogical links via the Family Connections System (46). While the Linkage Team originally used a proprietary software to probabilistically link data, this no longer met its needs, leading to the development of a new linkage strategy, Data Linkage System Number 3 (DLS3) (53). Linkage keys are then encrypted and different encryption keys are used for each request. The linked data extracts are then released directly to applicants with the applicable ethical, research, and data governance approvals. However, sometimes a secure third-party environment must be used (53).

Established in 2006, CHeReL is hosted by the Cancer Institute NSW (11). Although CHeReL does not store clinical data centrally, it retains a Master Linkage Key for the core datasets, which includes a unique Person Number and an encrypted record number from the data source (47). Where researchers request access to data that are not included in the Master Linkage Key, they must approach the data controllers themselves (45). Both the CDL and SA NT DataLink also store an index of linkage keys but do not hold the clinical data (21, 47). At CHeReL, the proprietary software ChoiceMaker is used to link data (45). However, a pilot project, Lumos, is gathering and linking data from GPs with other NSW data collections using PPRL techniques and secure File Transfer Protocol with support for Transport Layer Security (54). These data will then be made available to researchers in the Secure Analytics Primary Health Environment (SAPHE), which is a custom-built, secure, remote-access computing environment and cloud repository (54).

3.1.11 Canada

Canada has 10 provinces each of which have jurisdiction over their own health data. Population Data British Columbia (PopData) (55), the Institute for Clinical Evaluative Sciences (ICES) in Ontario (56), the Manitoba Centre for Health Policy (MCHP), and, more recently, the New Brunswick Institute for Research, Data and Training (NB-IRDT) at the University of New Brunswick facilitate access and linkage to health and other administrative data in their respective provinces.

The pan-Canadian Health Data Research Network (HDRN) has also been developed, which will see source data remain within jurisdictional boundaries and aggregated results pooled across jurisdictions using a federated approach (57), as well as the development of the Strategy for Patient-Oriented Research (SPOR) Canadian Data Platform. Paprica *et al.* (2020) (58) have also developed minimum specifications for the establishment and operation of RDTs, including a legal basis; adaptive, flexible, and accountable governance; well-defined policies and processes in relation to data protection and risk management; data user requirements in relation to agreements and training; and ongoing public and stakeholder engagement.

For the landscape analysis for this project, we largely focused on the operations of PopData, ICES, and MCHP. Under the respective provincial legislation, each of these RDTs can legally collect personal health information for named purposes.

Founded in 1992, ICES has a network of seven physical sites across Ontario, including a central location at the Sunnybrook Health Sciences Centre in Toronto (56). ICES is a named entity under the Personal Health Information Protection Act, 2004 (PHIPA) and can collect personally identifiable data; it also facilitates both internal (ICES researchers) and external (researchers from third-party organisations) projects. A privacy impact assessment (PIA) is conducted for internal projects by the ICES Privacy & Legal Office, while the feasibility of external projects is assessed by ICES Data and Analytic Services (DAS); the ICES Privacy & Legal Office also ensures that external projects have ethical and other approvals that are not required for internal projects. ICES stores Ontarians' health data (including GP and novel data) in its repository, which consists of SAS datasets and Microsoft SQL databases with role-based access control. Only a restricted group of staff has permission to handle the fully identifiable data (59).

Records are linked to the Registered Persons Database (RPDB), which includes every individual who has been issued an Ontario Health Insurance Plan (OHIP) card and is updated on a monthly basis. However, not every dataset collects the OHIP number, and data sources are linked deterministically and probabilistically using Automatch software and following the Fellegi-Sunter method. Data are standardised by implementing the New York State Identification and Intelligence System (NYSIIS), phonetic conversion, and blocking files in order to optimise the scanning process.

ICES researchers access the data via the Research Analysis Environment (RAE), which houses more than 5 terabytes (TB) of data, uses more than 80 central processing unit (CPU) cores and more than 1.2 TB of random-access memory (RAM), and is stored in a secure, isolated network at ICES Central where it can be accessed by other sites via a private network (59). It provides analytic tools such as Python, SAS, Stata, and R, and the systems use a mix of Red Hat Enterprise Linux and Windows servers on an Active Directory domain. The back end consists of virtual machines running on Cisco Unified Computing System (UCS) server blades and NetApp all-flash storage managed with VMware.

The Health Artificial Intelligence Data Analytics Platform (HAIDAP) is an extension of the RAE, which is a HPC cluster in a private cloud environment built to accommodate greater computational resources and specialised software packages for AI, machine learning, and natural language processing. The ICES Data and Analytic Virtual Environment (IDAVE) provides external researchers with remote access, and each user is able to load additional software from an approved catalogue. ICES DAS assesses requested outputs from the IDAVE for reidentification risk, and vetted files are emailed to the researchers. Through mechanisms such as focus groups and a public advisory council, members of the public are involved and engaged in the data-intensive health research at ICES.

PopData was established in 2009 and is a multi-university data and education resource physically located at the University of British Columbia (60), and is a named entity under British Columbia's Freedom of Information and Protection of Privacy Act (FIPPA). The policies that PopData implements are created in consultation with a Data Stewards Working Group and follow the Research Data Access Framework, which requires research projects to be in the public interest and have ethical approval (60).

The Data Access Unit facilitates researchers to submit their applications via an online Data Access Request, which is reviewed by the relevant data stewards. Following approval, researchers must complete privacy training and sign a number of agreements. Matching variables are linked to the Population Directory, which is maintained by PopData using software developed by PopData staff employing a combination of deterministic and probabilistic linkage and clerical review for competing matches. The linkage identification (ID) is then converted to a PopDataID. PopData houses the content data on an internal server that is separate from the identifiers and stores these data in a secure, climate-controlled room to which only a limited number of staff have access.

When data are provided for a research project, the PopDataID is replaced using a research project-specific key. It takes an average of 4–6 months for data to be made available in the Secure Research Environment (SRE) from anywhere in Canada, using two-factor authentication. The SRE contains a comprehensive set of software including SAS, Stata, R, and Python. Exceptions may be agreed with the data stewards for release of data to external safe havens as well. Users of the SRE self-vet their analytic output, and these exports are systematically scanned for type, size, and content, with suspicious files blocked for manual review and some files checked at random. Copies of all file transfers are then archived. At the time of conducting the landscape analysis for this report, the Health Data Platform BC was also being developed on SeRP in British Columbia in order to facilitate access to other government data using a federated approach (61).

Established in 1991, MCHP is located at the University of Manitoba in Winnipeg (62) and is a named entity under Manitoba's Personal Health Information Act (PHIA). The centralised Manitoba Population Research Data Repository includes more than 90 datasets from government departments (Manitoba Health; Manitoba Education and Early Childhood Learning; Manitoba Justice; etc.), provincial laboratories, clinical programmes, community and social outreach organisations, and Indigenous governance bodies (63). The need for consent to use these data for research is usually waived under the PHIA. Researchers must be accredited and then submit a feasibility request to MCHP. All research projects must then be reviewed by individual data providers, the Manitoba Health Information Privacy Committee, and the University of Manitoba Health Research Ethics Board.

The data provider sends the demographic data and an internal reference number to the Information Management and Analytics Unit of Manitoba Health, which is used to match each individual to their existing nine-digit Personal Health Identification Number (PHIN) using a custom-developed software package called LINKPRO (63). The PHIN is encrypted using a consistent, standard algorithm and is permanently stored with each record. The data provider also sends the internal reference numbers and the variables of interest to MCHP, and these are then linked to the encrypted PHIN.

The repository data are stored in a SAS-based SQL server with user- and project-level access controls. Analytic systems are supported on Microsoft Windows servers, with tools for data acquisition and handling, MCHP internal analytics, and remote access within Manitoba (63). Remote access is supported by Microsoft Windows-based computers and requires unique individual accounts with two-factor authentication as signed by MCHP (63). The platform supports SAS as the default analytic environment, although Stata and R software are also available. Remote access from physically secure locations can be arranged with the appropriate approvals. Information taken from the MCHP analytic systems must be aggregate or statistical in nature and is manually reviewed. Researchers preparing to publish findings using the Manitoba Population Research Data Repository data must submit their findings for review (63).

4 Key learnings from other models

Key learnings arising from the review of national health data sharing and linkage platforms and models elsewhere are summarised in Sections 4.1–4.15 and are highlighted throughout this report.

4.1 Governance

Access to linked health data in the countries discussed in Chapter 3, e.g. Australia, Canada, Finland, France and the UK, is provided at the national and, in some cases, regional level. A national or international network is often then created to connect organisations and other RDTs, such as HDR UK. In the UK, all or some of the components of the data linkage models are situated within the NHS. Other countries' data linkage models are hosted by universities or were established within, or by, the countries' departments of health. Still others use a combination of government and university models, such as SA NT DataLink, where the linkage staff are employed by the government but those managing the content data are employed by the university for data governance purposes. Finally, many of the more recent RDTs in Europe have established new not-for-profit organisations such as the French HDH, Health RI in the Netherlands, and Findata in Finland.

4.2 Lawful basis

While some countries (e.g. Scotland and Wales) rely on existing legislation and public interest to promote scientifically and ethically robust research, other countries (such as Finland, Canada, and Australia) introduced legislation that provides for secondary use of data collections, the use of health data by named entities, and linked data research on the basis of ethical approval, respectively. Notably, Findata becomes the data controller of the data that it receives, while SAIL becomes a data processor. Explicit consent is rarely sought for linking national health datasets due to impracticalities such as the data being derived from routine public service delivery, meaning that there is no direct contact with the data subjects.

Therefore, the RDTs often rely on the GDPR provision for processing special categories of data for statistical purposes in the public interest and on authorising boards to ensure that the processing of such data is in the public interest. However, where researchers request to link routinely collected data with their own research data, explicit consent is usually required.

4.3 Research project approval

Research projects are usually first assessed for feasibility by a Research Support Unit (RSU) and, depending on the data sharing agreement, the researcher or the RSU may seek approval from the data controller. A research project application, including a description of the project, the data required, a PIA, and data storage, is then completed. This application is assessed by an approvals board and ethical approval is sought where necessary (along with the equivalent of a Health Research Consent Declaration Committee (HRCDC) declaration in the case of Scotland). Research ethical approval is a prerequisite for linked data research and accessing data via a safe haven. However, issues with inconsistent reviews by local RECs have led to national or regional RECs in some countries reviewing requests. Input from public advisory committees may also be sought (e.g. CHeReL in Australia) if the public interest of a project is not immediately clear.

4.4 Researcher accreditation

It remains common practice to require researchers to complete safe researcher, information governance, or privacy training in order to be allowed to access data via safe havens. Researchers may also be required to demonstrate a proven track record and association with an approved public institution. Some RDTs allow industry access either themselves, via a partnership with researchers from an approved institution, or via summary- or aggregate-level findings for projects in the public interest and not solely for proprietary reasons. The researcher usually also signs data access, confidentiality, and/or privacy protection agreements.

4.5 Advisory committees and accountability

Most RDTs are independent not-for-profit entities, situated in one or more universities that work closely with, or are funded by, the country's government and health service. These RDTs are then independently audited and have boards of directors that act as sounding boards and that provide valued input into their activities.

4.6 Public engagement practices

Most RDTs have established public or community advisory panels made up of a diverse group of people. These panels meet several times per year (with or without RSU staff and researchers present) in order to provide feedback on activities, research studies, new data opportunities and partnerships, business plans, and policies and procedures, and to guide the development of research questions and public-facing content and activities. Additionally, members of these panels have been appointed to other, more general advisory boards, IGRPs, RECs, and interview panels. In some of the countries studied, such as Wales, there are clearly defined roles and responsibilities, as well as education on data linkage and governance, for members of these panels.

Other activities aimed at increasing public engagement may be organised by a public engagement officer. These activities often include public education forums, presentations to diverse groups across the country, workshops and focus groups, deliberation methods, citizens' juries, public surveys, and encouraging public and patient involvement (PPI) in research projects.

4.7 Location and responsibilities of RSUs

Most RSUs are co-located with the data hub, with the exception of eDRIS in Scotland, which is part of Public Health Scotland. Similar to the RSU proposed in the HRB Report, many international RSUs assist researchers with project design and application; certify that researchers have completed the necessary training; advise on the availability, strengths, and weaknesses of datasets; provide expert advice on coding, terminology, and metadata; ensure that all permissions and agreements are in place; prepare metadata and documentation; and conduct statistical disclosure controls. Some RSUs deliver researcher training themselves, while others certify that researchers have completed an approved data protection and/or safe research training programme. Additionally, they may assist researchers with their analysis, or perform analysis, if required, for government or private organisations. They usually also check inbound files (such as code, statistical packages, and datasets) to the safe haven for viruses and malware, as well as ethical compliance.

Depending on the data controller agreements in place, the RSU may liaise with the data custodians regarding each research project application, or will facilitate researchers to liaise with the data custodians. The RSU must also have a comprehensive understanding of individual datasets, especially where data are centralised. RSU staff typically provide content for public-facing websites, which provide a data catalogue and metadata, and inform the public about data use and research findings.

4.8 Data processing

A catalogue of available datasets and metadata is usually made openly available online. In Norway, researchers can use the metadata to build their cohort, while in Sweden, the RUT only provides the metadata and data access points as opposed to processing the data itself. Some data hubs store pseudonymised copies of national or regional datasets, and those in Canada store personally identifiable and clinical data separately from each other. Other data hubs, such as the National Safe Haven in Scotland, Findata in Finland, and CHeReL and SA NT DataLink in Australia, gather these data from the data controllers on a project-by-project basis. Some data hubs also store non-health-related government or administrative data. Hybrid models also exist that initially gathered data on a project-by-project basis and now store some key pseudonymised datasets in a central repository. Centralisation of these data for research can be a more efficient process compared to on-demand, project-specific requests for data on a case-by-case basis, as it minimises demands on data providers and allows data quality to be monitored. However, it requires effort and resources to gain an understanding of the datasets, assess and improve the quality and standardisation of a large volume of data, store the data, and update large quantities of expanding data. The success of WADLS was also attributed to the trust that was built up with data controllers.

In some cases, researchers and/or data controllers may also upload their own dataset into a data hub for a specific research project or for use by other researchers. Data sharing agreements between data hubs and data providers are put in place which determine the data processing that may occur, with or without notifying the data controller about every project. Federated models are also used, which mitigate the need for data to leave the source, but using these models requires high-quality data.

4.9 Software and hardware

Data hubs utilise software and hardware in order to receive, store, clean, and standardise data. Some RDTs use specifically developed systems which include tools that de-identify, link, clean, standardise (e.g. natural language processing (NLP) tools), and analyse data (including AI), as well as the ability to store genomics and images. These include the customisable SeRP used in the UK, Australia, and Canada; the open-source research data management platform (RDMP) at Tayside and Fife Safe Haven in Scotland; and the CARES in Australia. Proprietary software has also been used, such as IBM Db2 (10), SAS, and Microsoft SQL databases. Data extraction from EHRs or GP records is also usually completed automatically by software.

4.10 Security and data protection

Data hubs employ many security and data protection measures, including: access control limits; closed-circuit television (CCTV) monitoring; multi-factor authentication; access logs; audit trails of data access, edits, and erasure; daily backups; hypertext transfer protocol secure (HTTPS) upload facilities; separation of identifiable and clinical data; climate-controlled rooms; no Internet connection; monitoring of Internet traffic and blocking attack attempts; antivirus software; internal and perimeter firewalls; network segmentation; weekly automated penetration and vulnerability scanning; data encryption; rotation of data off-site; and locked compartments. Data hubs usually also have International Organization for Standardization (ISO) 27001 certification.

4.11 Trusted third party and split file approach

Matching or linking the same individual – or, in some cases, individuals from the same family or household – across datasets is usually conducted using the split file approach or separation principle. This approach means that matching variables are separated from the variables of interest by the data controller and sent to a trusted third party (TTP). The TTPs used by SAIL in Wales and the National Safe Haven in Scotland are situated within the NHS (10), whereas other TTPs are in-house but located in a physically separate area than those with access to the variables of interest. Although RDTs such as ICES can legally receive fully identifiable clinical data, they still only grant specific individuals the highest level of access to identifiable personal data.

4.12 Linkage process

Matching variables are usually cleaned and standardised prior to linkage using fully or semi-automated processes, such as Soundex codes (which reduce strings (e.g. names of individuals) to four characters) and algorithms that remove spaces and capitalise names. A combination of two types of linkage methods are utilised for identifying the same individual across different datasets:

- ***Deterministic linkage*** looks for exact matches between datasets (e.g. unique identifiers).
- ***Probabilistic linkage*** estimates the probability that two records relate to the same person based on the chosen threshold.

These linkage methods often use 'blockers' (which require specified variables to have some degree of similarity before the records are compared) and experimentation with the cut-off weights for probability, as well as manual review when necessary. Datasets are compared with each other, or with a comprehensive population register such as a population directory or census data. A linkage quality report is usually produced regarding the strategies used and outcomes for each linkage step (e.g. linkage weights) so that researchers can report these and take them into account when interpreting their findings. The index of links from datasets is often stored by the data hub if it is centralising the dataset, or by the TTP if it operates a distributed model (such as CHeReL and SA NT DataLink in Australia) and has the data controller's permission. However, other RDTs destroy the links after every project. These linkage methods are also employed using encrypted information with or without the use of a TTP, which is referred to as PPRL. However, PPRL requires a huge amount of collaboration between the data controllers, as well as high-quality data. TTPs use their own software, open-source software, or proprietary software to link records. Some RDTs had to develop their own systems in order to enable comparison of more than two datasets at any one time, with up to 70 different datasets being compared at the same time.

4.13 Pseudonymisation and/or anonymisation

Personally identifiable data are encrypted or replaced with a pseudo-identifier, which allows them to be combined with the relevant variables of interest. The data hubs usually receive the encrypted linking field and linkage key from the TTP and may encrypt these further. However, in Scotland, the TTP sends the pseudo-identifiers to the data controllers to replace personally-identifiable data rather than directly to the data hub.

4.14 Safe havens and secure analysis environments

Most RDTs provide user access via safe havens. Some RDTs allow data to be transferred to another secure user environment that meets their stringent security measures if absolutely necessary, or they allow the release of anonymised aggregate-level data without the use of a safe haven. Additionally, safe havens have been made available to researchers to store and access research data that they have collected and control. Use of a safe haven and the preparation of the data view incurs a cost that takes into account the number of datasets, variables, and users, and the length of time the data view is to be stored. Only researchers listed on the original application, and who have been approved and completed the required training, can access the safe havens using multi-factor authentication. Traditionally, all safe havens were on-site, but now most RDTs facilitate virtual access via a virtual private network (VPN) and proprietary platforms such as VMware and Citrix. However, virtual access may be restricted to those with a strong, proven track record in research and with approval from the data controllers, in a specific location, or with a Windows or other desktop. At the end of the project, access to the safe haven is terminated and the data view is usually hibernated or archived for a specified time period (14, 27, 48, 64).

Safe havens have pre-installed statistical packages such as SQL querying tools, Microsoft Office, SAS, SPSS Statistics, Stata, R, and Python. Novel and advanced analytics (e.g. neural networks, NLP, machine learning tools) are now also being requested and made available in some safe havens, as well as business intelligence tools such as Tableau, R Shiny, and Microsoft Power BI. However, some applications are made available for a small licensing fee, or users may request to bring their own licence for a software package or import their own non-data files (e.g. codes), which must be checked. In addition, different levels of computer power are made available as well as access to HPC.

Shared project spaces may also be enabled in order to facilitate collaboration through databases, file storage, wiki pages, and Git repositories, as well as access to wider support and help materials. Other extensions to safe havens have included data translation systems to shift between data formats and ontology mapping systems. Dedicated file spaces usually exist within the safe haven for researchers to store the findings they wish to export.

4.15 Output checking practices

Output checking and disclosure control of data is usually conducted by staff at the RSU or equivalent. The extent and level of disclosure control checks are usually determined on a project-by-project basis using different organisation-specific policies, national statistical guidance, and/or stipulations of the governance panel and data custodians. They often require limiting the number of row-level table cell counts, individuals, or events, including those that allow for back-calculation of cells. At WADLS in Australia and MCHP in Canada, any publications – such as conference abstracts, posters, presentations, manuscripts, and student theses – may be scrutinised by data custodians on request, with a focus on compliance with ethical approval, disclosure control, data quality, and appropriate acknowledgements. Other RDTs (such as SAIL in Wales) require researchers to include a predetermined acknowledgements statement and to submit all copies of their related publications to the RDT.

5 Advances in technology

Technology is evolving at a rapid pace and many of these technological advances will influence the development of the technical infrastructure needed to support a Data Access, Storage, Sharing and Linkage (DASSL) model.

5.1 Cloud computing and storage

Cloud computing and storage is based on using the Internet to access remote servers. This offers many advantages, including scalable capacity for growing health and related datasets; a reduction in the initial capital expenditure required compared with bespoke, on-premises physical infrastructure; usually built-in resilience to reduce the impact of hardware failures; and centralised system administration and maintenance. Large multinational technology companies, such as Amazon, Google, and Microsoft, have become major competitors in the provision of cloud-based solutions for both public and private sector needs. The solutions on offer can range from public clouds that are delivered via the Internet and shared between organisations (or 'tenants') to private clouds with computing and storage resources dedicated solely to one organisation, and hybrid clouds that combine public and private clouds (65). Depending on the solution, the concepts of a private cloud solution and an on-premises infrastructure can be somewhat blurred; in both cases, resources are dedicated to a single organisation, which either maintains the hardware itself (on-premises) or outsources it to the provider (private cloud).

The security and privacy of cloud-based solutions is often a concern for processing sensitive data and these are dependent on the nature of the provider (66). The multi-tenancy nature of public clouds, where different people and organisations share the same infrastructure over the Internet, raises obvious concerns about keeping sensitive data safe from unintentional disclosure and malware. In contrast, private clouds offer dedicated resources along with additional controls that are much more suited to handling sensitive data. However, the level of control over the physical infrastructure and corresponding trust in the cloud provider are potential issues in choosing between a private cloud solution and on-premises infrastructure, particularly where sensitive data are involved.

Major cloud providers, such as Microsoft, have countered this by establishing and publishing the principles, methods, policies, and certifications that ensure security, privacy, and compliance of their services in safeguarding customer data (67). The Data Protection Commission in Ireland has published recommendations on secure cloud-based environments and on the measures to be applied for GDPR compliance, with additional advice and best practice provided by the European Union Agency for Cybersecurity (ENISA; formerly known as the European Union Agency for Network and Information Security), the National Institute of Standards and Technology (NIST) in the United States, and the European Data Protection Supervisor (EDPS) (66).

The Norwegian Health Analysis Platform has employed the use of a Microsoft Azure open cloud solution because of the benefits, including the pace of development; scalability and performance; reduced investment needs; easier and better management; information security; and functionality. Cloud solutions have also been used by safe havens internationally such as ERICA in Sydney, which is provided by Amazon Web Services (AWS) (48). As most security issues with cloud solutions are reportedly caused by human error and misconfiguration, these can be mitigated by making project spaces configuration and operation 'point and click' only (68).

The Health Data Hub in France has also been using a Microsoft Azure cloud solution (32), but it now plans to move to a solution provided by a European or French company due to concerns regarding data transfers between the EU and the United States (34). Other data hubs continue to store health data using on-premises hardware (48). The Cloud First Policy established by eHealth Ireland ensures that all future procurements in the Irish health service are developed as 'cloud-first' solutions (69). The epilepsy electronic patient record (70) is one such example of a cloud solution used in the Irish health service with data stored on Microsoft Azure.

Provision of cloud-based services has been dominated by large companies in the United States, leading to EU member states (including Ireland) committing to the establishment of a competitive and resilient European cloud (71). The HealthyCloud project, funded by the EU's Horizon 2020, is generating a strategic agenda for the European Health Research and Innovation Cloud, which could become part of European Open Science Cloud (EOSC) and provide the biomedical and health research community with the technical infrastructure and services necessary to support the development of innovative diagnostics methods and medical treatments (72).

5.2 Open-source software

Open-source software (OSS) has become prevalent in computing and its adoption for enterprise and government functions is steadily growing. The 'free' nature of such software can pose a dilemma for many organisations in terms of sustainability (i.e. investment in software that is free but that may come without support, versus investment in proprietary software that typically comes with support contracts) (73). Beyond this, OSS may offer advantages over its proprietary counterparts, such as the following:

- The transparency of open-source codes facilitates better trust in how the software functions, particularly in relation to security.
- For community codes (i.e. OSS developed by a community of software developers, often in a particular domain), there is more scrutiny of the code and any bugs discovered can be resolved relatively quickly.
- OSS gravitates to open standards where possible; hence, interoperability should improve.
- OSS can help avoid vendor lock-in where organisations become over-dependent on a particular proprietary solution.

Conversely, OSS has some obvious shortcomings that have led to stakeholder concern and preference for proprietary software. An organisation may require additional in-house expertise to maintain OSS, rather than outsourcing such support to third parties. OSS may also have reduced features, lack proven security standards, or have inferior user interfaces compared with its proprietary counterparts. However, many large-scale healthcare platforms have adopted OSS, such as the COVID Tracker app in Ireland; OpenPrescribing in the UK, which provides a search interface for the raw prescribing data files published by the NHS (74); and openEHR, which provides an open standard specification for the management, storage, retrieval, and exchange of health data in EHRs (75). Although many RDTs internationally have used commercial products from companies such as Microsoft, IBM, and Amazon, OSS has also been developed for data linkage, storage, and analysis (19, 53, 76).

5.3 Federated data services

Different federated models exist and remain under development, which often require the data to remain at the source (77). The proposed European Health Data Space (EHDS) will be a federated infrastructure that supports analysis, with the data likely remaining within the country in which they were collected. Federated services usually provide value if they are based on common standards that ensure transparency and interoperability (78). Gaia-X is another European project that aims to develop federated and secure data infrastructure in order to link cloud service providers and users together (78). Gaia-X aligns network and interconnection providers, cloud solution providers, and HPC, as well as sector-specific clouds and edge systems. There are a number of instances of Gaia-X being used in the healthcare context, such as the COVID-19 Dashboard in Germany and the Personal Health Train in the Netherlands, where the researcher travels to the data source (39).

5.4 Artificial intelligence

AI refers to systems that display intelligent behaviour by analysing their environment and taking action – with some degree of autonomy – in order to achieve specific goals (79). AI can have a huge impact on healthcare by improving hospital workflows, optimising the assignment of human and other resources, enhancing the efficiency and effectiveness of clinical trials, and supporting the discovery of new medicines (80). The potential of AI for health has been further reinforced by the COVID-19 pandemic (80). Machine learning is an application of AI that sees machines accessing and learning from data, while NLP, another subset of AI, is used to identify and translate human language.

AI capabilities are advancing rapidly, and while huge benefits are expected from AI – both for healthcare and for other sectors – it also raises a number of ethical and data protection concerns (81). In 2018, the European Commission set up the independent High-Level Expert Group on AI, which subsequently published ethics guidelines to promote trustworthy AI that is lawful, ethical, and robust (82). The *White Paper on Artificial Intelligence* followed in 2020 to demonstrate how the European Commission would support and promote the development and uptake of AI (83) and this coincided with the publication of the *European strategy for data* (84) and *Shaping Europe's digital future* (85).

Following public and government consultations on the *White Paper on Artificial Intelligence* (86), the recent proposal for the European Commission's risk-based Artificial Intelligence Act (87) and a coordinated plan on AI for EU member states (80) were also published. Key components of promoting and supporting AI include improving the availability of high-quality data (which are diverse and non-discriminatory and can be reused and combined) and developing HPC capabilities (80). The EHDS will support the training and testing of AI algorithms. Testing and experimentation facilities in health for AI and robotics technologies are scheduled to be set up by the end of 2022 through the Digital Europe programme (88). In 2021, Ireland's National AI Strategy was published in which it outlines the strategy for building public trust in AI, leveraging AI for economic and societal benefit, and enabling AI via education and infrastructure (89). The National AI Strategy highlights the use of AI for health but recognises some legal gaps and the need for impact assessments, codes of practice, and ethical guidelines.

RDTs have used AI in a number of processes, including data pre-processing, cleaning, linking, and anonymisation/de-identification, as well as statistical disclosure control. This has included streamlining the manual review of record linkage results, which is a resource-intensive process traditionally requiring visual inspection and expert domain knowledge (90). NLP has also been used to abstract relevant information from free text and to improve the consistency and standardisation of terms for research purposes. For researchers who are using machine learning to perform analyses and training algorithms for clinical use, access to large national health datasets can be enabled via safe havens, such as the HAIDAP in Canada (59). However, members of the public have had concerns regarding the use of AI, and their support is often conditional on transparency in terms of how data are used (91).

5.5 Encryption

Data encryption encodes or scrambles messages or files so that they can only be read by the intended party with the necessary key. Encryption methods have been used for many years for data security when transmitting or moving data, and these methods are advancing (e.g. homomorphic encryption; polymorphic encryption and pseudonymisation) (48). Additionally, data encryption is now enabling PPRL, where only encrypted data leave the data controller (92). PPRL methods have been successfully employed in some scenarios (77, 93) and are useful when data controllers are reluctant or unable to release personally identifiable data. The data from each of the data controllers are encrypted and then matched or linked by a TTP or two individual entities (94). However, data inaccuracies (e.g. misspellings, missing data) create challenges to linking the encrypted data.

Therefore, Schnell *et al.*, (2009) (95) proposed using Bloom filters, which separate the string variable into bigrams (e.g. two adjacent letters or numbers) and process them separately. The Centre for Data Linkage (CDL) at Curtin University in Australia has evaluated these methods using its open-source linkage software LinXmart across RDTs in Australia (i.e. the CDL, CheReL, and WADLS) and Canada (PopData), as well as in some real-world projects (76). While PPRL compared favourably with clear text linkage when a unique identifier existed across the datasets (74), the linkage quality depends on the data controllers correctly encrypting the data, and on the quality of the original data (92, 96). Manual review of linked records, and records that are similar but should not be linked, are a key component in the data linkage process and is used to assess the data linkage quality. However, manual review cannot be conducted using encrypted data, i.e. without disclosure of some personally identifiable data (92, 96). Therefore, clear text data remains to be the preferred input for record linkage internationally.

5.6 Blockchain

Blockchain is a decentralised or distributed ledger technology that allows for the storage of data that are permanent and immune to fraud, without the need for a central or trusted authority (97, 98). Different blockchains exist (e.g. Ethereum, Hyperledger Fabric), which may be private or public, and open source or proprietary (98, 99). While the most well-known application of this technology is in the domain of cryptocurrencies (e.g. Bitcoin), the technology can be generally applied to other systems where transactional records of any kind can be updated and kept safely in a distributed manner (e.g. keeping a record of micro-contracts). In the health sector, blockchain is being discussed for consent management (100).

With the exception of Estonia, where every citizen's health record is secured with blockchain technology (99, 101), there have been minimal real-word implementations of blockchain in healthcare (97, 98). Other uses or potential uses of blockchain in healthcare have included data sharing between Electronic Medical Records (EMRs), access control, auditability, distributed computing, data storage, and data aggregation (98). Additionally, a number of projects are currently investigating the use of blockchain and smart contracts to allow individuals to dynamically manage their consent preferences; examples include the EU Horizon 2020 My Health My Data project (102), the Dwarna web portal in Malta (99), and a prototype consent management solution being developed at Monash University in Australia (103).

The advantage of blockchain is that it enables greater openness, transparency, and trust (97, 98, 101), as data are linked, time-stamped, and validated across the network and cannot be deleted (98). This may encourage and facilitate data sharing across organisations and allow data subjects more control over the use of, and access to, their health records (99, 101, 104). However, it is also reportedly a potential conflict with the GDPR, as data subjects have the right to be forgotten (99). Other hurdles faced by this nascent technology include its scalability (98, 99), cost, performance speed, high energy consumption (97, 98), and exposure to security flaws (97). However, costs can reportedly be offset by the associated benefits (101), and some types of blockchains, such as those based on the Hyperledger Fabric, offers higher transaction rates, lower network latency, and lower energy demands than conventional blockchains (104). Additionally, sensitive data can be stored off the main blockchain (97, 99), and private blockchains can be used (99) to address security and privacy concerns. It is unlikely that blockchain will have major relevance to the DASSL model, at least initially. The most likely use of blockchain will be in fine-grained consent management.

5.7 Synthetic data generation

The collection of health data requires time, money, and effort, but access to these valuable data can be restrictive, particularly for researchers who are not involved in collecting the data. One way to augment the value of such data would be to generate a synthetic dataset based on the original that does not contain any sensitive personal information. This synthetic dataset should ideally be shown to have similar characteristics or patterns as the original. The DASSL proof of concept (PoC) technical infrastructure involved the generation of synthetic datasets in order to simulate linkage of such data for a number of case studies.

There are various approaches to generating synthetic data based on real datasets. These range from simple perturbations implemented on original data to more recent machine learning methods. In order to quantify the degree to which a dataset may be anonymised from the actual data, the concept of *k*-anonymity has been proposed (105). Achieving *k*-anonymity means that each record contained in a synthetic dataset cannot be distinguished from at least *k*-1 other individuals in the original dataset. One of the easiest and most obvious ways to achieve this type of anonymity is via generalisation, i.e. transforming values of potentially identifiable information to more general values in order to make several people indistinguishable from one another. This idea can be pushed to the extreme where all records are generalised, which will result in synthetic datasets that reflects only high-level patterns observed in the real datasets.

This approach to anonymisation can distort the records significantly, and over-generalisation means that any patterns from the original dataset may be lost in subsequent analysis; in other words, the synthetic dataset loses value. It is important to note, however, that all anonymisation methods necessitate generalisation to some degree.

In order to reduce over-generalisation and to retain characteristics of the original data, various methods have been proposed (106) that involve the use of semantic graphics, decision trees, fuzzy regression models, Gibbs sampling, support vector machines. In many cases, the methods try to retain the fundamental features of the real data by replacing the values in each record with those generated by computer simulations (e.g. from probability distributions). While maintaining the characteristics for individual fields is achievable (e.g. a normal distribution for age from a real dataset), retaining the complex and temporal associations and relationships in a synthetic dataset is usually the most difficult aspect.

In terms of the tools available, one popular and publicly available tool is the synthpop library (107), which is implemented in the R statistical package. The tool was developed to generate synthetic datasets for various longitudinal studies in the UK examining administrative data and observing individuals and their families across several decades. Synthpop works by ingesting the original content data (personally identifiable information will be removed) and then creating a dependence tree of variables and their distributions and other aggregate statistics. It then generates the least dependent variable first (e.g. sex) and, from the resulting data point, it generates the next dependent variable using conditional probability and so on until all variables have been generated for that data row. Thus, the tool preserves the dependence and correlation between data. The software library includes a number of synthesis methods – from parametric (linear/logistic regression) to non-parametric (classification and regression tree) – that can be used to generate synthetic datasets.

Finally, more advanced ways of generating synthetic data are being proposed using some of the latest machine and deep learning methods inspired by AI. Deep generative models, such as variational autoencoder (VAE) and generative adversarial network (GAN), have already proven to be effective in generating synthetic data; the latter model, applied to images, has bee responsible for generating realistic artificial human faces. However, these types of methods are still being introduced and tested by the health informatics community.

6 EU and international developments

The timing of this PoC project aligned well with a number of encouraging European initiatives that require the secure sharing of health and related data. In addition to the upcoming European Health Data Space legislation (discussed further in Section 13.1), this PoC project also aligns with a number of other EU developments, described below.

The sharing of health data worldwide is being encouraged by a number of organisations, including the World Health Organization (WHO) and the Organisation for Economic Co-operation and Development (OECD) (108). The COVID-19 pandemic highlighted the importance of having member state mechanisms in place to provide secure access to health and related data across borders for policy and planning; to inform treatment, care, and public health responses; and to enable research and innovation in the public interest. While the data may not always be able to leave Ireland, federated analytics could be facilitated by a national DASSL service.

6.1 Population Health Information Research Infrastructure

The Population Health Information Research Infrastructure (PHIRI) is laying the foundations for a distributed infrastructure on population health in order to facilitate the sharing of cross-country population health information and the exchange of best practices on the reuse of data (109). The objectives of PHIRI are to provide a Health Information Portal with Findable, Accessible, Interoperable, and Reusable (FAIR) catalogues of health data. This includes the services and tools necessary for researchers to link different data sources, provision of structured exchange between countries on COVID-19 best practices and expertise, promotion of interoperability, and tackling of health information inequalities. The development of a standardised catalogue of health and related datasets is recommended for a national DASSL service, and this could feed into the Health Information Portal developed as part of PHIRI.

A single national DASSL service supporting access to and analysis of our national health data resources would also support the population of this Health Information Portal, as it would provide information on the national 'node', legal and ethical guidelines, research networks, projects underway, and publications.

6.2 European Health Data Space

The EHDS is being developed with the aim of fostering the exchange and sharing of different types of health data, including electronic health records and registries, in a federated manner (110). This will support healthcare delivery, as well as health research and policy-making. The Joint Action Towards the European Health Data Space (TEHDAS) is producing many findings on the sharing and use of health data, including interoperability standards, public engagement, and a recommendation for specific legislation (111). An exemplar of cross-border data sharing is being trialled between Findata and the French Health Data Hub (112). These findings will be extremely important for a national DASSL service in Ireland, as this could operate as the 'node' in Ireland for the EHDS. The Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space will also be of critical importance to a national DASSL service. For example, it will provide guidance for health access bodies regarding the provision of pseudonymised data to researchers.

6.3 Gaia-X

Gaia-X is a European project that aims to develop federated and secure data infrastructure that can link many cloud service providers and users together (78). As the EHDS will be a federated infrastructure, Gaia-X may be of great importance to the EHDS and, therefore, the development of a national DASSL service. The Gaia-X ecosystem consists of three architectural layers that are interconnected: the data ecosystem, the infrastructure ecosystem, and the Federation Services. Gaia-X aligns network and interconnection providers, cloud service providers, and HPC, as well as sector-specific clouds and edge systems. Future-proofing a national DASSL service to interact with the EHDS and other European infrastructures may require consideration of Gaia-X in the future.

6.4 European Open Science Cloud

EOSC is a federated and open environment which supports the publication and sharing of data, tools, and services for research, innovation, and educational purposes, including in the medical and health sciences domain. This environment could be important to a national DASSL service in the context of supporting research and analysis, as well as sharing of data, tools, and services produced as outputs and/or imported into a national DASSL service.

6.5 North–South initiatives

The COVID-19 pandemic further demonstrated the importance of the movement of health and related data across the island of Ireland. Continued development in this area would support and progress the delivery of healthcare and critical healthcare decisions as well as the government's Shared Island initiative. The EU adequacy agreement with the UK for the GDPR could support this movement of data, as well as a national DASSL service collaborating with the Health and Social Care Honest Broker Service in Northern Ireland, but challenges remain as a result of Brexit. Federated analysis (i.e. analysis performed on separated datasets which remain with their respective data controllers), as is being proposed for the EHDS, could support data sharing between Northern Ireland and a national DASSL service in the Republic of Ireland. However, the lack of data flow between the data trusts with federated analysis would restrict the ability to link records of individuals attending healthcare both north and south of the border. This would limit the full potential of data linkage on the island of Ireland for public benefit. Therefore, it is recommended that the governance and potential to share data across the entire island of Ireland is considered.

7 Health and related datasets

Datasets are a collection of related sets of information that are usually gathered for a primary specified purpose, although they are often also used for one or more other relevant purposes (often referred to as secondary purposes).

During the course of the development of this PoC, many different types of datasets and data were identified across Ireland. The primary purpose for which the dataset was originally collected, the type of data collected, the quality of the data, and where the data are stored will impact on how the data can be linked and utilised for secondary purposes within a national DASSL service, and these considerations are discussed in Sections 7.1–7.3.

7.1 Dataset type and purpose

The main types of different health and related datasets in Ireland have been categorised in this report as clinical records, administrative health datasets, patient registers, longitudinal cohorts, operational, and research-specific. Non-tabular datasets were also categorised separately as they would have different governance and technical requirements within a national DASSL service compared with tabular datasets. Each dataset type/purpose comes with its own characteristics which can impact on how it could be shared, stored, linked, and analysed within a national DASSL service.

7.1.1 Clinical records

Healthcare professionals primarily and traditionally collect health or clinical records for individual patient care and management or for audits, but these records hold critical information for use by policy-makers and researchers. With appropriate governance and a lawful basis, these computerised clinical records could be used within a national DASSL service. However, the following considerations around the use of these types of datasets should be taken into account:

- Only computerised clinical records (such as the electronic patient record (EPR) at St James's Hospital), as well as those for specific conditions or cohorts (e.g. the epilepsy EPR, Maternity and Newborn Clinical Management System (MN-CMS)) and GP clinics, could be used.
- The EPR does not currently cover the entire population (e.g. at present, not all epilepsy patients and newborns/pregnant women are recorded on the epilepsy EPR or the MN-CMS, respectively).
- There is a relatively limited time period covered, as many patient record systems have only recently been computerised, with some older legacy records remaining on paper.
- Free text is popular within clinical records, but can be challenging for computers to interpret and analyse and can be more difficult to anonymise (but NLP could assist with this).
- Structured data fields and coding standards are more readily usable elements within clinical records, if and when they are consistently completed by healthcare providers.
- Several coding standards are used across different clinical records, which can create a challenge for harmonising the data unless standards are mapped to one another.
- The deployment of different systems, vendors, and data dictionaries/formats across hospitals and primary care settings creates challenges for aggregating and correctly interpreting data.

7.1.2 Administrative health datasets

Administrative datasets are those which collect data primarily for reimbursement and operations of the health service, such as the Hospital In-Patient Enquiry (HIPE) and the Primary Care Reimbursement Service (PCRS). The traditional reimbursement purpose of these datasets has resulted in some limitations and advantages which should be considered in using them for linked data research:

- Many have been collected for long periods of time at a national level.
- Many only cover public patients or hospitals.
- Some (e.g. the national HIPE dataset) do not require, and therefore do not collect, personally identifiable information that would allow record linkage.

7.1.3 Patient registers

Patient registers collect information on a group of people based on their particular disease, condition, exposure, or health-related service, such as the National Cancer Registry Ireland (NCRI), the Cystic Fibrosis Registry of Ireland, or some National Office of Clinical Audit (NOCA) datasets. These registers may be collected nationally and also connected with international registers (e.g. the Irish Epilepsy and Pregnancy Register). Some characteristics of patient registers which differ from other datasets should be considered if using such registers in a record linkage study:

- Some (e.g. the Cystic Fibrosis Registry of Ireland) require informed consent and may not include every person with the disease, whereas the NCRI, for example, has a lawful basis to collect data on all individuals with cancer.
- Some are for a specific region of Ireland only or may not be maintained and up to date due to a lack of resources.
- Not all patient registers collect personally identifiable information at present, such as many of the NOCA audits, and thus do not support record linkage.

7.1.4 Longitudinal cohorts

Longitudinal cohorts are collected in waves from the same purposive sample of a population. In Ireland, there are two well-known longitudinal cohort datasets: The Irish Longitudinal Study on Ageing (TILDA) dataset and the Growing Up in Ireland (GUI) dataset. Similar datasets in the UK are regularly linked with clinical records and other health datasets in order to validate the data collected and expand on the data available from reliable sources. Due to the nature of these datasets, the following should be considered when linking these datasets:

- Personally identifiable information collected to allow follow-up of a population can be used for record linkage.
- A purposive sample is representative of the population of interest.
- Several waves (or years) of data collection are available on each individual, creating very large and detailed datasets.

7.1.5 Health surveys

The Central Statistics Office (CSO) and Health Service Executive (HSE), among other organisations, regularly collect surveys on the health of the population (e.g. the Irish Health Survey, the National Disability Survey, the Healthy Ireland Survey). Similar to longitudinal datasets, these are usually consented datasets, but they differ from longitudinal datasets in that the same people are not necessarily followed up on, and the participants may change in each edition of the survey. These datasets could be linked with other datasets, but some considerations should be taken into account:

- Anonymous surveys would not support record linkage.
- Only a purposive sample of the population would be available.

7.1.6 Specifically collected health research studies

With the progress towards FAIR data and Open Science, more research datasets are being stored in repositories for reuse, including the Irish Social Science Data Archive (ISSDA) (113) and EOSC. Additionally, researchers often want to link their own dataset with one or more of the routinely collected datasets discussed above. If the researcher is linking their own research data, they likely have access to the individuals' personal information. Therefore, aspects for consideration in the linkage of this type of dataset include:

- The population would usually be volunteer participants, which could be challenging to link with a dataset with only a purposive sample.
- Informed consent for individuals is often feasible when it is the researcher's own dataset, so the governance and approvals process may differ slightly.
- The researcher also becomes the data provider and would need to share their data with the national DASSL service.

7.1.7 Operational health datasets

For this report, operational datasets refer to those which do not capture personal information but rather staffing levels, waiting list numbers, number of notifiable incidents, etc. These include data held by the HSE and the National Treatment Purchase Fund. These datasets can be linked to patient-level data using location to provide important service information. Considerations for the use of operational datasets in a national DASSL service include the following:

- When used in isolation, operational datasets are anonymous and often open and available online; however, if they are linked to individual-level data, they must be managed appropriately (e.g. a hospital name may need to be replaced with a pseudonym and appropriately anonymised in order to ensure that a researcher could not reidentify the individual using available information).
- Location could also be considered an identifying factor and should be pseudonymised.

7.1.8 Health-related social datasets

Many other datasets are also crucial to health research and policy-making, including those related to socioeconomic status, education, criminal justice, and housing. In Ireland, many of these important datasets are held by the CSO and other government departments. Linking these datasets to other health datasets provides important insights into social inequalities and areas for improvement. However, important considerations for linking of these health-related datasets must be taken into account:

- Linkage of health data with housing data, employment data, criminal record data, etc. may create additional data protection concerns.
- Personal identifiers collected may differ across datasets (e.g. the individual health identifier (IHI) versus the Personal Public Service Number (PPSN)).

7.1.9 Imaging

Diagnostic and medical imaging are critical data for health research as well as for clinical use. They differ from the typical tabular data discussed above (Section 7.1.1 to 7.1.8) in terms of the types of analysis which can be run and the requirements for storage. The National Integrated Medical Imaging System (NIMIS) has made great progress in the centralisation of these images in Ireland. However, the use and linkage of images from NIMIS or other imagery data providers requires some additional considerations compared with the more traditional tabular data:

- Personal identifiers would need to be shared with a national DASSL service in a comparable format to other tabular datasets (e.g. comma-separated values (CSV), Microsoft Excel) and removed from the images themselves by data providers.

- Additional storage and analysis requirements for these data would require more computing power.
- Use of different machines to capture the images can impact on analysis and interpretation.

7.1.10 Genomics

While genomics are being used for clinical and research applications in Ireland, there is currently no national policy or biobank (114), and to date, Ireland has been merely an observer of large-scale genomics sequencing initiatives such as the European 1+ Million Genomes initiative (115). There is huge potential for genomics to be used for research and public benefit using a national DASSL service, but certain considerations must be taken into account:

- Use and linkage of genomics data may create additional ethical and data protection concerns.
- In some cases, genomics data may be considered too sensitive to leave the data source.

7.2 Data quality, utility, and fit for purpose

Metrics used for data quality often differ based on the purpose for the collection of the data and the specific research question being asked of the data. The Health Information and Quality Authority (HIQA) has developed a framework to evaluate data quality (116) and has carried out a number of evaluations on major national health datasets, e.g. HIPE, PCRS, the Computerised Infectious Disease Reporting (CIDR) dataset (117). These have revealed a general lack of coordination between data custodians, and a lack of robust governance arrangements to ensure quality of data and effective use of information.

Across national data collections there remains considerable challenges such as a large variation in data quality, duplication of data, accessibility problems, lack of completeness and sub-optimal use or sharing of information. This landscape has led to considerable negative impacts, such as the lack of a system to identify vulnerable cohorts of patients in the roll-out of the COVID-19 vaccination programme. There have been calls for a reform of the national health information system, including national strategic leadership to establish a clear, coordinated approach to collect data that is 'fit for purpose' which benefits the health and social care system (117).

Information on quality and utility is essential for linked datasets for research purposes within a national DASSL service. Analysis of health datasets requires knowledge of the completeness, validity, and reliability of data collection; data coverage; and consideration of potential bias. Furthermore, the utility of a national DASSL service will depend on a number of critically important health datasets from primary (e.g. data from GPs) and secondary care (e.g. HIPE), which are often not readily linkable to other datasets. How these datasets are collected and shared will need to be considered in order for them to be amenable for record linkage and this would greatly enhance the value of these datasets for health and social care and for research purposes. Finally, it is recommended that researchers using a national DASSL service be provided with information on the quality of data provided to them, covering the areas outlined in Sections 7.2.1–7.2.3.

7.2.1 Completeness and accuracy

One aspect of data quality is the level of completeness of the dataset. If data fields are left empty and not completed, this will impact on the accuracy of any findings from analysing these data. Where a data field is not consistently completed, the percentage of individuals for whom the data field is completed is important to the researcher. However, it should be noted that not every data field is critical to all research questions. Therefore, the level of completeness of each data field, as opposed to the overall dataset, is important information to be included in the metadata.

Additionally, the validity and reliability of the data entry is critical to deriving accurate findings from the data. Health data fields could be entered by hundreds, if not thousands, of different individuals in some cases. Therefore, clear data dictionaries, education, and training to assist those inputting data will help to improve data quality. Measures employed to improve data quality should also be shared with the researcher.

7.2.2 Population and time coverage

Population and time coverage of the dataset will impact on the questions that can be asked of the data and how the data should be linked with other datasets and later interpreted. The researcher needs to be aware of the length of time the data are available for. The data fields and the coding systems used within the dataset may also change over time, and this information should be shared with researchers within the metadata in order to allow them to determine if the information they require in order to answer their research question is available. Furthermore, coding systems and data fields change and are updated over time.

While older versions of coding standards are sometimes mapped to each other, this is not always possible and creates many challenges to the aggregation of data using different formats. Additionally, if the population is only a purposive sample or from a specific region or healthcare organisation, this needs to be acknowledged in the interpretation of the findings.

7.2.3 Bias and other ethical factors

While routinely collected data usually provide a more comprehensive overview of a population compared with data collected for a research study, bias can also exist within routinely collected data, especially when linked with other datasets, and this potential bias should be considered (92). Bias can enter the dataset from the initial healthcare interaction or at a later stage during the coding of the clinical interaction, during the linkage or analysis of the data, or at all of these points in time. This bias may relate to gender identity, socioeconomic status, age, religion, or race/culture/ethnicity, among other factors. If a specific cohort of individuals does not appear in a linked dataset created by a national DASSL service, this should be explored in order to determine whether this is because people with these characteristics do not require that type of healthcare or do not present to the healthcare service, or whether the data recorded in relation to them are of low quality (e.g. there is no PPSN available and hence lower quality linkages conducted using names and addresses) and therefore could not be linked to other datasets. Issues created by biased data can be magnified when machine learning is applied to the data to create a model for use in policy and clinical decision-making. Therefore, linked data should always be assessed for potential bias, with this information provided in the linkage quality report to the researchers or other users who do not have access to the unlinked data and personal identifiers.

7.3 Metadata and data formats

Metadata describe the data collected within a dataset. Availability of metadata is important to encourage and support the use of health data for research in the public interest and ensure correct interpretation and understanding of the specific dataset. Metadata are usually described within a data catalogue which contains information on the available datasets, including data dictionaries and standards. Data catalogues, dictionaries, and standards identified during this PoC project are discussed in Sections 7.3.1–7.3.4.

7.3.1 Data catalogues

Several health and related data catalogues were identified in Ireland; these were developed by the government, HIQA, and the CSO (118-120). However, a single comprehensive, up-to-date, and maintained list of all of the datasets and searchable metadata which can be accessed via a national DASSL service is recommended. This would provide all stakeholders with insight into the available data as well as the quality and standards employed, helping to reduce duplication in data collection and supporting some European projects.

7.3.2 Metadata standards and common data models

Metadata standards establish a common method for displaying health information, understanding data semantics, and the correct and proper interpretation and use of data by humans and computers. Some metadata standards – such as the Data Catalogue Vocabulary (DCAT) (121) and the Data Documentation Initiative (DDI) Alliance (122) – are used by data catalogues to increase the discoverability of datasets and allow federated searches of datasets across catalogues in multiple sites. The DCAT has been a popular choice in Europe (123), an example of which can be seen in the Register Utiliser Tool (RUT) in Sweden (36). It is recommended that a future data catalogue for a national DASSL service considers the application of a metadata standard, and data controllers would need to be educated and supported to convert their metadata into this standard.

Other metadata standards support a common method for transmitting, storing, retrieving, and displaying health information, including Digital Imaging and Communications in Medicine (DICOM) for medical imaging, which is applied within NIMIS in Ireland. Another metadata standard used for text data is the Fast Healthcare Interoperability Resources (FHIR) published by Health Level Seven (HL7) (124), which HIQA highlighted in the report *Guidance on Messaging Standards for Ireland* (125). This standard is becoming increasingly popular across Europe and, similar to the DCAT, it has been highlighted by TEHDAS as a potential standard for the EHDS, as it supports interoperability of clinical systems both nationally and internationally (123). Additionally, the recent Goldacre report in the UK (2022) has recommended the use of FHIR and HL7 to support research and analysis of health data (23). Similarly, the Goldacre report and the TEHDAS findings highlighted the Observational Medical Outcomes Partnership (OMOP) common data model to support secondary use of data (23, 123). This standard has previously been successfully used to support federated analysis of international data for COVID-19 research (126).

In Ireland, the recently published *Dataset Specification Management Process Report: Standardising Data for The Future* also aims to provide a standardised uniform process that facilitates new and existing dataset specification (127). Overall, metadata standards are recommended, and European guidance should be considered in order to align with upcoming initiatives.

7.3.3 Data dictionaries

As part of the metadata, a data dictionary for the specific dataset needs to be provided in order to allow the user to interpret the data correctly. This usually includes the names and descriptions of the data fields and any coding systems used. It was noted during this PoC project that while many data dictionaries are available online, others were only available on request directly from the data controllers. Data controllers should be incentivised and supported to share their data dictionaries publicly via a national data catalogue. Additionally, more consistent application of coding systems (e.g. use of '1' for yes, '2' for no, and '0' for not completed) and standardised terminologies would facilitate the integration and combination of these datasets. The HSE's National Health and Social Care Data Dictionary (NHSCDD) may support this, as it aims to provide a list of key health service terms and concepts, including agreed definitions and protocols for inclusion in any new projects or applications being introduced into the HSE (128). This includes the creation of the datasets, alignment of existing datasets, and assisting vendors through the development of evolving dataset specifications, thus promoting more consistency. However, individual datasets will likely still require their own data dictionaries where nuances exist within the dataset which do not appear in the NHSCDD.

7.3.4 Standardised terminologies and coding standards

Standardised terminologies, vocabularies, and coding standards provide a common understanding of clinical terms, and their use is usually reported in data dictionaries This supports consistent collection of data, and thus the overall quality of the data; consistent interpretation of data; interoperability; and the combination of datasets. In the identified datasets, the International Classification of Diseases (ICD) was the most commonly used coding standard, but other standardised terminologies were also employed in clinical records, including the International Classification of Primary Care (ICPC), the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), and Logical Observation Identifiers Names and Codes (LOINC). New versions of these standards are released as updates become available, and some datasets have employed several different versions of the ICD. This can create challenges when aggregating data over different time periods. Additionally, different terminologies may need to be mapped across one another; for example, where a condition or medication is recorded using different terminology across datasets. The researcher or user of a national DASSL service may need support to correctly and appropriately map terminologies and coding systems or for the terminologies to be officially mapped to one another.

8 Development of the PoC technical infrastructure

The DASSL PoC infrastructure was developed to include the key components that support the key technical functions of the DASSL model:

- **Data ingress from data providers**
- **A secure environment for record linkage by the data linkage unit (DLU) with access to personal identifiers**
- **A secure environment for curation, preparation, storage, and output checking of pseudonymised and anonymised data by the Research Support Unit (RSU)**
- **A safe haven for researcher analysis of pseudonymised data, and**
- **Data export of checked outputs for the researcher.**

8.1 System roles

A variety of people need to interface with the technical infrastructure, from system administrators (who manage the infrastructure on a day-to-day basis) to research support staff (who assist the approved users/researchers) (Table 1). It is important to define these (anticipated) roles clearly and explicitly in the design of the architecture, in particular with respect to access privileges to the network and to hardware resources (i.e. access to resources is highly restricted by default and only given to those who have a reasonable need for a limited time).

Table 1 DASSL PoC roles within the technical infrastructure	
Role	**Description**
Researcher	Accredited researchers who have completed the training and approvals required in order to access the platform
Data provider	Responds to data requests on behalf of the data controller and sends the relevant data to the DLU (personally identifiable data) and the RSU (content data)
Infrastructure team	Systems staff operating and managing the environment, including servers (e.g. hosting the analysis platforms or virtual machines) and networking resources
RSU	Responsible for preparing datasets, setting up secure analysis platforms (safe havens), and performing statistical disclosure controls in a secure processing environment. The RSU (optionally, with different sub-roles) may also manage centralised, pseudonymised datasets that are consistently updated by data providers.
DLU/trusted third party	Receives personal identifiable data from data providers, conducts record linkage, and sends linkage keys to the RSU for construction of linked datasets, along with information/statistics about the data linkage process
Secure file transfer service	Manages secure transfer of data in and out of the system
Hardware vendor staff	Staff who may require occasional access to install, configure, and maintain the hardware
Software vendor staff	Staff who may require occasional access to install, configure, and maintain software packages

8.2 System architecture and components

For the purpose of this PoC, the technical environment to simulate the DASSL systems and data flows was implemented in a public cloud environment, utilising Amazon Web Services (AWS). This facilitated convenient deployment of system components and virtual machines (VMs) that serve different functions in the environment. The overall architecture of the system environment is visually represented in Figure 4.

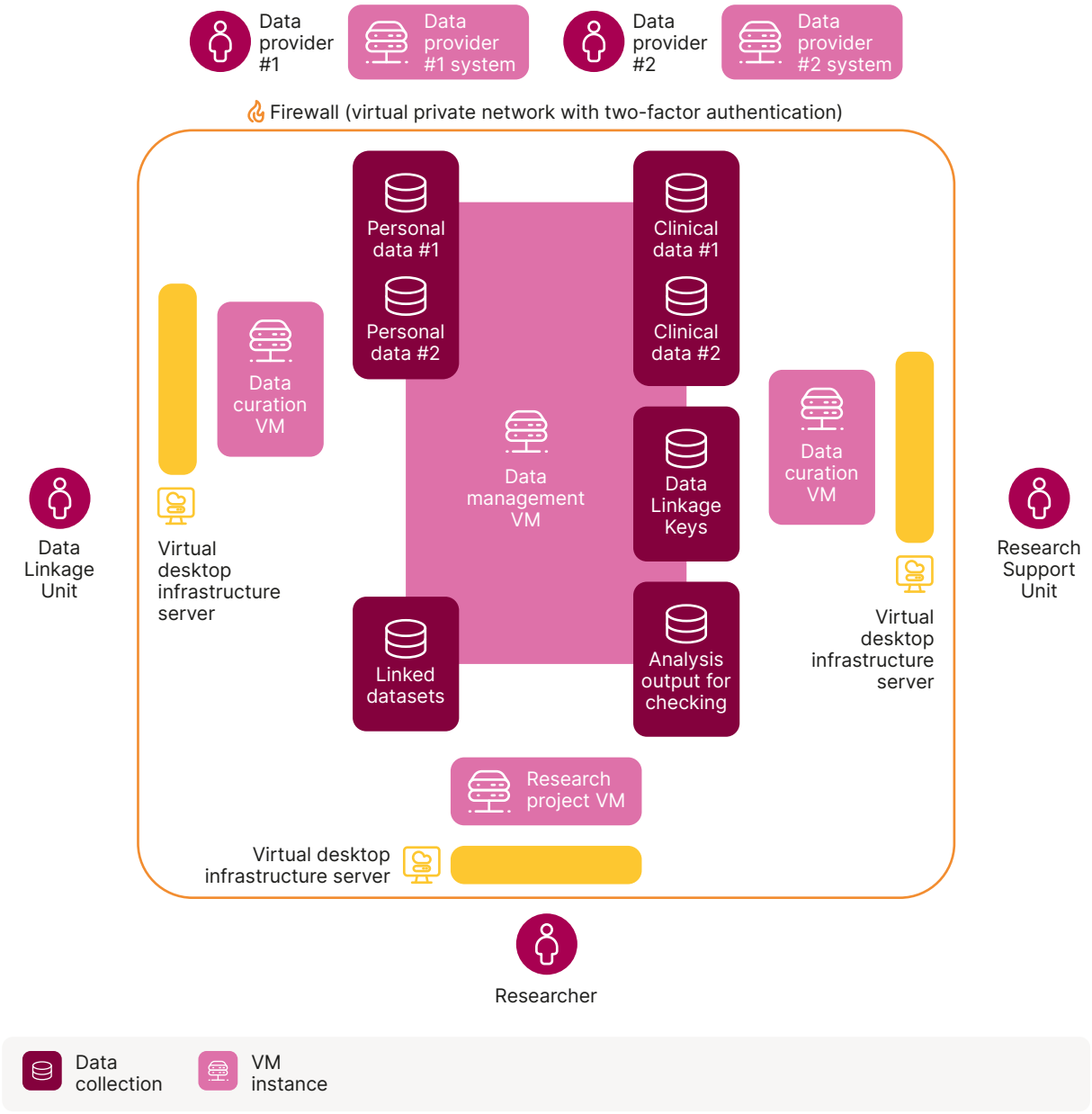


Figure 4 Overview of key system architecture and components for the DASSL technical infrastructure

The system has the following key components:

- A perimeter firewall that protects the entire technical DASSL PoC infrastructure from unauthorised external access. This is complemented by firewall rules on individual VMs, as well as AWS security groups, which only permit access on a strict need-to-use basis.
- Access to the PoC environment via a virtual private network (VPN) is mediated by two-factor authentication (2FA) and by deployment of an identity and access management service, which defines the user roles and access privileges to the different systems. For the PoC, the WireGuard software (129) was used to set up the VPN server/clients, complemented by the Keycloak (130) server to implement 2FA and identity management.
- A virtual desktop infrastructure (VDI) server that acts as the gateway for different user roles to interact with the desktops of the VMs designated for each role. For the PoC, the open-source Apache Guacamole application suite was deployed as the VDI server, which readily supports remote client graphical user interface desktop connections via HyperText Markup Language 5- (HTML5-) enabled browsers without the need for dedicated clients for different operating systems. It also supports different host operating systems (such as Windows and Linux) which is important for the research analysis VMs.
- A data management VM that enables data flows all the way from ingress of data from data providers to the research output produced by the researcher, which must undergo statistical disclosure control by the RSU. For the PoC, the Nextcloud data management platform was used to facilitate data ingress and subsequent sharing between the different roles via data partitions with pre-configured read/write access permissions.
- VMs that are set up for different key roles within the environment:
- Data linkage VM: A system that enables the DLU to carry out data linkage with strict read-only access to personally identifiable data sent by data providers. Once the linkage process is completed, linkage keys are sent to the RSU via a shared partition.
- Data curation VM: A system that allows the RSU to prepare linked datasets using content data from data providers and linkage keys from the DLU via read-only partitions. The VM has write access to another partition on the data hub that provides linked datasets for specific research projects.

- Research project VMs: Individual systems created and configured per project that allow the researcher to carry out analyses on the linked datasets provided by the RSU. The VMs have read-only access to the linked data and write access to a separate partition that hosts the research outputs that are to be checked by the RSU as part of the statistical disclosure control process.

8.3 Data ingestion from data providers

Once a research project has been approved, the first step is the ingestion of the relevant data into the DASSL environment. For illustration purposes, Figure 5 shows datasets from two data providers for a simple case study, but the same principles can be extrapolated for multiple datasets from different providers. For the PoC infrastructure, an open-source data management platform, Nextcloud, was deployed to facilitate the data providers to share data with the RSU and DLU, and for subsequent data sharing between different DASSL units. For each project, the data providers share data via the following process (Figure 5):

- Data providers connect and log in to the Nextcloud platform via a VPN connection and using 2FA.
- Two folders are made available by the RSU and DLU respectively for separate uploads of personal identifiers and content data (each data provider can only see the data they uploaded).
- The file containing personal identifiers is uploaded by the Data Provider to a specific folder (created by the DLU for each project and for each data provider) via a secure web interface with end-to-end encryption; these datasets are subsequently shared only with the DLU.
- As above, the file containing content data is uploaded by the Data Provider to a project- and provider-specific folder created by the RSU; these datasets are subsequently shared only with the RSU.

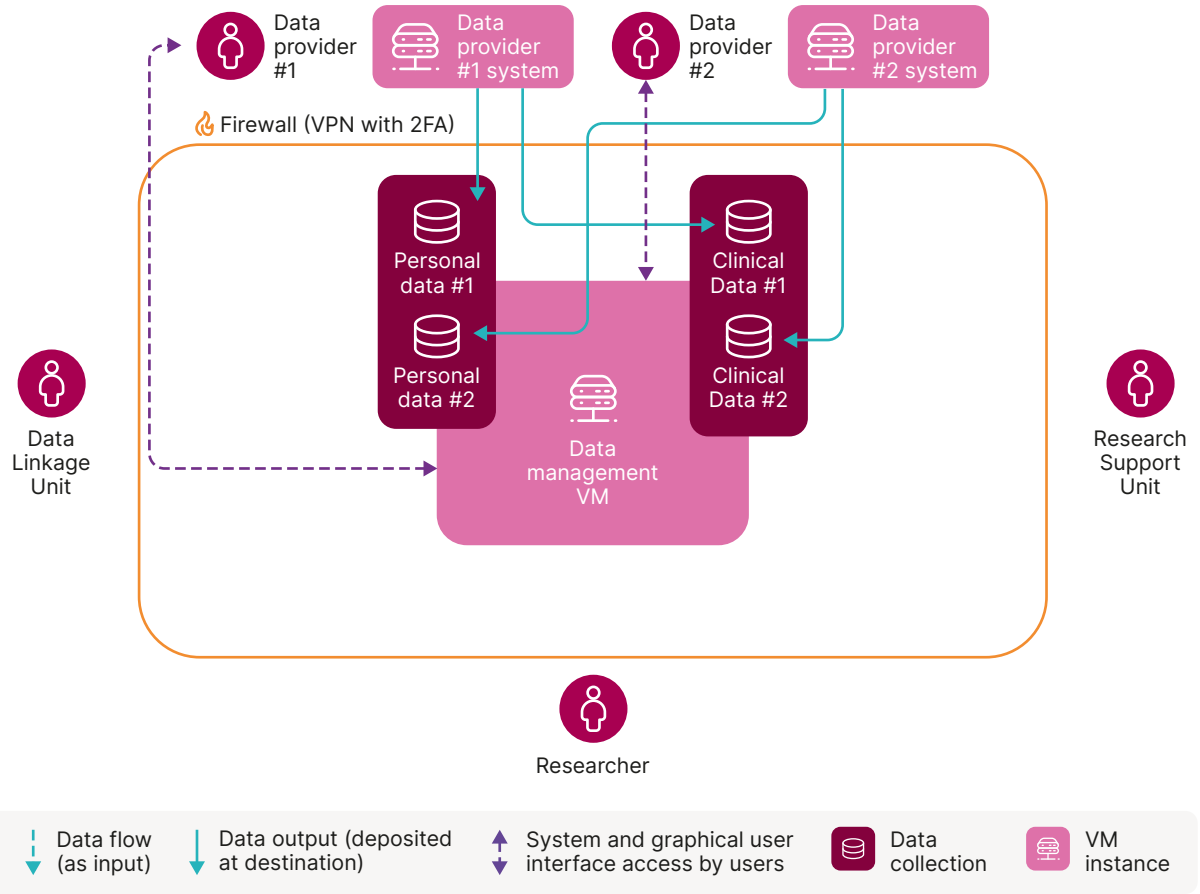


Figure 5 Illustration of the ingestion of data from two data providers into the DASSL technical infrastructure

While the DASSL PoC deployed the same instance of Nextcloud to host both the personal and content datasets for each data provider, access restrictions were put in place in order to ensure that the datasets were only accessible by the DLU and RSU, respectively. Nevertheless, for a production system, the datasets could be held in separate data stores or data management services. In addition, user feedback and international best practices suggest that the process of the data providers sending files securely would benefit from data upload requests sent by the DASSL team for individual projects (e.g. individual URLs where data designated for a project are to be uploaded), upon receipt of which the data provider would authenticate and upload the files. This ensures a much more user-friendly process for the data providers and should reduce errors made in relying on the data provider to correctly place the uploads in an appropriate folder, particularly in instances where data providers upload multiple datasets for different projects. Commercial third parties also provide such file transfer services (along with end-to-end encryption and other security features), which offer further convenience.

8.4 DLU secure processing environment

The DLU has graphical user interface desktop access (via the VDI server) to a dedicated data linkage VM, which has read-only access to the personal data from each project and from the relevant data providers (Figure 6). In addition to accessing personal data, the data linkage VM contains record linkage software; for the purpose of the DASSL PoC, the R package privacy preserving record linkage (PPRL) was deployed (see Section 10.8). Record linkage, particularly probabilistic linkage, can be notoriously computationally intensive, but some methods (such as blocking procedures) can be used in order to reduce the number of linkage comparisons. While the PoC was able to use a modest system (up to 8 cores and 16 gigabytes (GB) of memory) for conducting probabilistic linkage of synthetic data, a production system should have considerably greater resources that should scale up as the size of the datasets increases, e.g. about 100 cores with 256 GB of memory. A SQL server (e.g. MySQL, PostgreSQL) was also installed for the PoC; this is recommended in a production environment for managing personal data, which are relatively limited in the number of fields containing data that should conform to predefined types/formats (e.g. text strings, dates, numerical values). This type of relational database provides better support for data provenance that could also facilitate the establishment of a population spine (see Section 10.6). For the purpose of the DASSL PoC, the DLU shares the linkage keys with the RSU via a dedicated shared folder on the data hub. Only the DLU has write access to this folder, whereas the RSU has read-only access.

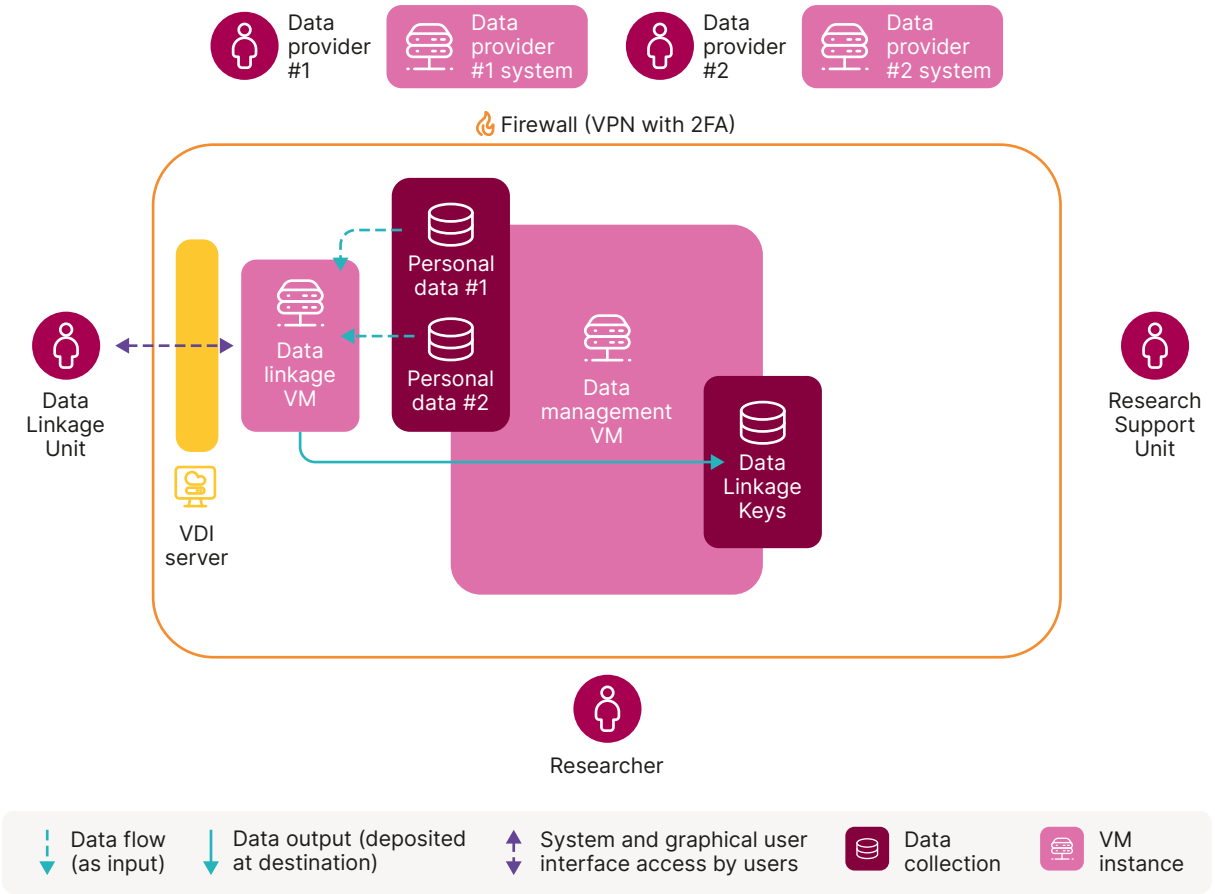


Figure 6 Illustration of DLU accessing VM to perform record linkage and share linkage keys with RSU

8.5 RSU secure processing environment

The RSU has multiple roles within the DASSL environment, but only the routine data preparation task is described here (please refer to Section 8.7 for a description of the RSU's role and requirements in conducting statistical disclosure control). The RSU also has graphical user interface desktop access (via the VDI server) to a dedicated RSU VM, which has the following levels of access to various partitions within the data hub (Figure 7):

- Read-only access to content data provided by the data providers
- Read-only access to project-specific data linkage keys generated by the DLU, and
- Write access to a partition, shared with the researcher, where content data linked by the RSU is provided.

A primary function to be carried out by the RSU is to utilise the linkage keys (sent by the DLU) to combine the content records of individuals (sent by data providers) for a particular project and to further anonymise the dataset where required. The RSU requires data management software applications in order to join and curate the linked datasets for the researcher. For the DASSL PoC, R and RStudio were used, but there should be some flexibility for the RSU to use other tools or frameworks (e.g. Python) to perform this duty, according to the skills and knowledge of the RSU team. For tabular datasets, a relational database system (specifically a SQL server/client set-up) was integrated into the PoC design, and the RSU would benefit from using this database in a production environment to better support and capture data-processing operations and provenance. The RSU then deposits the pseudonymised, linked datasets to a shared partition on Nextcloud that can be read by the project researcher(s). The hardware requirements for the RSU team are relatively modest for relatively straightforward data joining and curation activities – the equivalent of a typical workstation was sufficient for the PoC (quad-core VMs with 16 GB of random-access memory (RAM)). However, this will need to scale according to the number of RSU staff and the number of projects to be supported concurrently.

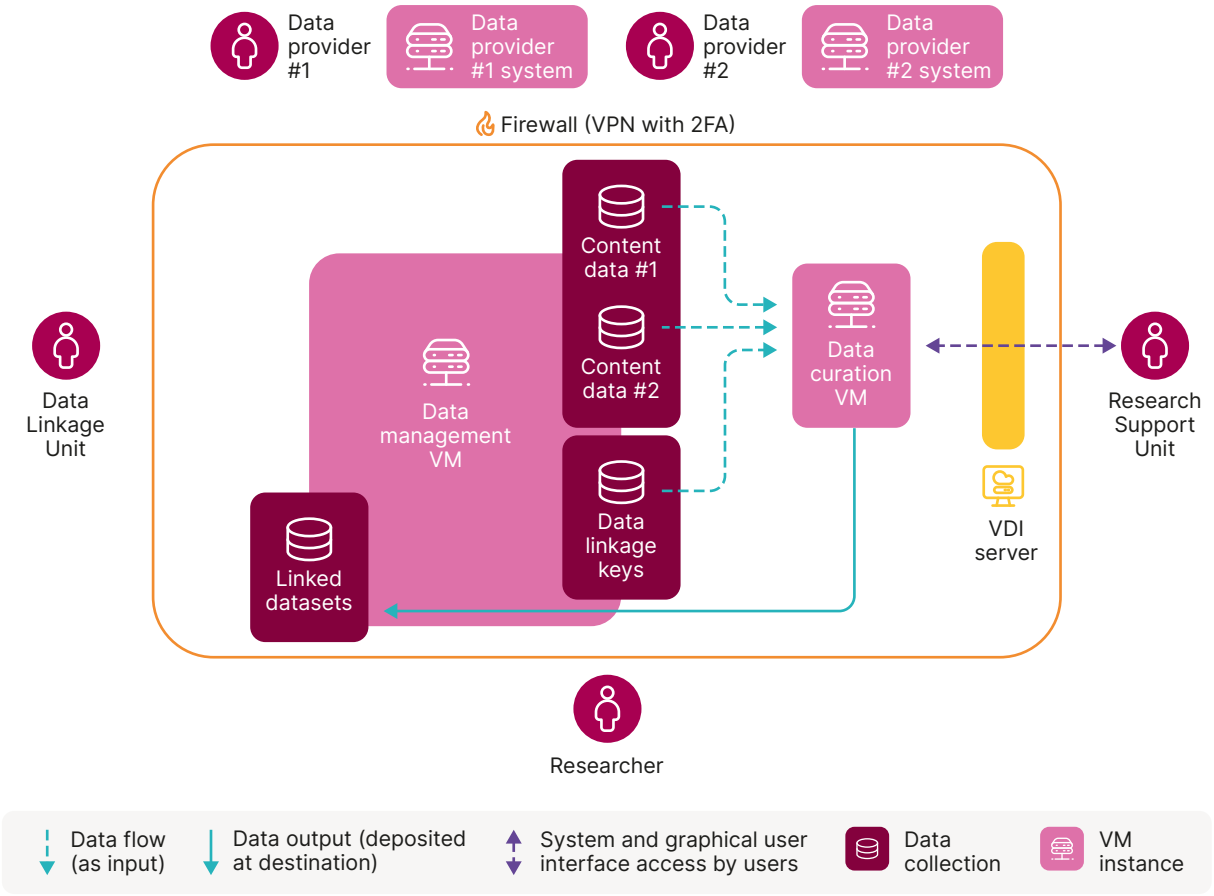


Figure 7 Illustration of curation and combination of datasets, and subsequent sharing with the safe haven

8.6 Researcher secure processing environment (safe haven)

Figure 8 illustrates the next step, where the researcher gains desktop access (with a graphical user interface via the VDI server) to a dedicated research project VM, which has read-only access to the linked datasets curated by the RSU on a project-specific partition on the data management platform.

The hardware and software requirements of the research project VM should be flexible and scalable according to the needs of individual projects and researchers. For the production environment, it is anticipated that these VMs should host different operating systems (Windows and Linux) and should have, at a minimum, routine statistical software applications (e.g. R, SAS, SPSS Statistics, Stata, Excel/ LibreOffice), which may include commercial software packages. Any outputs that the researcher wants to export are written to another partition on Nextcloud.

The research project VM has write-only access to the outputs partition on Nextcloud – all other means for exporting data from this environment are disabled where possible. The VM is prevented from making connections with the external Internet, and the use of copy and paste is disabled so that data cannot be directly copied onto the researcher's local system. These data subsequently undergo statistical disclosure control in order to preserve privacy. This set-up provides the primary interface for the researcher to examine, process, and analyse the linked, pseudonymised data. Once the researcher is connected to the VPN and authenticated via 2FA, only a web browser is required to access the desktop environment of the VM. Figure 9 shows an example of this interface where Case Study 1 (CS#1) data are analysed using RStudio in a Windows environment. This provides an intuitive way for the researcher to gain access to the DASSL infrastructure while security safeguards remain in place (see Section 9.4).

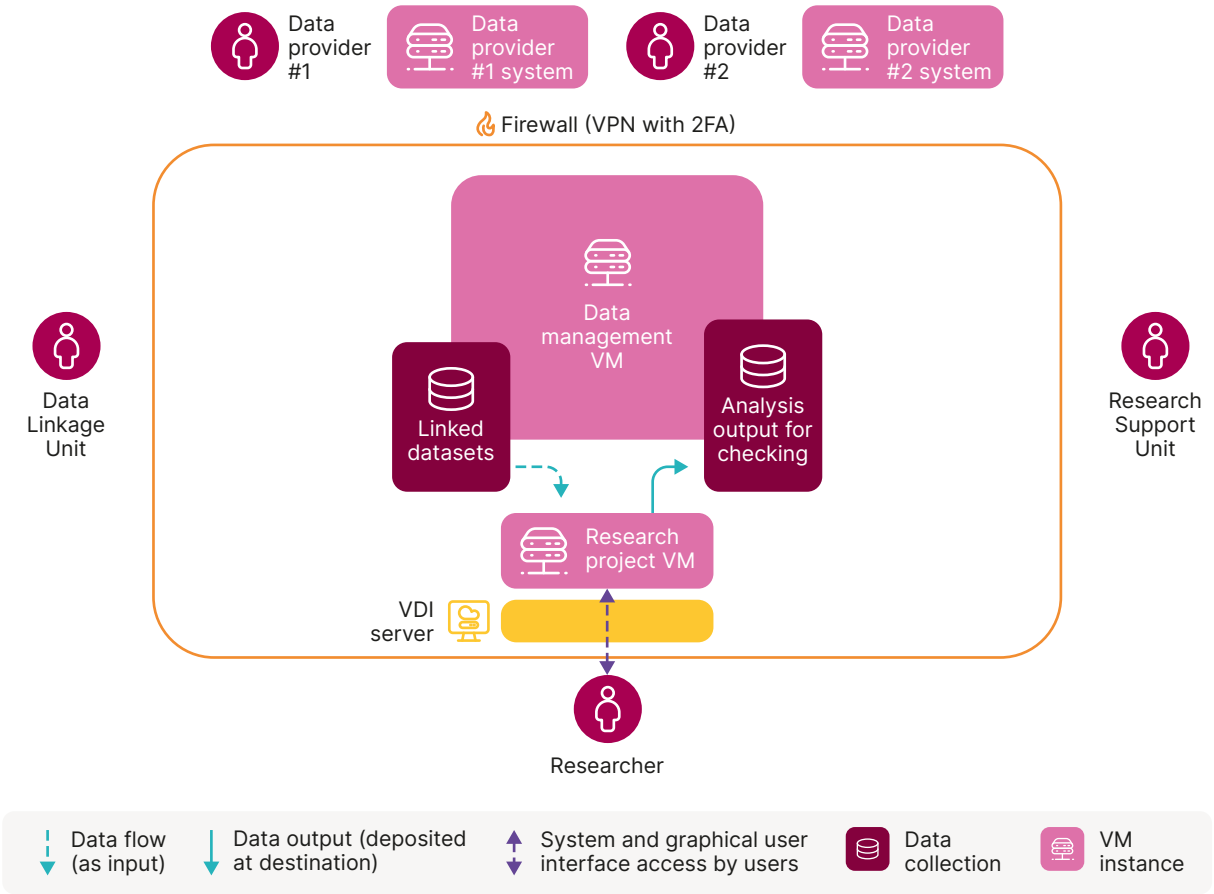


Figure 8 Illustration of researcher access to the research project VM and exportation of analysis output for checking by the RSU

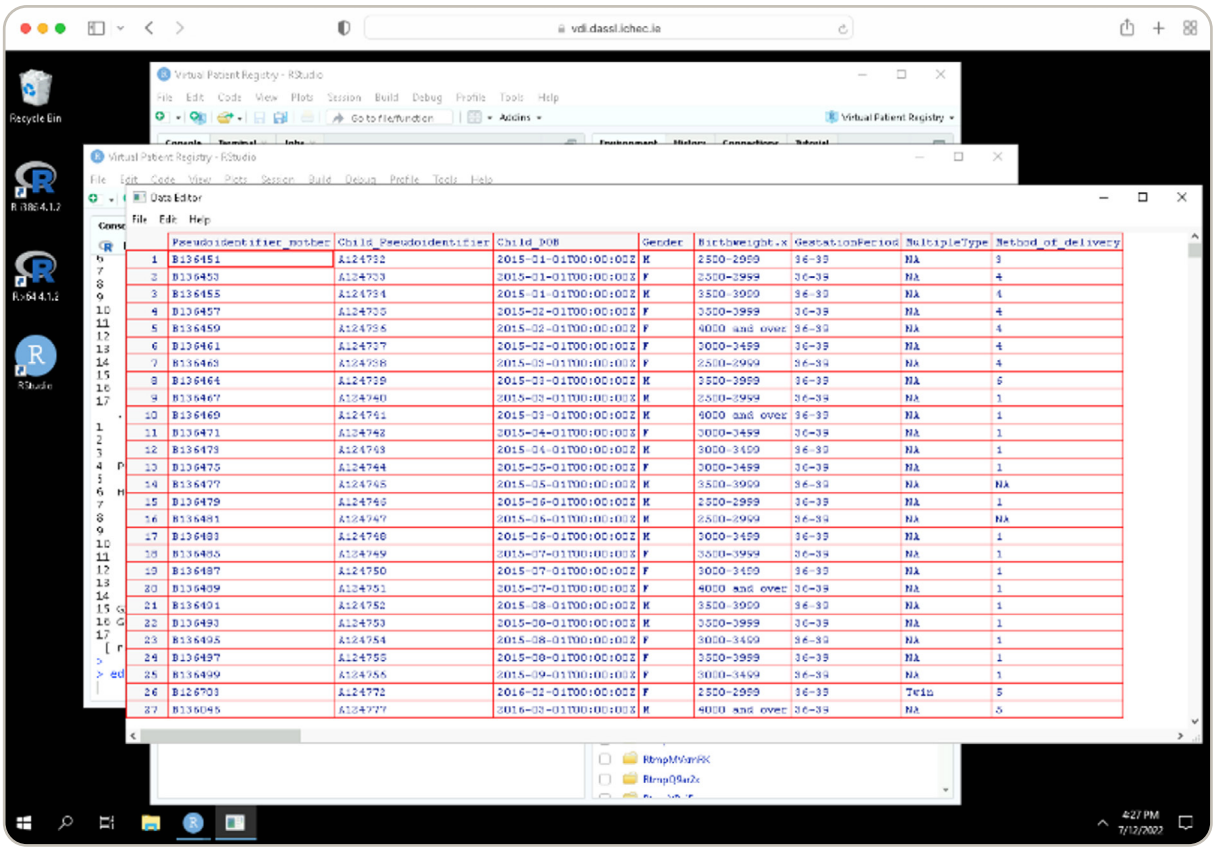


Figure 9 Interface for the researcher to conduct data analysis using RStudio in a Windows environment

8.7 Output exportation

In the final step of a typical DASSL project workflow, any outputs produced by the researcher that are designated to be exported outside the DASSL environment must undergo statistical disclosure control, which is shown in Figure 10.

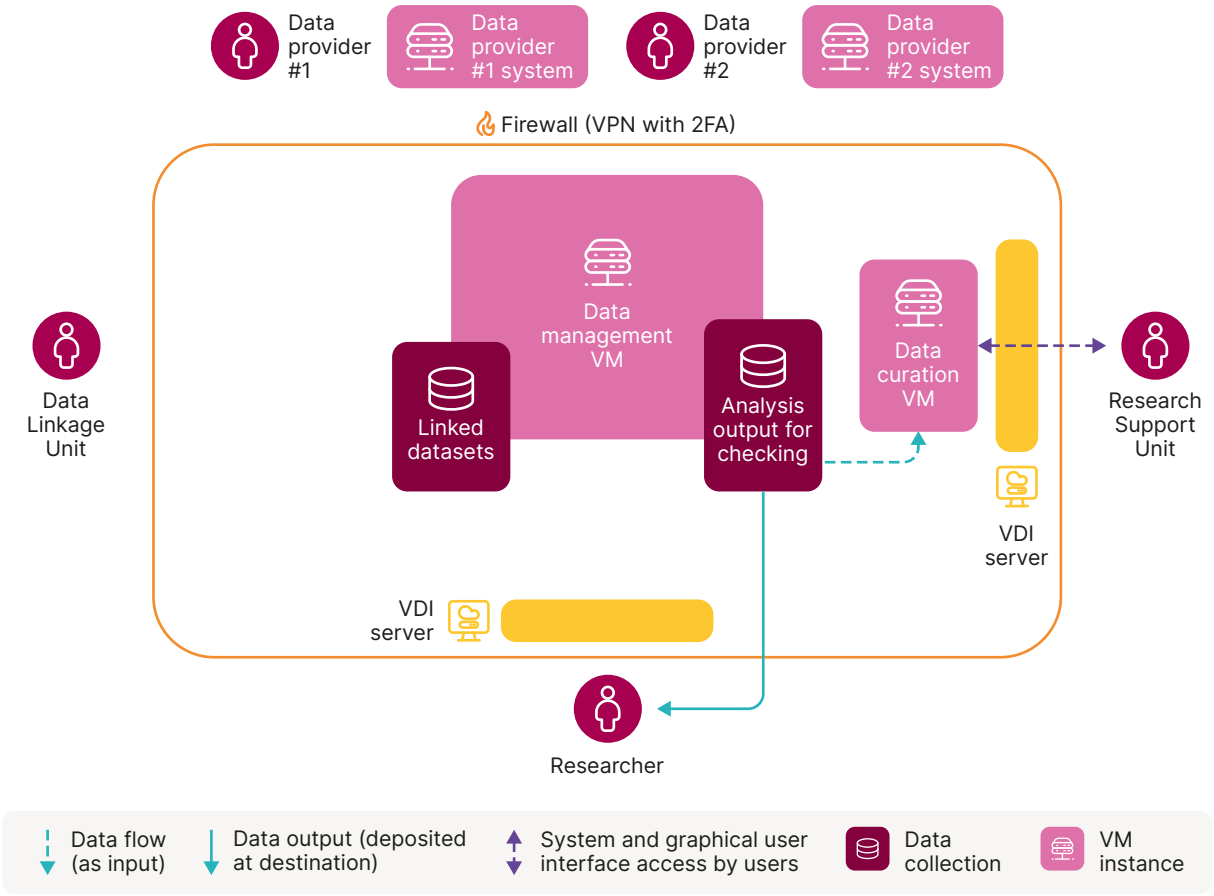


Figure 10 Illustration of the statistical disclosure control process on researcher outputs by the RSU

The researcher writes the output (i.e. research output/findings) onto a dedicated partition on Nextcloud. This is then analysed by the RSU to ensure that disclosure of such outputs poses no risk to the privacy of the data subjects, i.e. that the outputs are anonymised. For the PoC, the R software library sdcMicro (131) was used to perform this step. However, there are also commercial alternatives that could be deployed, such as Tau-Argus (132) and SAS-based implementations used by the CSO, which has carried out an assessment of best practices in this area (133). The implementation of statistical disclosure control using such tools can be compute-intensive, hence the RSU should have considerable computing resources for this type of workload, to the order of about 100 central processing unit (CPU) cores and at least 256 GB of RAM. After statistical disclosure control, the approved output data are exported to the researcher via a designated download URL (generated by the RSU), which provides the only means of exporting data from the DASSL system.

9 Security, data protection, and privacy

The DASSL PoC infrastructure was developed to incorporate privacy by design, with both organisational and technical security measures in place.

The measures introduced as part of the PoC, as well as those recommended for a national DASSL service, are discussed in Sections 9.1–9.6.

9.1 Data Protection Impact Assessment

Under the GDPR, a Data Protection Impact Assessment (DPIA) is mandatory for any new high-risk processing projects in order to identify, and mitigate against, any data protection risks (134). Although the DASSL PoC infrastructure only processed synthetic health data, it was deemed important to complete and update a DPIA throughout the development process in order to promote privacy by design. A DPIA for the PoC was developed with the Data Protection Commission and the Data Protection Officer at the University of Galway. However, a DPIA cannot account for all the risks associated with individual projects using different data sources. Thus, it is expected that new projects processing linked national datasets would require their own DPIAs, which can use the national DASSL service DPIA as a template.

9.2 Five Safes

The Five Safes (Figure 11) is an internationally recognised framework on which RDTs often base their data sharing and linkage models (7). The Five Safes is not mutually exclusive but provides a framework for data management services to make decisions based on five key elements: Safe Data, Safe People, Safe projects, Safe Settings and Safe Outputs. Many of the critical components of the Five Safes are organisational rather than technical in nature and are discussed in Sections 9.3 and 9.4.

Figure 11 Five Safes framework

9.3 Organisational security and data protection measures

Many organisational measures can be put in place to promote the Five Safes. These various measures are discussed in Sections 9.3.1–9.3.5.

9.3.1 Safe data

The separation principle ensures that only the data controller ever has access to both the personal identifiers and corresponding content data. The data controller splits the personal identifiers from the content data and shares the two files with the DLU and RSU, respectively. Additionally, the data minimisation principle is applied during the creation of the researcher data view by the RSU, with aggregation of data variables and removal of variables that are not required to answer the research question. As one researcher may be conducting more than one project via a national DASSL service, pseudo-identifiers used for data subjects should also change for each project.

9.3.2 Safe people

Organisational measures can be put in place in order to ensure that only 'safe persons' can access a national DASSL service. Data sharing agreements must be put in place with any data controller sharing data with the system. All staff at a national DASSL service, as well as researchers, will be trained in information governance and data protection, and will usually sign a confidentiality agreement or declaration of secrecy. Only researchers who have signed the agreement should be granted access to the safe haven, and their organisation and an experienced lead researcher may need to take responsibility for any misuse of data or noncompliance with the agreements. Additionally, division of responsibilities is recommended in order to avoid any conflict of interest (i.e. the DLU, RSU, and Information Governance Review Panels (IGRPs) conduct linkage, content data curation, and approvals, respectively, and should be in separate organisational units, under different line management).

9.3.3 Safe projects

As discussed in Section 13.5, the governance and approvals processes are critical to ensuring safe projects. Clear project approvals processes, as well as appropriate boards with a diversity of representation, will help ensure that all projects undertaken within a national DASSL service are safe.

9.3.4 Safe settings

Security and risk management policies will be critical to ensuring that the DASSL infrastructure is secure and compliant with standards. This includes defining the level of access that each role within the system should have. Similar policies should be in place for physical access to the organisation, with strict policies for visitors/contractors. The organisation operating a national DASSL service should receive and maintain ISO 27001 security management accreditation. This standard is met by many of the international data linkage models as well as by organisations in Ireland that process health data. Regular external audits of the organisation should also be undertaken.

9.3.5 Safe outputs

Any findings that the researcher wants to export from the safe haven will be assessed for statistical disclosure control in order to ensure that the data within the outputs are anonymised to their greatest extent. There are existing policies and guidance on this process produced by the CSO (135) that could be readily adopted as best practice in Ireland.

9.4 Technical security and data protection measures

The technical measures supporting the Five Safes are largely focused on ensuring a safe setting. These measures are discussed in Sections 9.4.1–9.4.6 in relation to access and environments.

9.4.1 Access control

A number of technical measures have been used in the DASSL PoC in order to ensure approved and secure access only. To gain access to the DASSL PoC environment, all users (i.e. the DLU, RSU, researcher, data provider) must use a VPN, which creates a secure encrypted connection between computers over the Internet, providing a private encrypted tunnel for all data and communications within this network. The Internet protocol (IP) address of the device being used to gain access to the VPN acts as an additional layer of security, by limiting VPN access for a whitelist of IP addresses of known devices or networks. Allowing data providers to upload data by whitelisting their IPs, as opposed to simply requiring use of the VPN, was tested for the PoC; however, this was ranked as a less secure option. The whitelist may include, for example, IP addresses associated with university campuses or research institutes whose computers connect to the Internet via fixed IP address ranges; this measure is already in place for the National Safe Haven in Scotland. However, this may cause inconvenience to researchers who may be working remotely and may not have persistent (fixed) IP addresses.

Apart from IP address whitelisting, 2FA provides additional security. 2FA is a two-step verification that requires two different authentication factors to verify users. For the DASSL PoC, each user must provide their username and password as well as an authentication token, which changes every 30 seconds using a 2FA app on their mobile phone (Figure 12). A physical token or text message to a mobile phone could also be used, which may be a less secure option, as a physical token may not be password-protected, and a text message could potentially be intercepted.

Finally, secure access is facilitated by the use of an identity and access management service to ensure that each user role (i.e. the DLU, RSU, researcher, data provider) only has access to the appropriate system and corresponding data (e.g. the RSU can only gain access to its own VM, with access to content data and linkage keys but not to any personal data).

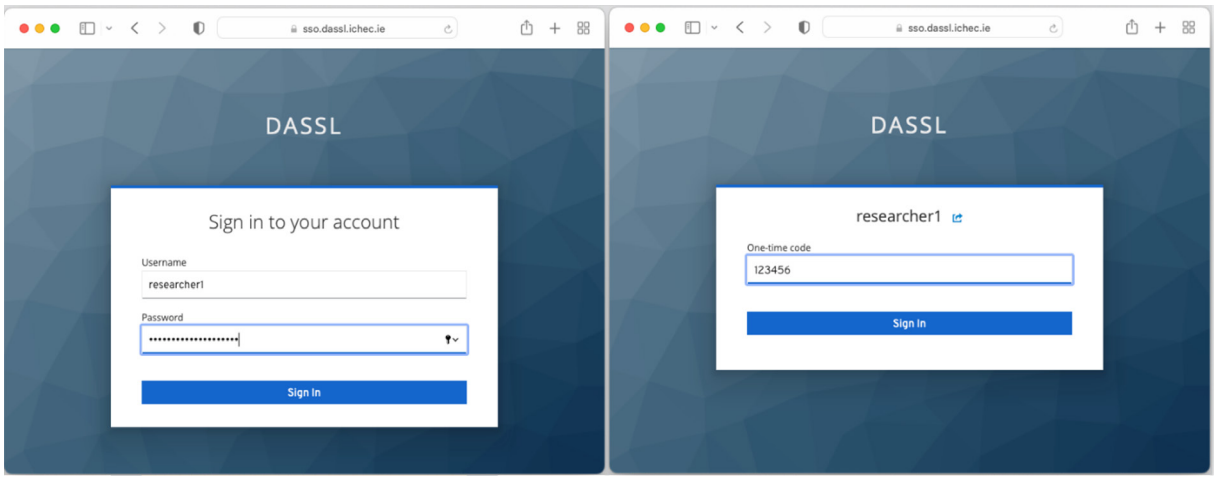


Figure 12 Visual of using 2FA to access the DASSL PoC system

9.4.2 Network security controls

The environments operated by the RSU, DLU, and researchers must be securely locked down in order to prevent unauthorised access and/or modification (including removal and export) of data either maliciously or unintentionally. All incoming and outgoing VPN traffic should be monitored and filtered based on predefined network security policies, using host-level and web application firewalls, and with all files and data ingested into the system being tested using antivirus software, rootkit hunters, and antimalware software. Intrusion detection and prevention services, which monitor and detect suspicious network or system activities, provide an additional layer of security for DASSL systems. International best practice also supports regular penetration and vulnerability testing for safe havens on an annual or biennial basis. For the DASSL PoC, Internet access, data download, and copy-and-paste functions have been disabled for the RSU, DLU, and researcher roles, whose respective systems (or VMs) should be provisioned based on a security template and policies that are audited regularly in order to ensure compliance.

9.4.3 Data protection

Data in transit and at rest should also be secured using encryption and anonymisation techniques. These are discussed in further detail in Sections 9.5 and 9.6. There are measures in place where DASSL systems prevent data export to external systems (Section 9). Additionally, technical measures can be applied to the outputs in order to assess whether the data are anonymised and can be released to the researcher outside of the safe haven.

9.4.4 Physical and environmental security

Safeguards should be in place at the physical location (typically a data centre) where the hardware and software infrastructure is situated. Access to the data centre must be strictly controlled (e.g. only approved staff may enter buildings) and there should be physical and electronic surveillance systems (e.g. CCTV) in place. The physical security measures and policies must be reviewed on a regular basis.

9.4.5 Operational security

Standard operating procedures and policies must be in place to cater for change management, e.g. system and software changes and upgrades. An electronic ticketing system is recommended for tracking and documenting such changes. Antivirus and antimalware software should be updated and executed on a regular basis. There should also be clear guidelines, set out in writing and disseminated to users via training and documentation, for the reporting of any real or perceived breach to security. Data backups should be implemented for all systems for disaster recovery purposes. For research reproducibility purposes, research VMs and data could be archived (in encrypted form) subject to data controllers' agreement and/or information governance approval. It is not anticipated that the DASSL infrastructure should be operated as a high-availability service (≥99% availability), i.e. system downtimes should be scheduled for routine maintenance, but should nevertheless offer a robust and reliable service with typical service availability of ≥95%.

9.4.6 Audit controls

A comprehensive auditing system must be in place in order to track the following activities:

- Access to different DASSL systems, including 2FA attempts
- Regular antivirus and antimalware scans
- Firewall traffic summaries, and
- System backups and updates.

Although these logs can be centralised, it is recommended that the original logs be sent directly to multiple parties in order to prevent modification and tampering.

9.5 Encryption

Data encryption encodes or scrambles messages or files so that they can only be read by someone with a corresponding key to unscramble them. End-to-end encryption is an important component of a national DASSL service as it secures data in transit. This is particularly relevant for the initial import of data from data providers into the DASSL environment. However, there are abundant tools and services, both open-source and commercial, that will facilitate this type of secure data transfer. As an additional layer of security, the data provider may also encrypt the data prior to sharing them with a national DASSL service.

Encrypting data at rest will add a further layer of security, but at the expense of added hardware/software resources and cost. This was not implemented for the PoC but could be considered for the national DASSL system, where data are stored on encrypted volumes by default. This mainly protects against malicious intrusion by third parties (rather than by DASSL users/staff who already have access to unencrypted data), which should largely be mitigated by other measures (e.g. the separation principle of personal and content data, locked-down environments, and access restrictions). While physical security measures are in place to secure the location where data are stored, other safe havens do not typically enforce encryption of data at rest. However, this additional layer of security would be important should a public cloud environment be used to store data. Encryption may also be employed during the linkage and pseudonymisation process, as discussed in Section 9.6.

9.6 Pseudonymisation and anonymisation

Anonymising data irreversibly prevents the identification of the individuals to whom the data relate by means of singling out individual data subjects, association by linked datasets, and inference of individuals, whereas it is possible to reidentify the data subject from pseudonymised data using the underlying or related data, so they must be treated as personal data under the GDPR (136). Both types of data will be used as part of a national DASSL service, and techniques to pseudonymise and anonymise data will need to be employed.

9.6.1 Pseudonymisation techniques

The data providers, DLU, and RSU can apply pseudonymisation techniques to allow datasets to be combined while protecting the privacy of the data subjects. Several techniques can be used to pseudonymise personal identifiers according to the European Union Agency for Cybersecurity (ENISA) (137). A counter or a random number generator can replace the identifiable information with incremental or random numbers, respectively. This simple method provides privacy as the numbers do not relate to the individuals' personal information. However, the complete pseudonymisation mapping table would need to be securely stored, repeated numbers must be avoided, and scalability may be a challenge.

For a national DASSL service, a counter or random number generator may work fine for the data provider step and for a distributed model, which does not need to store the pseudonyms on an ongoing basis and/or reidentify individuals. Other methods, such as hashing and encryption, can overcome the challenges discussed above. Hashing transforms the personal identifier into another pseudo-random value and is a one-way function, unlike encryption, which is a two-way function that allows another person to use the key to unlock the data. As the RSU does not need to reidentify data subjects, hashing may be the preferred option. However, simple hashing is sensitive to brute force and dictionary attacks. To overcome this risk, and to provide an extra layer of security, the hashing process can be conducted by adding project-specific 'salt', i.e. an extra piece of random data appended to identifiers, so that the hashed pseudo-identifiers are different for each project.

9.6.2 Anonymisation techniques

After combining the pseudonymised datasets for the researcher, the RSU may apply anonymisation techniques to the content data. This would be an additional data protection measure that is applied in order to ensure that the data minimisation principle is adhered to, or where a person could potentially be reidentified based on their attributes when their information from several datasets is combined. Generalising or diluting the attributes of the data subjects is often performed, which involves modification of the respective scale or order of magnitude of the data (i.e. year of birth rather than full date, a county rather than a town) (138). Aggregation, suppression, and *k*-anonymity techniques are methods that aim to prevent a data subject from being singled out by grouping them with at least '*k*' other individuals. These methods are often employed for statistical disclosure control. However, anonymisation of imaging and genomics data as well as AI algorithmic models can be different and more challenging due to the nature of these data and models, meaning that more specialised approaches and advancing technologies need to be reviewed. Finally, synthetic data (see Section 5.7) is another potential form of anonymised data that could be openly provided by a national DASSL service.

10 Record linkage

The DASSL PoC explored linking records across datasets.

To simulate real-world data in Ireland, synthetic personally identifiable information (i.e. first name, surname, Personal Public Service Number (PPSN), individual health identifier (IHI), date of birth, sex, ethnicity, nationality, address, electoral division, small area code, and Eircode) was generated from openly available data from the CSO and other sources. Further information on the generation of synthetic data is available in Section 5.7. Spelling errors, data entry errors, missing data, and name and address changes were introduced into some of the datasets in order to test the record linkage software and processes. The record linkage process includes the cleansing of data, linking across datasets, and sharing of the linkage key with the RSU. The linkage processes undertaken and tested, as well as the learnings from these tests, are discussed in Sections 10.1–10.8.

10.1 Linkage types

In addition to linking individual persons across datasets, families and locations (e.g. households, healthcare organisations or professionals, and local areas) can also be linked together. Linking of family members provides extremely valuable information but it requires the relationship to be recorded (see Section 12.3). This family linkage could exist in the dataset being linked (e.g. if linking mothers with their children, CSO Births could be used), it could be built into the population spine (as in the Western Australia Data Linkage System (WADLS) (53), or it could be based on address (as with the Secure Anonymised Information Linkage (SAIL) Databank in Wales) (139). In Ireland, Eircode would likely be required in order to support the SAIL methodology, as the same address may be used across many different households (which may also have the same surnames), especially in rural areas. Conversely, the WADLS methodology may create additional ethical issues around storing family linkages.

Healthcare organisations or providers can also be linked, and this is relatively simpler, as they are usually consistently coded with a limited number of possible matches (see Section 7.1.7). However, pseudonymisation or anonymisation of the healthcare organisations/providers would be recommended for research purposes in order to avoid inappropriate comparisons of individual providers and hospitals.

Linkage of location based on small area code or electoral division also allows evaluation of the impact of the area (e.g. level of deprivation, environmental factors) on healthcare outcomes (see Section 7.1.8). Notably, as the healthcare organisation/ provider and electoral division may be considered less sensitive and identifiable than an individual's address, this information could be provided to the RSU directly as content data and pseudonymised for research purposes, whereas the DLU retains the ability to use addresses for record linkage purposes.

10.2 Data preparation, cleansing, and harmonisation

The personally identifiable data received by the DLU need to be in a standard format to facilitate accurate record linkage. This includes common file formats (e.g. CSV), data fields (e.g. first name, surname), and data formats (e.g. 'DD/MM/YYYY', and 'ODonnell', 'O'Donnell', or 'O Donnell'). As discussed in Section 7.3.4, application of consistent interoperability standards by data providers could reduce the resources required for this step, or render it not required in some cases. However, this step is critical in Ireland at present. The responsibility for cleansing and standardising the data could fall on either the data provider or the DLU, depending on the available resources and the expertise of the DLU and data provider to complete this step.

To enable recombination of the content data with the linkage key created by the DLU, the data provider could also be asked to give an ID to each individual on the dataset and to apply the same ID to their corresponding content data on the file sent to the RSU. Alternatively, the DLU can use the row number as the ID, but this requires that both the personal identifiers and the content data remain in the same order. Additionally, internal linkage of each dataset may also be completed prior to linking across datasets in order to identify duplicates in the dataset resulting from an error or because several rows correspond to one individual's healthcare interactions (e.g. the Primary Care Reimbursement Service (PCRS) has one row per drug prescribed).

10.3 Linkage methods

Linking can be undertaken using deterministic and/or probabilistic methods. While deterministic methods require the exact data to appear in both datasets, probabilistic methods take into account dynamic variables (e.g. name and address changes), missing data (e.g. only townland available in dataset), and errors (e.g. incorrect name entered or name misspelled). As not every health and related dataset in Ireland has a consistently entered unique identifier, the probabilistic method is largely required at present. The result of probabilistic linkage is a set of pairs – matching records between two or more datasets – together with a similarity score.

Given a certain threshold, every pair with a similarity above that threshold is considered a correct match and returned in pseudonymised format to the RSU. The cut-off scores for similarity in our case studies have been derived empirically by the PoC DLU after conducting record linkage and the best matches for each record inspected, but these would need experimentation in order to automate/assist the process further within the DLU of a national DASSL service (51, 140).

International data linkage models have also used algorithms such as Lexicon matching and Soundex codes, which index names based on how they sound rather than how they are spelled (e.g. 'Smith' and 'Smyth' would be categorised as the same) in order to combat misspellings. However, as in Wales (10), Irish-specific variants would need to be introduced (e.g. 'Cliona' and 'Clíodhna'). Additionally, other challenges that could be overcome with algorithms or clerical review of matches include the different versions of names that can exist. For example, 'Margaret' could be 'Mags', 'Maggie', 'Mairead', 'Peg', 'Peig', or 'Peggy' on another record, and the same address could be recorded as 'Dingle', 'Daingean', 'An Daingean', or 'Daingean Uí Chúis'.

10.4 Blocking strategies

From a performance point of view, deterministic linkage can quickly increase in complexity: without any further context, linking 'n' records from one dataset to 'm' records from another will require comparing n*m records for similarity. When dealing with large datasets with thousands or millions of records (e.g. the Hospital In-Patient Enquiry (HIPE) dataset), this is no longer desirable or doable in a reasonable time, and probabilistic matching is also a compute-intensive task. As such, the larger the datasets to be matched, the exponentially greater the calculations required. It is therefore critical that a 'blocking' strategy be put in place, which would group records in smaller matching 'blocks' that will be evaluated against each other. For example, if it is expected that date or year of birth data are more reliably collected compared with name data only, which may be more at risk of spelling errors, then the exact date may be used as a blocking field, and each record from Dataset A will only be matched to records from Dataset B that have the same date/year of birth (140, 141). As described in the learning from the case studies (see Case Study #1 in the annex to this report), this strategy is sufficient to run probabilistic linkage even on large datasets, as there is often a way to dramatically reduce the potential matches.

When devising a blocking strategy, it should be kept in mind that typos or incorrect data in the blocking fields will result in records being erroneously excluded as candidates for matching. As much as possible, blocking fields should be reliable/high-quality fields where very few errors are expected.

For example, an address field would be a very poor blocking field (it could be written differently or it could have changed), as would a name (multiple spellings, ambiguities arising from middle name variations and usage), but a year of birth should be more reliable. Parallelising the linkage across multiple cores/nodes/computers is also a way to reduce the overall time required for the linkage.

10.5 Unique identifiers

Unique identifiers used in Ireland include the PPSN, IHI, General Medical Services (GMS) number, Long-Term Illness (LTI) number, Drugs Payment Scheme (DPS) number, and Eircode. Additionally, some identifiers are used by specific hospitals (medical record number (MRN)) and data controllers, but these are specific to that dataset, allowing them to be used to link individuals within that dataset only. To support deterministic linkage using unique identifiers, the same unique identifier needs to be used on every dataset, or, alternatively, a population spine that maps the identifiers could be used (e.g. the IHI register contains the PPSN).

A benefit of using the IHI as the unique identifier for linkage is that anyone who interacts with the health service can be provided with one, which follows them for life. The HSE Health Identifiers Service is currently making strides in the application of the IHI to critical health datasets (such as cancer screening and COVID-19 datasets), with high matching rates achieved. However, the IHI would likely not be available on social datasets, but could be linked to these datasets with the use of a population spine. On the other hand, the PPSN has now been used for some health and social datasets, but challenges occur for those who do not have a PPSN (e.g. asylum seekers, partners of individuals with a work visa, newborns) and where individuals have multiple PPSNs (e.g. for marriage or tax reasons). Of note, the IHI register works to link multiple correct PPSNs related to the same individual. With the use of any unique identifier, incorrect integers could be entered for an individual (e.g. a partner's PPSN could be used), and in these cases, other personally identifiable information needs to be used for linkage.

Where the same unique identifier is consistently used across all datasets, the linkage is done deterministically and becomes trivial, as is the case in Finland. This could negate the need for a population spine. Additionally, the need for a securely separated DLU (i.e. a trusted third party) may not be necessary, as the data providers could encrypt the identifiers (if they have the expertise and resources to do this), supporting privacy preserving record linkage (PPRL) (76).

However, as long as important legacy records only include names, addresses, etc., more than one unique identifier is used across datasets, and errors are possible on entry of that unique identifier, probabilistic matching and a separated DLU will likely be an important data protection mechanism in a national DASSL service.

10.6 Population spine

A population spine is a register of individuals. In the context of probabilistic linkage, the existence of a population spine constitutes a major benefit, as it represents a clean superset of the personally identifiable data of known individuals, ideally collected over time. This means that a population spine could, for example, track past addresses of an individual, allowing record linkage to work properly irrespective of whether the dataset is being linked using a current or previous address. As Ireland does not have unique identifiers consistently applied across datasets, a population spine is critical for high-quality linkage and in order to enable a centralised data hub model, which otherwise would not be possible in Ireland. The linkage process differs depending on whether a population spine is available or not. Where a population spine is not available, the DLU would need to be given a clear, step-by-step process for linking each dataset, whereas with a population spine, every dataset is simply linked to the population spine.

An existing register could be used as the spine, as is the case in the UK and Canada. In Ireland, the Department of Social Protection holds a register of every individual with a PPSN. The IHI register was built from this and now also includes some people without a PPSN (e.g. those in the COVAX dataset), and eventually it may contain newborns who do not yet have a PPSN (e.g. from the Maternity and Newborn Clinical Management System (MN-CMS)). A new population spine could also be built for the specific purpose of a national DASSL service, like what was developed at South Australia and the Northern Territory DataLink (SA NT DataLink) and the Centre for Health Record Linkage (CHeReL) in Australia.

These organisations created and maintained a spine from the health datasets themselves as well as from the birth, death, and marital registers. Creation of a population spine in this manner requires a lot of time and resources to perform clerical review, but it stores links from the datasets used to build the spine. In addition to a population spine, an address spine (e.g. Eircode database) could be used, as in Wales – but, of note, Eircode is also stored within the IHI register. Therefore, the IHI represents a ready-made and maintained population spine in Ireland that could be used for a national DASSL service.

10.7 Clerical review and linkage quality

Clerical review of record linkage involves RSU staff manually assessing a subset of the linkages, which is a labour- and resource-intensive process even with assistance from computer software (90). The process is used to alter the probabilistic matching threshold (e.g. predefined data fields need to match exactly reinforced by inexact matches from other data fields in order to be considered a positive match) as required and subsequently assess the quality of the linkage for the researchers (i.e. estimate false positives and negatives, determine if certain cohorts of people are not linked) (1, 8, 9). For example, clerical review may assess matches that are close to the matching threshold to see whether they refer to the same person or not. This can be challenging in cases such as twins, where the surname, sex, date of birth, and address could all match. Where unique identifiers are consistently applied correctly, there will be 100% matching, and neither clerical review nor a population spine is required.

For research purposes, the researcher may be interested in seeing both the linked and non-linked data (e.g. to assess whether there are inherent biases in the linked dataset), but this may not be compliant with the data minimisation principle or approved by the ethical and information governance boards. Thus, a linkage quality report could be provided to the researchers that provides summary information on the linked records as well as those that have not been linked.

10.8 Linkage software

Different software packages for record linkage were identified during the landscape analysis phase of this PoC project (Table 2). Most international data linkage models have developed their own linkage software, many of which are open source. For the purpose the PoC, a software package that met our requirements for this project was selected using R. PPRL provides a toolbox for record linkage that combines the functionality of the Merge ToolBox software package and privacy preserving techniques (142). From a combination of literature reviews (based on gold-standard, manually curated real datasets) and practical implementation, the performance of algorithms/methods used by PPRL is comparable to those used by other record linkage tools in terms of linkage accuracy (143). However, it is recommended that further testing of linkage software be conducted on real health data for a national DASSL service.

Table 2 Linkage software packages
Statistical software packages (e.g. Stata, SAS)
Programming languages (e.g. R, Python, SQL) and dedicated record linkage libraries (e.g. PPRL package for R)
Choicemaker (proprietary software used by CHeReL)
Data Linkage System Number 3 (DLS3) (developed by WADLS)
Freely Extensible Biomedical Record Linkage (Febrl) (open-source software used in combination with LinkageWiz and SQL scripts by SA NT DataLink).
LinXmart (open source and included in the Secure eResearch Platform (SeRP)
Matching Algorithm for Consistent Results in Anonymised Linkage (MACRAL) (SQL-based and used by the National Health Service (NHS) Wales Informatics Service (NWIS)
AutoMatch (used by the Institute for Clinical Evaluative Sciences (ICES))
G-Link (used for some sites in Canada)
LINKPRO (used by the Manitoba Centre for Health Policy (MCHP))

11 Content data management, preparation, and access

As the personal identifiers are being linked by the DLU, the RSU will receive the content data within the data hub. A number of steps are involved in the processing of the content data, from collection from the data providers and storage, to the creation of the data view and release of the findings to the researchers.

11.1 Content data collection and storage

The content data received from the data provider may be gathered on a project-by-project basis (i.e. distributed model), stored in a pseudonymised format for future projects (i.e. centralised model), or a combination of both (i.e. hybrid model). While some countries (such as Finland and Scotland) operate distributed data linkage models, Wales, Australia, and Canada operate centralised and hybrid models. A number of advantages and disadvantages to storing pseudonymised datasets on an ongoing basis are outlined in Table 3. Some data trusts have started as distributed models and over time, with data provider and public trust, have began storing some of the data centrally (52). Of note, a federated model where data remain with the data controller is another possibility for a data sharing model, as proposed for the European Health Data Space (EHDS) and for the Personal Health Train in the Netherlands (39).

However, this would only work where data are stored in a consistent format, and it would be very difficult to link individuals in a federated model, as the analysis needs to be run on siloes of data. In addition to data controller and public trust in the centralisation of data, the selection of the technical infrastructure provider could impact on this decision. In France, the Health Data Hub was provided via Microsoft Azure, and the use of this provider has reportedly contributed to the Health Data Hub withdrawing its request to the French Data Protection Authority (*Commission Nationale de l'Informatique et des Libertés;* CNIL) to store the health database (144).

Table 3 Advantages and disadvantages of a centralised versus distributed model*		
	Centralised	**Distributed**
Project turnaround	Project turnaround is more efficient.	Project turnaround depends on data providers, and can be slower.
Resources	Maintaining, storing, and updating datasets requires staffing resources and infrastructure within the data hub. Preparing and cleansing data is done once and the data are used many times. The DLU links the data once.	Datasets are destroyed after each use; thus, fewer resources are required. Preparing and cleansing the data is done for every project. The DLU links the data for every project unless permitted to store the links within a data spine (e.g. CHeReL).
Data protection	Personal data are being stored, and this requires necessary technical and organisational security. A lawful basis is required for storing, sharing, and linking personal data.	Personal data are only stored as long as required for each project. A lawful basis is required for sharing and linking data.
Trust	There may be concerns over data centralisation and associated risks.	There may be fewer concerns when ongoing storage of data is not permitted.

*A hybrid approach can also be used.

11.2 Researcher data view preparation

The RSU creates the data view to be accessed by the researcher in the safe haven. The data fields and level of granularity of the data view will be determined during the approvals process. Where the data controller only shares the approved project data with the RSU in a distributed model, the amount of processing and subsequent resources required by the RSU are reduced.

However, if the data provider shares the entire dataset (i.e. in the case of a centralised solution or where the data provider does not have the resources to create a project-specific dataset), the RSU will need to remove the data fields that are not required for each research question in order to comply with the data minimisation principle and approval boards' requirements. In each case, the RSU will need to combine the datasets using the linkage key provided by the DLU, and where the researcher does not have approval to see non-linked individuals, the RSU may then remove these columns from the data view. Once the data view is created, the RSU will apply a project-specific identifier to each individual in the dataset. Whether a distributed or centralised model is employed, the data view is usually archived at the end of the project in order to allow the results to be checked at a later date, if required (e.g. for publication).

11.3 User access to content data

Researchers or other users can access the safe haven either in person at a physical national DASSL service, or virtually via secure mechanisms. Under certain defined and approved circumstances, international or other users are usually permitted to access the data within an external safe haven. For example, in Finland (145) and Australia (146), if genomics data need to be linked with other data from the data trusts, the genomics data do not leave the biobank but the tabular data are shared with the researchers in their own secure environment. When using the safe haven of a national DASSL service, the researcher needs to liaise with the RSU to capture the research requirements and to ensure feasbility of the intended data linkages and analyses. A number of commonly used software and analytical packages should be available in every case, but the safe haven may need to meet additional requirements, and this could incur additional cost.

Furthermore, the researcher may wish to bring in their own code or other data, which would need to be sent to the RSU and checked for viruses before being imported into the safe haven. The researcher can then access the data, analytical packages, and any additional information they import into the safe haven. Once the data have been analysed and the outputs/findings are ready, the researcher must place these in a folder for exporting. In the PoC, as with the CSO and internationally, the outputs are shared with the RSU for disclosure control (see Section 8.7 for further information). However, at Population Data British Columbia (PopData) in Canada, this responsibility is left to the researcher. Of note, at ICES in Canada and at NHS Digital in the UK, staff at the RSU or equivalent can in some cases perform analysis on linked data for research purposes.

This leverages the expertise of the RSU who work closely with the datasets on a regular basis which has the potential to better exploit the value of linked datasets for research. However, the governance and approvals process for such internal access and analysis may differ from that subjected to by an external researcher.

11.4 Output sharing and publication

Once the findings have been assessed and shared with the researcher external to the DASSL system, the findings should be shared publicly. Therefore, all outputs (including publications) should be made available, which could include publication on a national DASSL service's website. Data controllers or the RSU may be allowed to request to review any publications in order to ensure correct interpretation of the findings and proper acknowledgement of the contributions of a national DASSL service and the data controllers.

Table 4 Inclusion criteria for case studies

Datasets	Record linkage	Study designs	Research questions	Analyses	Data management models
Administrative	Linkage types	Large-scale whole population	Virtual patient register/prevalence	Descriptive analysis	Centralised (i.e. pseudonymised data stored on an ongoing basis)
Patient registries	Individual persons	Cohort studies	Trends	Inferential analysis	Distributed (i.e. data gathered on a project-by-project basis)
Clinical records	Family and/or households	Purposive sampling	Identification of predisposing factors	Correlations (relationship between variables)	Hybrid
Longitudinal cohorts	Addresses/locations	Case control	Identification of outcomes	Regressions (cause and effect)	
Health surveys	Linkage methods	Methodological research	Cost evaluation	Analysis of variance	
Operational	Deterministic using unique identifiers	Two-group comparison	Data validation	Cost analysis	
Health-related social data	Probabilistic using names, addresses, dates of birth, etc.		Characteristics of individuals		
Genomics	Population spine		Policy/clinical guidelines evaluation		
Imaging	Existing		Predictive analytics		
	Purpose-built		Development of new tools, diagnostics, or treatments		
	None				

12 Testing of the PoC infrastructure: Case studies

In order to test the infrastructure and demonstrate its potential, use cases for a national DASSL service were developed and run by a PoC DLU, PoC RSU, and PoC researcher.

Synthetic data were then generated to mimic real health and related datasets in Ireland for these case studies. This was followed by using each of these case studies to test the infrastructure, with many learnings identified from this process.

12.1 Identification of use cases

In order to ensure a broad spectrum of research questions with which to test the infrastructure and demonstrate its potential, inclusion criteria were set based on the DASSL landscape analysis, stakeholder engagement, and international examples (Table 4). Criteria included the types of datasets and data, record linkage methods, data management models, study design, analyses, and research questions.

12.2 Synthetic data generation

Synthetic data were generated for each case study and these data were processed via the DASSL PoC demonstrator. The synthetic data were largely based on existing datasets, with the exception of the synthetic genomics dataset. Following some investigation into tools to generate synthetic data based on existing trends within real data, the R package synthpop (107) was selected. The Python package Synthetic Data Vault (SDV) (147) was also used by collaborators at the Royal College of Surgeons in Ireland (RCSI) to generate a dataset. StyleGAN (or Stylized Generative Adversarial Network), published by Nvidia (148), was chosen to generate synthetic images. An anonymised online repository of 'normal' (i.e. absence of COVID-19) and 'COVID-19' lung X-rays (149) was used to train the model, requiring the use of graphics processing units and Kay (a national supercomputer provided by the Irish Centre for High-End Computing (ICHEC)).

While the project initially intended to apply these packages to real health datasets, due to ethical and data protection challenges, it was decided that relatively small synthetic versions of the datasets would be created based on the data dictionaries and national statistics provided by the data controllers. Additionally, for the personally identifiable data, census data from 1901 and 1911 were used in combination with other CSO files on common names in Ireland to create synthetic individuals using synthpop. The synthetic data package then used these datasets to learn the trends and expand the synthetic datasets, in some cases into millions of rows. There is also the potential for a national DASSL service to produce synthetic data for research, as offered by the French Health Data Hub, but a synthetic version of the linked data would need to be created as opposed to creating synthetic versions of the initial datasets and then trying to link them.

Five case studies were selected. These case studies are described briefly below, but more detailed information for each case study is provided in the annex to this report. For each case study, the annex provides details on:

- The background to the case study
- The datasets involved (and their quality and utility)
- The record linkage process
- Data preparation for viewing by the researcher
- Data analysis and interpretation of findings, and
- Lessons learned.

12.2.1 Case Study #1: Virtual patient registry (foetal valproate syndrome)

The aim of this case study was to test how the DASSL PoC could be used to evaluate the impact of sodium valproate on women and children, and the impact of the introduction of the PREVENT Programme. Synthetic versions of the following clinical records, administrative datasets, and patient registers were linked in order to answer this research question:

- The PCRS
- CSO Births
- The National Perinatal Reporting System (NPRS)
- The National Ability Supports System (NASS), and
- The epilepsy electronic patient record (EPR).

This case study was chosen to assess probabilistic linking in a distributed model without the availability of a population spine.

12.2.2 Case Study #2: Identification of social risk factors (mental health and addiction)

The aim of this case study was to demonstrate the use of the DASSL PoC to explore any potential risk factors in childhood for self-harm, suicide, psychiatric conditions, and alcohol and drug issues later in life by linking health and social data. In order to answer this research question, synthetic versions of the following datasets (which included a longitudinal cohort, patient registers, a statistical register of deaths, and an address dataset) were generated:

- The Growing Up in Ireland (GUI) 1998 Cohort
- The National Psychiatric Inpatient Reporting System (NPIRS)
- The National Self-Harm Registry Ireland (NSHRI)
- The National Drug-Related Deaths Index (NDRDI)
- The National Drug Treatment Reporting System (NDTRS)
- CSO Vital Statistics: deaths (CSO Mortality), and
- The Social Deprivation Index.

This case study was selected in order to examine the linkage of health data with related social data and address location using a hybrid model with probabilistic linkage and a purposively built population spine.

12.2.3 Case Study #3: Long-term outcomes and costs of healthcare initiative (hip fractures)

The aim of this case study was to demonstrate how the impact of introducing the Best Practice Tariff (BPT) for management of hip fractures in Ireland in 2018 could be evaluated by linking datasets using the DASSL PoC. For this case study, a synthetic version of a national clinical audit was linked with an administrative dataset, a statistical register, an operational dataset, and clinical records, as follows:

- The Irish Hip Fracture Database (IHFD)
- HIPE
- General practitioner (GP) EPR
- CSO Vital Statistics: deaths (CSO Mortality)
- Healthlink, and
- Hospital staffing levels.

This case study was selected in order to examine the linkage of very large datasets across both primary and secondary healthcare services, including healthcare providers, in order to identify the long-term outcomes of patients and evaluate a new healthcare initiative. For this case study, most of the datasets used the IHI and a population spine (the IHI register) for the record linkage, and a centralised model/ solution was applied.

12.2.4 Case Study #4: Predisposing genetic factors (cancer)

The putative research aim of this case study was to emulate the identification of novel gene mutations (in the APB gene) that correspond to incidences of late-onset colorectal cancer, relative to other known mutations of the gene that typically lead to early-onset cancer (in those aged under 40 years). Synthetic versions of the following datasets were developed for this case study:

- The National Cancer Registry Ireland (NCRI), and
- An artificial national genomics dataset.

This case study was selected in order to examine the combination and analysis of genomics data linked with tabular data. For the purpose of this demonstration, the PPSN was applied to the synthetic datasets, and only the data required for the research question were gathered (i.e. it used the distributed model).

12.2.5 Case Study #5: Image interpretation using machine learning (COVID-19)

The aim of this case study was to demonstrate the development and testing of an algorithmic model to identify the diagnosis and prognosis of a patient with COVID-19 and to determine whether receiving one or more vaccine doses impacted on this algorithm's ability to diagnose COVID-19. For the purpose of this case study, the IHI was consistently applied to each of the synthetic datasets and only the required data were shared by the data providers (i.e. it used the distributed model). The following datasets were used/generated:

- X-rays
- COVID Care Tracker (CCT)
- COVID-19 vaccination registry (COVAX), and
- Computerised Infectious Disease Reporting (CIDR).

This case study was selected in order to demonstrate the linkage of tabular data with medical images and the application of machine learning to linked data.

12.3 Learnings from the case studies

The case studies demonstrated many of the benefits, challenges, risks, and requirements of a national DASSL solution. They highlighted that, in addition to having technical infrastructure in place that supports a DASSL model, the quality of national health and related datasets is critical. The learnings are outlined below.

12.3.1 Data utility, quality, and fit for purpose

- Some datasets only capture public healthcare interactions (e.g. HIPE, the PCRS), whereas others collect data from both public and private healthcare organisations (e.g. the NPIRS).
- Population coverage can depend on whether the legal basis for data collection is explicit consent (e.g. the Cystic Fibrosis Registry of Ireland) versus a legislative basis for data collection (e.g. the NCRI).

- While multiple datasets may collect some of the same information (e.g. CSO Births and NPRS both collect information on births, and CSO Mortality and the NDRDI both collect information on deaths), combining the datasets can validate them against each other or expand on the information available from a single dataset.
- The time period of data collection and availability of data differs between each dataset, and some datasets may have some information available for longer periods than other information (e.g. the NASS has collected intellectual disability information since 1995 and physical and sensory disability information since 2002).
- Use of consistent coding standards (versus free text data entry) renders data aggregation for data view creation and analysis easier.
- Data formats of different vendor systems (e.g. GP EPRs) render it difficult to combine data across these systems.
- Some clinical records would need to be centralised and stored prior to being used by a national DASSL service (e.g. GP EPRs, Healthlink).
- Not all the data used in the case studies are currently available in Ireland (e.g. genomics).

12.3.2 Record linkage

- Probabilistic matching is required where there is no consistent unique identifier across all synthetic datasets.
- A blocking strategy is needed for linking large datasets (e.g. CSO Births).
- Matching of each dataset in a pairwise manner is required where no population spine is available.
- Not all datasets used in these case studies collect potentially linkable unique identifiers (e.g. HIPE, the IHFD).
- Linking of family members required familial relationship information to be collected in the datasets themselves.
- Linking of datasets that did not use the same unique identifiers required a population spine.

- Linking of location information between datasets is possible, and the RSU could conduct this linkage instead of the DLU if location information can be generalised (e.g. into regions) and is not considered too sensitive for the RSU to access.
- A population spine was required to facilitate the storage, maintenance, and subsequent linkage of pseudonymised datasets.
- The role of the DLU is largely redundant where consistent unique identifiers are employed across all datasets, which would negate the need for a population spine or clerical review.

12.3.3 Data view preparation

- Where the data provider only sent the required data to the RSU (i.e. the distributed model), this required less work by the RSU.
- In the case of a centralised model, the RSU needs to extract the approved data from the centralised storage location for each specific project.
- The RSU could perform some analysis on the data view prior to sending the data to the researcher in order to reduce the need to share sensitive information (e.g. to avoid sharing date of birth).
- Cleansing and harmonisation of data by the RSU helped to ensure the consistent use of codes.
- The data view will need to be checked for any other personally identifiable data.

12.3.4 Data analysis and interpretation of findings

- Information on population and time coverage, as well as on the quality and completeness of data, needs to be provided to the researcher.
- Many different statistical packages could be used for the analysis, but analysis of the images and genomics required more bespoke packages and libraries (e.g. Genome Analysis Toolkit, Tensorflow).
- Interpretation of the data by the researcher may need to be checked by the data controllers and/or RSU to ensure accurate based on their in-depth knowledge of the dataset.
- Combining individual-level data (NASS) with observation-level data (e.g. the PCRS, CSO Births) can be more challenging than combining individual-level data only.
- Data dictionaries and/or mapping of coding systems between datasets, where possible, is required in order to support researcher analysis, but this may not always be possible where versions of data fields or codes change.
- Exportation of certain outputs (e.g. an AI model) requires different considerations for output checking.
- The population size available to researchers depends on the datasets and whether one or more datasets is a purposive sample as opposed to a national sample.
- Researchers may request to import additional data or code into the safe haven.
- Allowing researchers access to free text that may assist in some research questions would require this free text to be assessed.

13 Insights into governance and approvals processes

Governance and underpinning legislation are critical to the development of a national DASSL service.

While defining the required governance was out of scope for the PoC, many learnings were gathered during the project and are shared in Sections 13.1–13.5 in relation to legislation, governance policies and boards, stakeholder boards, and project approvals.

13.1 Legislation

Existing and planned legislation in Ireland and across the EU will influence governance over a national DASSL service. This includes the GDPR and the Data Protection Act 2018, which encompasses the amended Health Research Regulations 2018. These regulations provide high standards of data protection for individuals and impose increased obligations on organisations processing personal data for the purposes of health research. However, at present, the GDPR does not appear to provide clear guidance on the lawful basis for linking health data for secondary uses, including research purposes. This has contributed to recommendations for additional EU legislation such as the proposed Data Act, Data Governance Act, the EHDS, and the Artificial Intelligence (AI) Act. A national DASSL service could support the operationalisation of the policy intent of this EU legislation in Ireland by providing the technical environment for the reuse and sharing of data using common data spaces for research, and for other purposes, such as cross-border federated analysis, business-to-government sharing, and developing AI tools for healthcare.

Across the EU, specific legislation at the national level has been, or is being, introduced to support and facilitate secondary use of health data (e.g. Findata was established as the responsible agency for linking named health and related datasets in Finland under the Act on the Secondary Use of Health and Social Data). Similarly, in Canada, the data trusts are named within each province's provincial legislation to allow them to legally collect personal health information for specified purposes. However, these organisations can usually release data (e.g. genomics data) to other entities if a need is demonstrated and if these other safe havens meet the required security criteria. Additionally, cross-sector linkage is usually supported by such legislation.

At present in the UK and Australia, linked data research is provided for under existing data protection legislation that does not specify the entities that can share and link data, but standards must be met by entities undertaking these responsibilities. For example, *A Charter for Safe Havens in Scotland* (12) defines the characteristics of a safe haven where personal health data can be processed. With the proposed Health Information Bill in Ireland, which could support a national DASSL service, Ireland will need to consider the current landscape and international learnings. The following considerations were identified during this PoC project:

- New legislation should support existing and upcoming national and EU legislation.
- A single new or existing entity to be responsible for linkage and providing access to health and related data could be named within legislation in order to avoid duplicated work, but there may be cases where external safe havens could/should receive these data.
- Storage of pseudonymised and/or identifiable data (including a population spine) may be required if the DLU exists within that named entity.
- Legislation should cover all potential data controllers across the public, voluntary, and private sectors.
- Provision of the sharing and linkage of all new and evolving data (e.g. images and genomics), as well as more advanced analytical techniques, will be important to future-proof the legislation, with any additional data protection concerns also addressed.
- The Health Research Consent Declaration Committee (HRCDC) covers consent declarations for research projects, but its remit does not extend to non-research or consented projects using linked, routinely collected health data, and consideration of the need for another information governance review panel or expanding the HRCDC's remit will be important (see Section). (13.5.3).

13.2 Governance boards and advisory committees

As with other public agencies, a governance board with expert members and representatives from key stakeholder groups and organisations is recommended in order to provide oversight of, and counsel on, the operations of a national DASSL service. In addition to a governance board, most international health data linkage centres also have advisory committees.

The most common and critical of these is a public advisory committee, which guides the agency on the public-facing content; priorities in health research; policies and procedures; and new data opportunities and partnerships (9, 150, 151). Members of public advisory committees in other jurisdictions also sit on governance boards, ethics committees, interview panels, and information governance boards (150, 151). Other advisory committees have included international and national scientific experts (13, 56), data controllers (60), and researchers (28).

13.3 Policies, agreements, and standard operating procedures

Many different policies, standard operating procedures (SoPs), and legal agreements would be required to ensure the security and operations of a national DASSL service. These should cover security and risk management in alignment with the relevant security accreditation (e.g. ISO 27001); public communications (i.e. public involvement and engagement policy); privacy statements; a Data Protection Impact Assessment (DPIA); data sharing agreements with data controllers; and data access agreements and declarations of secrecy with anyone who sees the data (e.g. researchers, RSU, DLU). The roles and access of users will also need to be outlined, as well as the provision of access to data controllers and researchers (see Section 8.3). As data controllers or custodians may have different internal policies in relation to sharing data, collaboration will be critical during the development of policies for a national DASSL service.

13.4 Research accreditation and training

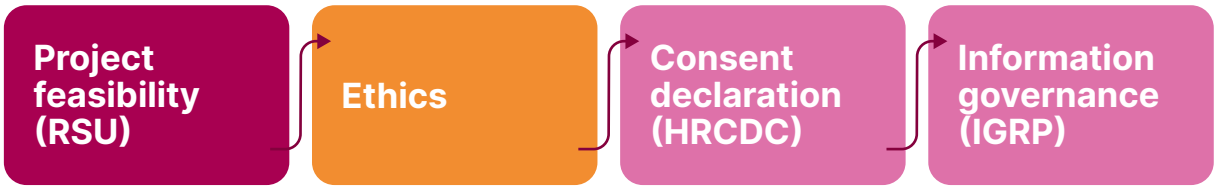
Researchers and other users should be trained and accredited in using a national DASSL service. Accreditation often includes being from a recognised research-performing institute, providing a curriculum vitae, and completing data protection and safe researcher training. Garda Vetting may also be included, but, as the researcher will have no access to the data subjects or methods of reidentifying them, this will likely not be required. Once the training is completed, this accreditation should last for a stated period (e.g. 3–5 years) before users need to retake the training course (unless there is a major system change). While an existing data protection course may provide much of the relevant training required for researchers, and can reduce the responsibility of the RSU to produce and conduct a training course, there will be additional elements specific to a national DASSL service (such as multi-factor authentication and the need to sign out from the system when idle) that may require the development of a specific training programme.

Access by international researchers meeting the same criteria is also usually allowed in other countries. However, additional consideration of whether and how researchers from the private sector could be accredited for access to a national DASSL service is required, with involvement from patients and the public.

13.5 Project approvals

In order to ensure safe projects, a project that requires use of a national DASSL service should be assessed by a number of boards. This will differ depending on the user and their use case.

1 A researcher requesting access to linked, routinely collected datasets without consent



Where consent is not feasible for a research project involving the linking of datasets, the researcher may follow the above project approval steps. However, further consideration is needed regarding whether the HRCDC alone should assess the information governance, or whether a specifically designated Information Governance Review Panel (IGRP) is also required to review these steps for linked data research. An IGRP will be required for other types of projects (mentioned in point 2 and 3 below) that could use a national DASSL service.

2 A researcher requesting access to link their own research data to routinely collected data with individual participant consent

If a researcher has collected data directly from volunteers, consent for linkage with other routinely collected datasets is feasible and likely to be obtained. However, even with consent, information governance over using the routinely collected data may be required.

3 Other secondary data (non-research) uses

Project feasibility (RSU) → **Information governance (IGRP)**

If the project is not for research purposes, research ethical approval and a health research consent declaration are not required. However, because assessing information governance will be important for these projects, an IGRP for non-research projects may be required.

13.5.1 Project feasibility

Prior to applying for full approval, the researcher needs to determine whether the project is feasible and whether the required data are available. This approval process would need to be conducted by individuals or entities familiar with the data, which would most likely be the RSU with or without input from the data controller. This is similar to how the CSO COVID-19 Data Research Hub and many other data controllers operate, which encourages researchers to discuss feasibility prior to submitting their applications.

13.5.2 Research ethics

According to the "Data Protection Act 2018 (Section 36(2)) (Health Research) Regulations 2018 (S.I. No. 314 of 2018)", health research that uses personal data requires research ethical approval. Therefore, any application seeking to conduct research using pseudonymised data within a national DASSL service would require ethical approval. A key learning from the Australian experience was that a single national/regional ethics committee should be established for linked data research (44). The National Research Ethics Committees Bill 2019, although not yet legislated for, has laid the foundation for the creation of the National Office for Research Ethics Committees and the first three National Research Ethics Committees (NRECs) for COVID-19, clinical trials, and medical devices (152). Linked data research requires a specialised knowledge and understanding of the related ethical implications. Additionally, patient registers and other national datasets do not have a single ethics committee, but this has been advocated for by Health Research Charities Ireland (153). Therefore, an NREC to review research projects using national datasets is recommended. This would streamline the approvals process and support consistent decision-making by an appropriate and informed committee.

13.5.3 Research consent and consent declaration

The Data Protection Act 2018 (Section 36(2)) (Health Research) Regulations 2018 (S.I. No. 314 of 2018) provided for the establishment and operation of the HRCDC. The HRCDC makes decisions on applications for consent declarations where obtaining consent is not possible but where the public interest of doing the research significantly outweighs the need for explicit consent (154). In most research project applications to a national DASSL service, consent would not be feasible within the current infrastructure and thus a consent declaration would need to be sought. This is a similar requirement to the feasibility requirement of the COVID-19 Data Research Hub. An application to the HRCDC must include a DPIA, ethical approval, and transparency arrangements, and must demonstrate engagement with the public and patients, as appropriate. Where a researcher has obtained informed consent to perform this linkage, a HRCDC consent declaration would not be required.

13.5.4 Information Governance Review Panel

Internationally, IGRPs assess the information governance issues of project applications, such as consent, privacy, and public interest, and weigh up the risks and benefits. IGRPs usually include representatives from different sectors (including the regulatory and public sectors) and operate a two-tier structure for assessing project applications. Where a researcher has consent to perform record linkage, the IGRP is often still required internationally, but this issue requires further consideration for the Irish context. Additionally, an IGRP can review non-research projects using linked, routinely collected data. As part of the CSO COVID-19 Data Research Hub, a Research Data Governance Board was established in order to review project applications. Its role involves determining the eligibility of applicants, assessing the validity of projects, and ensuring that other regulatory approvals are secured (e.g. HRCDC, ethics). The learnings from this initiative demonstrate how an IGRP could operate alongside the HRCDC and NREC. However, to avoid duplication of work and onerous application processes for researchers (involving up to four different approvals boards), further consideration as to the distinct roles of these panels and boards is required for a national DASSL solution.

13.5.5 Research project application

Research project applications would need to be completed for submission to these panels and boards. A single digital data access application form in Ireland would benefit a national DASSL service and could reduce the workload of both researchers and data controllers. Additionally, a single application form for all approval boards, if possible, would be more time efficient than the researcher repeating information on four different applications to the RSU, IGRP, NREC, and HRCDC.

However, this would require collaboration across data controllers and approval boards in order to ensure that all required information is included. Such collaboration was demonstrated by all actors in the course of establishing the COVID-19 Data Research Hub. Research project applications should also include a DPIA and evidence of public and patient involvement (PPI).

13.5.6 Project cost recovery model

Internationally, researchers and other users are usually charged in order to cover the cost of data preparation by both the data controllers and the RSU, as well as use of the secure locked down environment. A cost recovery model is commonly used, the cost of which differs depending on the project size (e.g. number of datasets/ controllers and pre-processing required), the user/purpose of the data (e.g. thesis candidate versus established researcher, or health service versus commercial user), the additional software packages required, and the computing capacity required. Examples from costing models in Scotland and Finland are provided in Table 5.

Table 5 International costing models

Scotland (electronic Data Research and Innovation Service (eDRIS))		Finland (Findata)	
Study (small–large)	NHS/public sector: GB£4,688–17,580 (Great British pounds) Academic institutions/charities: GB£6,096–22,860 Commercial/industry: GB£14,064–52,740	Data access and requests*	Thesis: €250 Normal: €1,000 Extensive: €3,000
Computing and disclosure (per year)	NHS/public sector: GB£804 Academic institutions/charities: GB£1,045 Commercial/industry: GB£2,412	Computing package (per year)**	8–64 GB RAM: €2,250–5,525
Indexing support	NHS/public sector: GB£2,344 Academic institutions/charities: GB£3,048 Commercial/industry: GB£7,032	Data processing	€115 per hour

* Cost for data requests can vary depending on dataset(s). Additional costs can be incurred for amendments and lapsed applications.

** Additional charge for customisation with packages, and users can pay per month.

14 Importance of stakeholder involvement and engagement

Stakeholder trust in a national DASSL service – as well as stakeholder involvement and engagement in its development – is critical, according to international learnings.

For the purpose of this PoC project, a stakeholder committee with representatives from the Health Service Executive (HSE), the Department of Health, the Health Information and Quality Authority (HIQA), hospitals, the CSO, academia, patient groups, and the public was established, and a similar committee for a national DASSL service would be recommended. Along with other additional meetings and workshops with different stakeholders and experts, this finding was extremely informative for the project. However, a national DASSL service would require further involvement of these key parties in its initial and ongoing development. This includes patients and the public, researchers and other users, data providers and controllers, policy-makers, healthcare professionals, and the private sector.

14.1 Public and patient involvement

While the public and patients are generally supportive of their data being used in the public interest, information and reassurance on the benefits and appropriate governance and data protection procedures in place will be critical for a national DASSL service. PPI was an integral component of the DASSL PoC, and further public engagement and knowledge sharing on the linkage and sharing of health data is recommended for a national DASSL service. Public advisory boards (to provide regular and ongoing input on developments in health data and communication with the public (150)), as well as ongoing research into public attitudes – especially in relation to new developments in AI and private access – and a public engagement officer and a communication and engagement policy are also recommended, based on international learnings (155). Other international initiatives (such as by the Joint Action Towards the European Health Data Space (TEHDAS) and Data Saves Lives (156, 157)), and those closer to home (such as by HIQA and the Irish Platform for Patient Organisations, Science and Industry (IPPOSI) (158, 159)), will also provide key learnings for a national DASSL service and how the public is engaged.

Finally, PPI in individual research projects is a prerequisite and highly recommended for applications to the HRCDC and many other bodies, and IPPOSI and the national PPI network are supporting this work.

14.2 Data controllers

In addition to governance and a lawful basis for sharing data, building trust and engaging with data controllers from the outset will be critical to facilitate the sharing and use of health data with a national DASSL service. There are many different data controllers across the public and private sector, many of whom have their own policies for data sharing and who control several datasets with different data dictionaries, formats, and coding systems. Data controllers' expertise will be critical in ensuring that the datasets are correctly analysed and interpreted, and the RSU will need to work closely with the data controllers. Additionally, while the PoC gained feedback from data controllers in relation to the overall concept and PoC technical infrastructure, further input on the approvals process and mechanisms for sharing and preparing data will be required from all data controllers (e.g. whether data are uploaded on a project-by-project basis or stored on an ongoing basis). In addition to one-to-one meetings, data controller engagement and involvement can be facilitated by appointing data controllers to advisory committees and governance boards, as well as via surveys.

14.3 Researchers and other users

Resourcing to establish a national DASSL service needs to support researcher requirements, both initially and into the future. This may include access to the platform (e.g. approvals process, cost, and time) and the availability of data and the granularity of those data, as well as the software packages and computing power available to researchers. Therefore, the ongoing requirements of researchers from various institutions across Ireland should be identified. For a national DASSL solution, this could be achieved by appointing researchers or other users to advisory committees and by gaining regular feedback from them via surveys or interviews.

14.4 Policy-makers and healthcare providers

For many use cases of linked data research, the findings will be extremely important to decision-makers, whether for policy or guideline development or for the management of individual patients. Therefore, the needs and requirements of this group will also need to be identified. For example, findings will need to be shared with these knowledge users in a useful and understandable format, and knowledge users may provide input on research questions of critical importance. This group will also need to be engaged with for a national DASSL service to succeed. The quality of the data and use of codes often depends on the healthcare providers inputting the data, and the policy-makers are critical to the funding, governance, and legislation required to support a national DASSL service.

14.5 Private sector

The private sector also comprises key stakeholders for a national DASSL service, either as data providers, potential infrastructure providers, or users of the findings and/or linked data. Many health data systems are provided by private vendors in the public and private health sector (e.g. Clanwilliam Ireland is the vendor for three of the GP systems used in Ireland, Ergo provides the epilepsy EPR, and Cerner is the vendor for the St James's Hospital EPR). Therefore, mechanisms for sharing data from these clinical information systems (such as the aggregation of data from GP practices) may need to be developed with these vendors. It was also noted during the development of the PoC that the HSE uses Microsoft Azure for its data lake and that many international data linkage models use services from Microsoft and Amazon, as well as from other companies, for their technical infrastructure.

Finally, the European Commission has recognised the importance of the private sector's access to health data in order to support the development and evaluation of drugs and other medical treatments. This has led to consideration of private access within the proposed legislation for the EHDS (160). While future private access to a national DASSL service should be considered, this will require extensive public engagement and consideration of the approvals and privacy processes in relation to what types of private access are permitted, and under what circumstances, in order to build and gain public trust.

14.6 International and national expertise

Finally, input from national and international experts on the different components of a national DASSL service will be extremely beneficial for its successful implementation and ongoing use. As part of this DASSL PoC, engagement with international and national experts informed the development of the technical infrastructure and of this report. To support ongoing engagement with international experts, some international data linkage models have set up scientific expert groups to advise on the operations and development of their systems (56, 161). Additionally, groups in Ireland that have expertise in the areas of record linkage and safe havens can contribute learnings and advice for a national DASSL service. These include the CSO and the Health Identifier Service who oversee the IHI in the HSE.

15 Benefits and risks of a national DASSL service

Establishing a national DASSL service would support a number of different uses and result in many potential societal benefits. However, with these benefits come potential risks that need to be identified and mitigated against where possible.

15.1 Benefits

There are many societal benefits that could be driven by a national DASSL service that supports access, sharing, storage, and linkage of health and related data, including, but not limited to, high-quality research and innovation in the public interest.

15.1.1 Quality and expansion of research

Facilitating researcher access to linked, routinely collected data expands the possibilities of research in Ireland, including large-scale whole-population analyses and longitudinal follow-up of individuals using real-life data. While clinical trials are beneficial for testing the efficacy of a drug or other intervention, they are conducted within controlled environments; routinely collected data are critical to determining and testing the impact of healthcare interventions on different types of patients. Additionally, analysis of the whole population – as opposed to just those who hear about, and consent to participate in, a research study – can reduce bias and provide more informative findings, such as for identifying the number of people with a particular disease.

For example, the Secure Anonymised Information Linkage (SAIL) Databank researchers to analyse linked datasets and identify the number of people living with Idiopathic Intracranial Hypertension in Wales and their characteristics which was previously unknown (162). Routinely collected data is also be used to validate patient registers (162) or subjective reports by participants (163) and expand on the information available. Finally, a national DASSL service would enable Ireland to collaborate with EU and other international organisations in order to expand the possibilities of health research internationally.

15.1.2 Population health and well-being

Research findings from linked health and related datasets can improve the delivery and planning of healthcare, with the overall aim of benefiting the health and well-being of the population. Additionally, the benefits and adverse effects of drugs and other healthcare interventions can be followed up longitudinally, as demonstrated in CS#1 and CS#2. An example from England shows how researchers linked data from primary care, hospital admissions, and death registries for 4.7 million inhabitants and identified that diabetes medications reduced the risk of cardiovascular mortality and hospitalisation for heart failure (161). It is critical, however, that the findings produced by a national DASSL service are transparent and openly shared with healthcare providers and policy-makers.

15.1.3 Data-driven policy and guideline decisions

A national DASSL service and more effective use of health data in Ireland would support data-driven decisions and evidence-based policy-making. Linked data research is used to identify the number of people with a disease or receiving a particular intervention, to evaluate the effectiveness of a treatment on cost and patient outcomes, and to identify risk factors for a health outcome, all of which can inform policies, guidelines, and service planning. For example, a study on people in Wales who have asthma found that those from deprived areas have worse outcomes and increased risk of death when compared with the rest of the population, which resulted in the provision of public health messaging and intervention within deprived communities to better empower patients to manage their asthma (162). Therefore, linked data research can contribute to health technology assessments and more efficacious delivery of healthcare. Additionally, a national DASSL service would help Ireland benefit from the cross-border data sharing initiatives in the EU and beyond, and maintain alignment with these countries in order to assist in global emergencies such as the recent COVID-19 pandemic.

15.1.4 Resources and support for data controllers

Provision of this technical infrastructure, along with clear governance and approvals, should reduce the resources required by data controllers in the provision of data to researchers and for other secondary use cases. Upcoming legislation (e.g. the EU Data Governance Act, AI Act, and the proposed EHDS regulation) will require data controllers to support researcher access to pseudonymised health data via safe havens or secure processing environments.

To avoid duplication of work and resources, a national DASSL service could support this requirement of data controllers. Additionally, data providers could benefit from the infrastructure in terms of linking their own data with other data of interest; for example, for following up on patients on registers, contributing to improving patient outcomes, and service planning.

15.1.5 Data protection and security

Data are currently shared in an ad hoc manner for many research purposes, and the technical and organisational security measures surrounding these methods can differ. The development of a national DASSL service would benefit data protection and ensure a consistently secure mechanism for sharing health data in the public interest. This would help ensure that all access to health data is compliant with data protection principles (including the data minimisation principle) and clear approvals processes.

15.1.6 Efficient user access

A national DASSL service could streamline resources for reviewing data access requests, providing data, and ensuring secure environments for accessing data. A single point of access for researchers with consistent application forms, but with stringent protocols and safeguards around data security and confidentiality, should also make the process of gaining access to data more efficient and less resource-intensive, problems which have often led to the delay and abandonment of research projects in the past.

15.1.7 Economic growth

There is also the potential for a national DASSL service to support economic growth. First, a national DASSL service could improve the competitiveness of Irish researchers' applications for international funding. For example, SAIL in Wales has contributed to GB£48 million in research income being secured (162). Data-driven healthcare policies and improved patient health outcomes have the potential to contribute to reducing disabilities and sickness in the long term, but can also help streamline health service funding, resulting in a more cost-effective healthcare service. Second, many researchers develop and test products while conducting their research projects. The outputs from these research projects often go on to become viable businesses and contribute to economic growth. A national DASSL service could also support the proposed European Data Act and the sharing of data between businesses and government. While private access to a national DASSL service would require additional considerations regarding governance procedures and public trust, it would result in many benefits for the public, including the development and evaluation of drugs and medical devices.

15.2 Risks

There are a number of risks inherent in rolling out a national DASSL service. However, these can all be mitigated, and should be considered from the outset.

15.2.1 Poor-quality data and research

If the data being analysed are of poor quality or are not fit for the purpose of research, the findings will also be of poor quality. As discussed in Section 7.2, the completeness, reliability, validity, and consistency of the data collected will impact on the results generated from those data. Therefore, researchers should be aware of the quality of the data fields that they are using from the metadata and the RSU. Additionally, the data controllers should be supported and resourced in order to both improve the quality of their data where required and to report on the quality of the data for research. Improving data quality must be a key priority from the beginning. Additionally, inaccurate interpretation of data can occur where there is a lack of understanding of the data and metadata made available to researchers. Data controllers have the intrinsic knowledge of the datasets, and close interaction and collaboration between the data controllers and RSU/researchers will be required, especially in the initial stages of a national DASSL service. Duplication of data can also risk errors, or, if the dataset is not closed and is still being edited by the data controller, there is a risk that the duplicated version becomes inaccurate, meaning that the results cannot be replicated. Therefore, it is recommended that only closed and clean cuts of data are released by the data controller, where possible.

15.2.2 Replication of bias and marginalisation in policies and service planning

If specific cohorts of individuals are not linked across datasets, they may be omitted from the interpretation of those data, which could impact on data-driven policies and planning. This can occur when a specific cohort of individuals presents less often for healthcare (as opposed to requiring less healthcare), and where poor linkage occurs in specific locations or for specific cohorts due to poor-quality data collection and/or dependency on dynamic variables that may change more for some individuals than for others (e.g. people experiencing homelessness, asylum seekers, the Travelling Community).

Replication of marginalisation or discrimination within policy is a risk of using routinely collected data without appropriate assessment, understanding, and interpretation of the data (164). As the researcher often does not have access to the personally identifiable data which would allow review of data subjects who are not linked and their sociodemographics, it is important that the DLU assesses the linkage quality and provides this information to the researcher where possible. This also highlights the importance of linking health data with other related data, including housing, education, and criminal justice system data.

15.2.3 Lack of public and patient trust and support

There are many examples of failed health data projects due to a lack of public trust and awareness of the project. In England, the care.data programme aimed to develop a national database of patients' medical records spanning primary and secondary care, but the project failed to adequately explain the benefits of data sharing and win the public trust (165). Concerns over lack of clarity regarding commercial access and the opt-out consent procedure contributed to this failure. Unfortunately, the recent cyberattack on the HSE in Ireland may have impacted on public trust in how their data are protected. However, international data linkage models have garnered public support and trust via a number of mechanisms, including public advisory committees, inclusion of PPI members on research review panels, transparent public communication and engagement policies, stakeholder discussions through public consultations, and ongoing involvement in relation to expansion in the scope of the models, including AI and private access.

15.2.4 Lack of data controller trust and engagement

The success of a national DASSL service in terms of its contribution to public benefits, policy-making, research, and economic growth will depend on data controllers sharing data in an effective and efficient manner. Data controllers' trust in the security and governance procedures and measures, as well as their involvement in the development and oversight of the DASSL service, can also be critical to a centralised model for the data within a national DASSL solution. In order to build data controllers' trust and engagement, they must be involved in the development of a national DASSL service and its policies from the outset, and they need the resources to prepare and share their data. While new legislation mandating the sharing of data may provide a lawful basis and requirement to do this, data controllers' trust and engagement is crucial. Mechanisms used to build data controllers' trust include ongoing engagement and feedback, participation in governance and advisory boards, and involvement in the approvals process for projects that would use the data that they control.

15.2.5 Security and privacy risks

Personal health data are sensitive data that need to be protected with appropriate technical and organisational measures. However, they are always at risk of privacy and security breaches, either maliciously (e.g. a cyberattack) or unknowingly (e.g. by staff or data users). The security measures employed within the DASSL PoC, as covered in Section 9, will reduce the risks of these attacks tenfold. These include use of the separation principle, firewalls, and locked down environments. Overall, the benefits and risks of sharing data will need to be weighed up for each individual project, but the national DASSL service would likely provide a more secure and safe mechanism for linked data projects than what currently exists in Ireland.

15.2.6 Recruitment and retention of skilled staff

Operation of a national DASSL service will require highly skilled staff for managing the infrastructure and the data. There is a huge demand for these skills, and recruitment and retention of staff will require competitive compensation and benefits packages. Additionally, much of the expertise required by the RSU is held by the data controllers. Therefore, consideration needs to be given as to whether staff from the data controllers need to be seconded, or whether the RSU needs to work very closely with the data controllers in order to build this expertise internally within a national DASSL service.

16 National roll-out: Recommendations and resourcing

Development of the DASSL PoC technical infrastructure has provided important insights and considerations for what a national DASSL service to enable research would constitute and require.

The practical experience of designing the PoC systems, constructing relevant case studies, generating synthetic datasets, engaging with stakeholders and experts, etc. has revealed many of the issues, challenges, and lessons that are relevant for the roll-out of a national DASSL service. The following Sections 16.1–16.4 provide an overview of both the technical and non-technical considerations that are key to the roll-out of the proposed national DASSL service and infrastructure.

16.1 Governance and data requirements

From discussions with stakeholders nationally and international peers who have had experience in developing and operating data trusts, the non-technical components will be critical to the success of a national DASSL service, and in many ways will be more important than the technical components. Many of these requirements can also be put in place prior to the development of the technical infrastructure. The following is an overview of the key requirements for the governance and data components of a national DASSL service:

- A clear, lawful basis for the sharing and linking of health and related data for research purposes and for other secondary uses
- Establishment of a governance board and policies for the development and operation of a national DASSL service
- Establishment and/or identification of existing appropriate project approval boards (e.g. RSU, NREC, HRCDC, and IGRP) for each type of secondary use case, with streamlining of applications and forms where possible

- Stakeholder and expert involvement from the outset and on an ongoing basis via advisory boards, surveys, interviews, etc.
- Requirement of data sharing or access agreements by all users, along with data protection training for researchers and staff
- A single updated and maintained standardised metadata catalogue describing all of the health and related datasets available for access via a national DASSL service
- Intrinsic knowledge of the dataset and data collected in order to enable combination of datasets
- A population spine for record linkage, and
- Incentives for datasets to start collecting personal identifiers and sharing data with a national DASSL service.

16.2 Technical requirements

In planning for the roll-out of a national DASSL service, key technical and associated requirements will need to be identified. Many of these requirements were identified during the PoC project, and these are discussed below. While the DASSL PoC deployed specific platforms and software, these are not necessarily what should be deployed in a national roll-out (especially in a rapidly evolving environment), but a national DASSL infrastructure should include components that satisfy the criteria that are guided by this PoC. Sections 16.2.1–16.2.8 describe these technical criteria required for the main functions of a DASSL infrastructure.

16.2.1 Infrastructure requirements

Procurement of the infrastructure should take place after the roll-out team has been established and a governance model for the DASSL solution starts to take shape, in consultation with key stakeholders such as data controllers. A cloud-based technology is recommended, as it would provide a great deal of flexibility and scalability, and this would benefit a national DASSL service that would need to quickly provide resources on an on-demand basis and cater for the variable number and needs of research projects. It is also recommended that the procurement process remain open to using either a public or private cloud, and this should be determined based on the best solution that addresses the technical requirements. The advantages and disadvantages of public and private clouds are discussed in Sections 16.2.2 and 16.2.3.

16.2.2 Public cloud

The use of public clouds for services on a national scale would typically involve a major multinational commercial cloud provider, such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud, etc. The cloud offerings from these companies offer a high degree of flexibility in terms of resource provisioning – one could maintain a baseline level of resources and scale up when needed, e.g. when the number of projects or computing/data requirements increase. The cost model for a public cloud setup would involve some degree of 'pay-as-you-go' mechanism for sudden bursts of demand, which typically lowers overall cost compared with a private cloud. Another advantage of a public cloud is the rich ecosystem of native tools/applications/services available which can be readily procured for some of the key DASSL functions (e.g. databases, identity management, configuration of virtual machines (VMs)).

Finally, utilising a public cloud solution should reduce the level of staffing and expertise required of the DASSL team, who would otherwise be needed to maintain the hardware and some of the software infrastructure. The main concern with the use of a public cloud has to do with data sovereignty, i.e. the lack of physical control over the hardware infrastructure on which sensitive national data are being held. Public perception of this issue must also be taken into account. However, it must be noted that public clouds can offer solutions that provide additional reassurances for the storage of sensitive data, e.g. that such data are held in infrastructure that is physically located in Ireland. Moreover, some national government departments and agencies already have existing public cloud solutions in place which handle sensitive data.

16.2.3 Private cloud

This involves deployment of cloud technology (e.g. OpenStack) on top of on-premises hardware infrastructure, ensuring physical control of the entire infrastructure, which directly addresses concerns regarding data sovereignty. This facilitates more fine-grained control over the infrastructure, as well as the development of in-house expertise and knowledge of all the internal system components. The use of a private cloud on on-premises hardware resources would require additional expertise to maintain the systems and software services. The level of scalability is somewhat reduced, as upgrading to larger resources requires additional planning for installation of new hardware where necessary. The cost of procuring and maintaining a private cloud infrastructure would also be relatively higher, due to the need for fixed capital investments (which remain the same regardless of utilisation) and staffing requirements.

16.2.4 Security requirements

Technical security elements will be required for a national DASSL service, and the following highlights some of the major criteria that should be addressed when considering a national infrastructure:

- The entire DASSL infrastructure should be protected by firewall software/hardware solutions that restrict network traffic to between authorised machines and user roles only. Firewalls should also be configured for each host machine. It is also recommended that web application firewalls be put in place to protect specific web applications that are deployed within the infrastructure.
- At a minimum, access to the DASSL infrastructure must be mediated via a virtual private network (VPN) and end-to-end encryption technology, in combination with two-factor authentication (2FA). This will require an identity management service to handle user roles and credentials.
- All DASSL systems must implement measures to reduce the risk of malicious code being installed: antivirus/antimalware scanners and rootkit hunters must be run regularly on these systems and upon changes/updates to these systems that involve external applications/data, e.g. when installing applications for research project VMs.
- All access to the DASSL infrastructure must be logged. All internal data access and processing transactions must also be logged.
- An effective means to provide and control graphical user interface desktop access to different systems, based on user roles, is the deployment of a virtual desktop infrastructure (VDI) server (e.g. the open-source software Apache Guacamole was used for the PoC, but other remote desktop tools/services exist that provide similar functionalities). Each user gains access to a desktop environment (provided for their role/research project) via this server once connected to the VPN.

16.2.5 Data ingress (from data providers): Technical criteria

This section sets out the technical criteria for the ingress of data from the data providers into the DASSL environment.

16.2.5.1 Hardware

- The data management platform to receive incoming data from data providers must separate personally identifiable data (accessible only by the DLU) and content data (accessible only by the RSU). Two independent platforms could be used to ensure separation of these datasets.

- The resources required for the data management platform are highly dependent on the datasets being ingested. However, typical health datasets (excluding images and genomics data) are in the order of megabytes or gigabytes (GB) in size. Taking into account multiple datasets from research projects, about 50–100 terabytes (TB) of storage should provide sufficient capacity for the initial DASSL roll-out, but this may need to be expanded as datasets grow in size.

16.2.5.2 Software

- While Nextcloud was used in the PoC, alternative platforms with similar features are readily available. The chosen platform will interact with the identity management service to define role-based access to different datasets/partitions and will have relevant safeguards in place to prevent unauthorised access.
- Ingress of data from data providers should be conducted via secure connections that support end-to-end encryption along with 2FA. The review of current best practices and engagement with stakeholders have indicated that a third-party secure file transfer service would provide the best balance between user friendliness and security, which is needed in order to cater for a diverse set of data providers with variable technical know-how.

16.2.6 DLU: Technical criteria

The following sets out the technical criteria in order for the DLU to perform data linkage across datasets for different research projects.

16.2.6.1 Hardware

- The DLU requires considerable computing resources in order to conduct probabilistic linkages, particularly if it involves large datasets that contain millions of rows of data. Due to their nature, these probabilistic linkage workloads can often be conducted in parallel computationally. Therefore, the DLU should have access to multi-core systems along with ample memory (i.e. random access memory (RAM)) in order to achieve reasonable turnaround times for probabilistic linkages. An initial system with 80–120 cores containing a total of 256 GB of storage should provide a suitable platform for the DLU.

- Apart from the DLU VM, where record linkage is conducted, the DLU should have access to a relational database (e.g. a Structured Query Language (SQL) server such as MySQL or PostgreSQL) where data can be managed per project, and specifically where the putative population spine can reside. This database server should reside on a separate VM with access strictly limited to the DLU. Since the database mainly handles a limited number of primarily text fields pertaining to personal data, a system with 8–16 central processing unit (CPU) cores and 5 TB of storage should suffice for the initial roll-out.

16.2.6.2 Software

- The DLU requires routine statistical software packages (R and RStudio were used in the PoC, but this could be expanded to include other commercial packages) for standard data processing and cleaning. This also includes standard spreadsheet software, such as Microsoft Excel or the open-source LibreOffice.
- For the specialised purpose of conducting record linkages, dedicated tools will be required. The PoC DASSL infrastructure deployed the privacy preserving record linkage (PPRL) software, an R library/package, for this purpose. As discussed in Section 10.7, there is some flexibility with regard to the choice of software being used for record matching – the performance of PPRL and comparable tools are largely similar; hence, the choice of software may come down to the experience of the DLU and further testing based on real Irish datasets.

The hardware requirements for a relational database entail a software solution for which there is no shortage of suitable options, from open-source (e.g. MySQL, PostgreSQL) to commercial (e.g. Microsoft SQL Server, Amazon Aurora) offerings. While the PoC created an Aurora instance for the DLU (due to its integration with the AWS environment), other SQL solutions are equally viable. Some consideration could be given to the support and maintenance of the underlying database software, which is typically included in commercial solutions. Similar support for open-source solutions would have to be carried out by the DASSL infrastructure team, or via a commercial third party whose function would be purely to provide such support and maintenance, while the software remains free and open source.

16.2.7 RSU: Technical criteria

The RSU is involved in a number of different roles, from data preparation and linking (when given keys by the DLU), to dataset provisioning and disclosure control. The following points set out the technical criteria for the RSU to perform these actions for different research projects.

16.2.7.1 Hardware

- Most of the workloads are not expected to be computationally intensive, except for those associated with statistical disclosure control. Therefore, we recommend that the DASSL environment provides two new VMs (rather than one VM, as was implemented in the PoC) to cater for either: (a) data preparation and curation duties, or (b) statistical disclosure control, which would require a more powerful dedicated system.
  - For general data processing, the VM will require approximately 8–16 CPU cores and 32 GB of RAM per RSU staff member.
  - For statistical disclosure control, it is recommended that the VM (in shared use by RSU staff members for the purpose of statistical disclosure control) provide up to about 100 CPU cores and a total of 256 GB of RAM for tasks that will require parallel processing capabilities.
- It is envisaged that the RSU may also take on responsibility for maintaining centralised pseudonymised datasets that are updated on a regular basis with input and permission from the data provider, where data may be readily linked and provisioned once the project is approved (i.e. the data provider does not have to send data for specific projects). These datasets, which tend to be more stable in structure and maintained frequently, should be held in a separate relational database (a SQL server). The size of this database system will depend on the datasets to be held and is expected to grow with time.

16.2.7.2 Software

- The RSU will consist of data scientists and specialists who will require some routine data science tools for data processing and manipulation, e.g. statistical packages such as R and RStudio; Python; and commonly used data science libraries such as pandas. The RSU would also need office suite software (mainly for a graphical user interface to spreadsheets software, but other packages – such as word processing and presentation software – are also convenient to have on these systems).

- Here, the arguments for the chosen relational database are the same as for the DLU's requirements for a SQL server. Many commonly used SQL solutions would be suitable for managing the centralised datasets used by the DASSL service.

16.2.8 Safe haven: Technical criteria

For the purposes of this PoC, researchers are the main end users of the DASSL infrastructure. They will have different needs according to the nature of the research in question. The RSU works with the researchers for each project to establish a specific environment in which to conduct the research (e.g. installation of the required software packages). The following hardware and software options are anticipated for research projects on such a system.

16.2.8.1 Hardware

- The expectation is that most research analyses involving tabular data should not be overly compute-intensive, and that standard workstation specifications (e.g. 8 CPU cores, 16 GB of RAM, 500 GB to 1 TB of storage) should fit the needs of most.
- Using cloud-based technology, the DASSL infrastructure should be flexible in allocating the relevant resources to projects according to their needs; however, there should be standard specifications to guide those who are unsure of their computing/data requirements.

16.2.8.2 Software

- Most research will consist of standard statistical analyses on linked, tabular data that will require different statistical packages, the most commonly used ones being R, SPSS Statistics, Stata, and SAS. It is recommended that such standard packages (some of them commercial) are provided on research VMs. Again, an office suite software will also be required for peripheral work on the data analysis. Finally, it is important to also cater for researchers who may prefer different operating systems (Microsoft Windows or Linux) for conducting their analyses.

Researchers will likely make requests to have specific additional tools or software packages installed on the research analysis VMs. The RSU should work with researchers on a case-by-case basis to facilitate this process, taking care not to compromise those VMs that are hosted within the DASSL environment (e.g. conducting antivirus and antimalware scans post-installation).

16.3 Skills and expertise profiling and requirements

Apart from the technical PoC infrastructure that we have examined, all the systems and environments described above require appropriate personnel with the relevant expertise to ensure the well-managed operation of the DASSL model. In establishing a DASSL service for research purposes, it is envisaged that the following high-level functions will be required, along with the skills profiles of the personnel that are required.

16.3.1 Project management team

It is recommended that a project management team be established to oversee the initial roll-out of a national DASSL service. This team should have a mix of skills representing the key DASSL functional units, as highlighted in Sections 16.3.2–16.3.6, with a sound understanding of the Irish health data landscape; the same team members may potentially go on to seed the different DASSL units, e.g. the RSU. The roll-out team should work with DASSL stakeholders in order to address further non-technical considerations or barriers, e.g. access to primary care data, collection of data that are not currently implemented, and data quality issues with particular datasets.

16.3.2 Infrastructure team

This team oversees the smooth operation of the DASSL hardware and software infrastructure, from setting up the different systems and diagnosing hardware and low-level problems, to system updates, security monitoring, and providing technical support to all users. The team will mainly consist of system administrators (both senior and junior), some of whom will ideally have enough software development and IT operations (DevOps) experience to start building the platform. The team should also include administrators with cybersecurity expertise (i.e. the security function will not be entirely outsourced).

16.3.3 Data linkage unit

The DLU has a specific and important role in carrying out data linkage. This team should be represented by data scientists and statisticians with the relevant know-how to conduct data cleaning and manipulation, ideally with some prior experience of records linkage.

For probabilistic linkage, where linkage performance is reduced due to reduced data quality, the use of an extra clerical review step should be considered in order to augment the linkages with manually determined links aided by software. Clerical review requires additional personnel, perhaps hired on a part-time basis due to the nature of the duties involved.

16.3.4 Research Support Unit

The RSU will play key roles in the proposed DASSL service and should consist of a large team of data scientists and individuals with related experience. It will be critical for the RSU to have in-house expertise on Irish health and related datasets, and prior experience in providing research support. In addition, the RSU is the key liaison with data providers and researchers, and will need to engage with both communities consistently in order to ensure up-to-date knowledge and to support change management (e.g. dataset updates). RSU staff involved in statistical disclosure control will require strong statistical backgrounds in order to ensure proper implementation of the process. The RSU may also need to either accredit researcher training or provide this training, as well as review research project applications.

16.3.5 Operations management and administration

It is envisaged that the proposed DASSL operation and infrastructure will be a distinct unit/entity in providing a national data linkage and research platform service. It is expected that it will be led by a small management team that will oversee the running of the service and will interface with the relevant boards/committees of stakeholders. This management team will also include staff to handle routine human resources (HR), finance, and other administrative duties.

16.3.6 Communications and outreach

It is vital that the DASSL operation is regularly engaged with the public, patient groups, and other stakeholders in order to maintain communication about how national datasets are being used for which purposes, and the results of the research projects enabled by the national DASSL service. This communications and outreach team should consist of both senior and junior communication personnel with prior experience in public relations and communications.

16.4 National DASSL service costing

In considering the national roll-out of a DASSL service for research, this report has highlighted the level of technical and personnel resources required in order to create a minimum threshold environment for operating an infrastructure that would cater for 20 or more research projects annually. While the initial phase of rolling out this service will likely see a small number of pilot projects, estimated costs for such an infrastructure must take into account the front-loaded investments that would cater for subsequent ramp-up in the use of the infrastructure, as well as obtaining the critical mass of staff and expertise necessary to establish the service and provide a reasonable level of support to researchers. It is also important to note that the estimated costs of investment in infrastructure and personnel are based on the experience of operating national services for researchers in Ireland (such as the national high-performance computing (HPC) service operated by ICHEC) and on consultations with international peers operating similar services. Finally, the estimates are best treated as guidelines on the scale of funding required, rather than as formal quotations from providers. Table 6 outlines the anticipated personnel costs associated with the different teams within a DASSL Research Data Trust (RDT), as described in this report.

Table 6 Cost estimates of the personnel required to establish and support a national DASSL operation for research purposes

Personnel category	Number of full-time equivalent employees	Cost estimate (per annum)
Management, administration, and finance; roll-out project manager; communications	3.5	€440,000
RSU (junior and senior data scientists and statisticians)	7.0	€594,000
DLU (senior technical staff and part-time clerical review staff)	4.0	€304,000
Infrastructure (senior system administrators and DevOps)	4.0	€370,000
Total per annum	**18.5**	**€1,708,000**

The cost of operating the DASSL infrastructure will depend on whether it will be hosted by on-premises hardware (private cloud) or on a public cloud solution. The estimates also assume that the infrastructure will be constructed from scratch, i.e. assuming that it will not take advantage of parallel investments or existing infrastructure where synergy and cost savings could be achieved.

Regardless of whether a private or public cloud solution is used, the following commercial/third-party services costs (Table 7) are anticipated, which will be necessary for the smooth operation and security of a DASSL operation.

Table 7 Estimates of commercial/third-party service costs for rolling out a national DASSL service

Commercial/third-party services	Cost estimate (per annum)
Third-party secure file transfer service	€20,000
Cybersecurity services (antivirus and antimalware; penetration testing; security audits; intrusion detection and prevention services)	€100,000
Third-party software licensing, e.g. for statistical packages (mainly for researchers; potentially for statistical disclosure control)	€25,000
Total per annum	**€145,000**

In relation to the hardware infrastructure on a public cloud, cost estimates were derived from standard public cloud charges for establishing and sustaining the main systems that would cater for the projected technical requirements of the various DASSL units, assuming support for approximately 20 research projects per annum (15 with low computing resource requirements, and with the remaining 5 requiring larger computing resources). Table 8 provides an indication of the main costs for maintaining a public cloud infrastructure.

Table 8 Infrastructure cost estimates for a DASSL environment in the public cloud		
Infrastructure component	**Units**	**Cost estimate (per annum)**
DASSL internal VMs and storage		
Data management VM (8 cores, 64 GB RAM)	1	€2,400
Data curation VM (48 cores, 384 GB RAM)	1	€14,000
Statistical disclosure control VM (96 cores, 384 GB RAM)	1	€24,000
Data linkage VM (96 cores, 384 GB RAM)	1	€24,000
SQL server	2	€2,800
200 TB storage (shared by different VMs)	1	€40,000
Research project VMs and storage		
Data analysis VM (Windows operating system, 8 cores, 32 GB RAM, 100 GB storage, about 10% utilisation)	15	€8,100
Data analysis VM (Windows operating system, 32 cores, 128 GB RAM, 500 GB storage, about 15% utilisation)	5	€23,200
Total per annum		**€138,500**

The use of a private cloud to implement the DASSL environment assumes the procurement of new hardware as a capital investment, which would subsequently be amortised over a number of years. This is generally 3–4 years for a compute cluster of an appropriate size (with approximately 500 CPU cores and 2 TB of combined RAM) that would cater for the needs of the DASSL environment and support about 20 projects per annum; such a compute cluster would provide sufficient resources for the same systems as those specified for the public cloud above.

The operational costs associated with this hardware (Table 9) include costs that are not incurred by the public cloud option, such as the electricity costs required for cooling, additional costs associated with maintenance, and support costs for the hosting hardware and to sustain an operational OpenStack server.

Table 9 Infrastructure cost estimates for a DASSL environment in a private cloud, using on-premises hardware	
Infrastructure component	**Cost estimate (per annum)**
OpenStack server (about 500 cores, 2 TB RAM aggregate, 200 TB storage) (capital investment of €150,000, amortised over 3 years)	€50,000
Electricity (mainly for cooling)	€15,000
Managed maintenance and support costs	€150,000
Total per annum	**€215,000**

In summary, the roll-out of a national DASSL service for research will entail an investment of approximately €2 million per annum. While the infrastructure costs are significant, much of the investment and the real value of the service will lie in the personnel, who will be key to the success of developing the services around the infrastructure.

17 Conclusion and next steps

The DASSL PoC project has provided critical insights into the requirements and considerations for the roll-out of a national DASSL service.

The next steps in planning and subsequently implementing a national DASSL service should now be considered under the following key components: governance and legislation; stakeholder involvement and engagement; staffing; health and related data; technical infrastructure; and funding and resourcing.

17.1 Governance and legislation

Once enacted, the Health Information Bill proposed by the Department of Health should provide a lawful basis for a named entity to receive, link, and provide access to health and related social data in Ireland. This would provide the legal framework for data controllers across different sectors to share data with the named entity responsible for a national DASSL service. A governance board will also need to be established to oversee this service, along with the development of policies and standard operating procedures (SoPs), prior to the roll-out of a national DASSL service.

17.2 Stakeholder involvement and engagement

Development of these policies and SoPs should involve engagement with data controllers and other key stakeholders. Advisory boards involving the public, data controllers, and users of the system can be established immediately, and a large-scale public consultation process and awareness programme is also recommended in advance of the development of a national DASSL service. Engagement of key stakeholders from the outset will help mitigate issues early on in the process, and ongoing engagement and involvement will be critical for a national DASSL service in order to respond to the expected changes and advances in technology and data availability on an ongoing basis. More learnings can also be gained from speaking to international and national experts in the areas of data sharing and linking.

17.3 Staffing

To begin the process of engaging with stakeholders and developing policies, an initial team will need to be established. Once the technical infrastructure has been procured, additional teams will also need to be developed, including teams of system administrators, data scientists, and statisticians. It is expected that the initial phase of development will have minimal project applications and that much of the work will involve getting familiar with the data and developing expertise.

17.4 Health and related data

The benefits derived from a national DASSL service will depend on the data available. Prior to the availability of the technical infrastructure and governance procedures, initial work can be done in relation to health and related data. Development of a metadata catalogue, along with discussions with data controllers in relation to the data, will be critical prior to undertaking research projects within a national DASSL service.

17.5 Technical infrastructure

Based on the findings in this report and the PoC technical infrastructure, the requirements for a national DASSL service should be gathered from key stakeholders. This will require decisions to be made regarding whether the data received from data controllers can be stored on an ongoing basis (i.e. using a centralised/hybrid model); security and risk management policies in relation to the data flow, balancing security with usability (e.g. should content data and personal identifiers be uploaded via different platforms?); and the capability to perform more advanced analytics that require greater computing power. Once the requirements are identified, procurement of the technical infrastructure will need to be undertaken, with consideration of both public and private cloud providers and their related advantages and disadvantages. The national system will then need to be built and will require ongoing development in line with advances in technology, research, and data.

17.6 Funding and resourcing

Finally, a national DASSL solution will require adequate ongoing funding in order to undertake the roles and responsibilities outlined in upcoming Health Information Bill in Ireland and EHDS regulation. Planning for these costs will need to occur well in advance, and the resourcing and funding required will largely depend on decisions made during the next steps that have just been described.

18 Bibliography

1. Han A, Isaacson A, Muennig P. The promise of big data for precision population health management in the US. Public Health. 2020;185:110–6.
2. Lawrence NR, Bradley SH. Big data and the NHS – we have the technology, but we need patient and professional engagement. Future Healthc J. 2018;5(3):229–30.
3. Lovett R, Fisher J, Al-Yaman F, Dance P, Vally H. A review of Australian health privacy regulation regarding the use and disclosure of identified data to conduct data linkage. Aust N Z J Public Health. 2008;32(3):282–5.
4. Young A, Flack F. Recent trends in the use of linked data in Australia. Australian Health Review. 2018;42(5):584–90.
5. The Australian Government Linked Data Working Group. Submission 46 to the Productivity Commission Inquiry - Data Availability and Use. Queensland: Local Government Association of Queensland; 2016.
6. Moran R. Proposals for an Enabling Data Environment for Health and Related Research in Ireland. Dublin: Health Research Board; 2016.
7. Paprica AP, Sutherland E, Smith A, Brudno M, Cartagena RG, Crichlow M, et al. Essential requirements for establishing and operating data trusts: Practical guidance co-developed by representatives from fifteen Canadian organizations and initiatives. Int J Popul Data Sci. 2020;5(1).
8. Open Data Institute. Data Trusts: How do we unlock the value of data while preventing harmful impacts? [Internet]. 2019 [cited 1 Nov 2021]. Available from https://theodi.org/wp-content/uploads/2019/04/ODI-Data-Trusts-B3-Leaflet-web-2.pdf
9. The Secure Anonymised Information Linkage (SAIL) Databank. The SAIL Databank: 10 Years of Spearheading Data Privacy and Research Utility, 2017-2017. Swansea: Swansea University; 2017.
10. Jones KH, Ford DV, Thompson S, Lyons RA. A profile of the SAIL databank on the UK secure research platform. Int J Popul Data Sci. 2019;4(2):1134.
11. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Serv Res. 2012;12:480.

12. The Scottish Government. Charter for Safe Havens in Scotland: Handling Unconsented Data from National Health Service Patient Records to Support Research and Statistics. 2015.

13. NHS Scotland. NHS Research Scotland: Health Data Research (HDR-UK) [Internet]. [n.d.] [cited 13 Feb 2020]. Available from: https://www.nhsresearchscotland.org.uk/research-in-scotland/data/health-informatics

14. Public Health Scotland. ISD Scotland: eDRIS Products and Services [Internet]. [n.d.] [cited 1 Apr 2020]. Available from: https://www.isdscotland.org/Products-and-Services/eDRIS/FAQ-eDRIS/#a1

15. National Health Service Scotland. Public Benefit and Privacy Panel for Health and Social Care. [Internet]. [n.d.] [cited 24 Apr 2020]. Available from: https://www.informationgovernance.scot.nhs.uk/pbpphsc/

16. Scottish Informatics Programme. A Blueprint for Health Records Research in Scotland. 2012.

17. The electronic Data Research and Innovation Service (eDRIS) team. Researcher guide: requesting outputs from Safe Haven and disclosure control [Internet]. 2018 [cited: 1 Apr 2020]. Available from: https://www.isdscotland.org/products-and-services/edris/_docs/Researcher-guide-to-disclosure-control-V1-2.pdf

18. University of Dundee. Health Informatics Centre [Internet]. [n.d.] [cited 18 Mar 2020]. Available from: https://www.dundee.ac.uk/hic/

19. Nind T, Galloway J, McAllister G, Scobbie D, Bonney W, Hall C, et al. The research data management platform (RDMP): A novel, process driven, open-source tool for the management of longitudinal cohorts of clinical data. Gigascience. 2018;7(7):giy060.

20. The Secure Anonymised Information Linkage (SAIL) Databank. SAIL Databank [Internet]. [n.d.] [cited 02 Apr 2020]. Available from: https://saildatabank.com/

21. Jones KH, Ford DV. Population data science: advancing the safe use of population data for public benefit. Epidemiol Health. 2018;40:e2018061.

22. Health Information Research Unit. Secure Anonymised Information Linkage (SAIL) Primary Care Practice Sign Up Pack [Internet]. 2012 [cited 2 Apr 2020]. Available from: http://www.wales.nhs.uk/sites3/documents/952/Maximising%20Use%20of%20Routine%20Data%20for%20Research%202012%2012%2003%20SAIL%20GP%20Sign%20Up%20pack-%20FINAL%20VERSION.pdf

23. Goldacre B, Morley J, Hamilton N. Better, Broader, Safer: Using Health Data for Research and Analysis. London: Department for Health and Social Care; 2022.

24. OpenSAFELY Home. [Internet]. [n.d.] [cited 30 Jun 2022] Available from: https://www.opensafely.org/"

25. Business Services Organisation. Honest Broker Service [Internet]. [n.d.] [cited 26 Aug 2021]. Available from: https://hscbusiness.hscni.net/services/2454.htm

26. Suomen itsenäisyyden juhlarahasto (Sitra). A Finnish model for the secure and effective use of data. Helsinki: Suomen itsenäisyyden juhlarahasto; 2019.

27. Findata. Services for customers [Internet]. 2021 [cited 11 Mar 2022]. Available from: https://www.findata.fi/en/services/services-for-customers/

28. Seppänen J. Findata: Secondary use of Finnish Social and Health Data - a new Act and Data Permit Authority. Presentation at: Healthcare Information and Management Systems Society (HIMSS) Europe 2019; 11-13 June 2019; Helsinki, Finland.

29. Arrêté du 29 novembre 2019 portant approbation d'un avenant à la convention constitutive du groupement d'intérêt public « Institut national des données de santé » portant création du groupement d'intérêt public « Plateforme des données de santé »

30. Villani C, Schoenauer M, Bonnet Y, Berthet C, Cornut AC, Levin F, et al. For a meaningful Artificial Intelligence: Towards a French and European strategy. 2018.

31. Health Data Hub. Health Data Hub [Internet]. [n.d.] [cited 30 Jun 2021]. Available from: https://www.health-data-hub.fr/

32. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare. Yearb Med Inform. 2019;28(1):195–202.

33. OpusLine. Health Data Hub: an ambitious French initiative for tomorrow's health [Internet]. 2019 [cited 20 Jun 2021]. Available from: https://opusline.fr/health-data-hub-an-ambitious-french-initiative-for-tomorrows-health/

34. Dillet R. France's Health Data hub to move to European cloud infrastcuture to avoid EU-US data transfers [Internet]. 2020 [cited 24 Aug 2021]. Available from: https://techcrunch.com/2020/10/12/frances-health-data-hub-to-move-to-european-cloud-infrastructure-to-avoid-eu-us-data-transfers

35. Guesdon M, Benzenine E, Gadouche K, Quantin C. Securizing data linkage in French public statistics. BMC Med Inform Decis Mak. 2016;16(1):129.

36. Swedish Research Council. Registerforskning.se [Internet]. 2021 [cited 24 Aug 2021]. Available from: https://www.registerforskning.se/sv/

37. Direktoratet for eHelse. Helsedataprogrammet (avsluttet 31.12.2021) [Internet]. 2021 [cited 30 Jun 2021]. Available from: https://www.ehelse.no/programmer/helsedataprogrammet

38. Health RI. About Health RI [Internet]. 2021 [cited 24 Aug 2021]. Available from: https://www.health-ri.nl/about-health-ri

39. Personal Health Train. The Personal Health Train Network [Internet]. 2021 [cited 29 Jan 2021]. Available from: https://pht.health-ri.nl/

40. PHRN. Population Health Research Network [Internet]. 2020 [cited 9 Oct 2020]. Available from: https://www.phrn.org.au/

41. The Centre for Data Linkage at Curtin University. LinXmart – Record Linkage and Management Software [Internet]. [n.d.] [cited 8 Jan 2021]. Available from: https://linxmart.com.au/

42. Australian Law Reform Commission. For Your Information: Australian Privacy Law And Practice (ALRC Report 108). 2008.

43. Tan KM, Flack FS, Bear NL, Allen JA. An evaluation of a data linkage training workshop for research ethics committees. BMC Med Ethics. 2015;16:13.

44. Australian Institute of Health and Welfare. About our data [Internet]. 2022 [cited 30 Aug 2022]. Available from: https://www.aihw.gov.au/about-our-data

45. Centre for Health Record Linkage. CHeReL [Internet]. [n.d.] [cited 3 Aug 2021]. Available from: https://www.cherel.org.au/about-us

46. Data Linkage Western Australia. Sample Selections [Internet]. 2020 [cited 16 Mar 2020]. Available from: https://www.datalinkage-wa.org.au/dlb-services/sample-selections/

47. Mitchell RJ, Cameron CM, McClure RJ, Williamson AM. Data linkage capabilities in Australia: Practical issues identified by a Population Health Research Network 'Proof of Concept project'. Aust N Z J Public Health. 2015;39(4):319–25.

48. Centre of Big Data Research in Health. International Collaborative Session: Secure data analysis environments What are the models? What do we call them? In: The International Population Data Linkage Conference 2018; 12-14 Sep 2018; Banff, Alberta, Canada.

49. University of New South Wales Sydney. ERICA – E-Research Institutional Cloud Architecture [Internet]. [n.d.] [cited 12 Jan 2021]. Available from: https://research.unsw.edu.au/erica

50. Brook EL, Rosman DL, Holman CDAJ. Public good through data linkage: Measuring research outputs from the Western Australian Data Linkage System. Aust N Z J Public Health. 2008;32(1):19–23.

51. Eitelhuber T, Davis G, Rosman D, Glauert R. Western Australia unveils advances in linked data delivery systems. Aust N Z J Public Health. 2014;38(4):397–8.

52. Eitelhuber T, Davis G. The custodian administered research extract server: "improving the pipeline" in linked data delivery systems. Health Inf Sci Syst. 2014;2:6.

53. Eitelhuber TW, Thackray J, Hodges S, Alan J. Fit for Purpose: Western Australia unveils its new data linkage system. Int J Popul Data Sci. 2018;3(3):435.

54. New South Wales Government. Lumos Technical Information Summary [Internet] 2021 [cited: 3 Jul 2022]. Available from: https://swsphn.com.au/wp-content/uploads/2022/02/LumosTechnicalInformationSummary.pdf

55. Hertzman CP, Meagher N, McGrail KM. Privacy by Design at Population Data BC: A case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest. J Am Med Inform Assoc. 2013;20(1):25–8.

56. ICES. About ICES Research [Internet]. [n.d.] [cited 19 Mar 2020]. Available from: https://www.ices.on.ca/Research/About-ICES-Research

57. Smith M, McGrail K, Schull M, Katz A, McDonald T, Paprica PA, et al. Pan-Canadian Real-World Health Data Network: Building a National Data Platform. Int J Popul Data Sci. 2018;3(4):984.

58. Paprica AP, Sutherland E, Smith A, Brudno M, Cartagena RG, Crichlow M, et al. Essential requirements for establishing and operating data trusts: Practical guidance co-developed by representatives from fifteen Canadian organizations and initiatives. Int J Popul Data Sci. 2020;5(1):1353.

59. Schull M, Azimaee M, Marra M, Cartagena R, Vermeulen M, Ho M, et al. ICES: Data, Discovery, Better Health. Int J Popul Data Sci. 2020;4(2):1135.

60. Ark T, Kesselring S, Hills B, McGrail K. Population Data BC: Supporting population data science in British Columbia. Int J Popul Data Sci. 2020;4(2):1133.

61. Wong S, Schuckel V, Thompson S, Ford D, Lyons R, Hier R. British Columbia's Health Data Platform: Unleashing the Power of a Data Environment Commons for Health and Health System Improvement. Int J Popul Data Sci. 2020;5(5):1477.

62. University of Manitoba. Manitoba Centre for Health Policy [Internet]. 2020 [cited 2 Nov 2020]. Available from: http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/about.html

63. Katz A, Enns J, Smith M, Burchill C, Turner K, Towns D. Population data centre profile: The Manitoba centre for health policy. Int J Popul Data Sci. 2019;4(2):1131.

64. Health Informatics Centre. Working with HIC's data linkage service [Internet]. [n.d.] [cited 2 Nov 2020]. Available from: https://www.dundee.ac.uk/hic/data-service/data-linkage-service

65. Data Protection Commission. Guidance for Organisations Engaging Cloud Service Providers [Internet]. 2019 [cited 2 Nov 2020]. Available from: https://www.dataprotection.ie/en/dpc-guidance/guidance-organisations-engaging-cloud-service-providers

66. Data Protection Commission. Five Steps to Secure Cloud-based Environments [Internet]. [n.d.] [cited 2 Nov 2020]. Available from: https://www.dataprotection.ie/en/dpc-guidance/five-steps-secure-cloud-based-environments

67. Microsoft. Microsoft Trust Center [Internet]. [n.d.] [cited 30 Nov 2020]. Available from: https://www.microsoft.com/en-ie/trust-center

68. Schneider M. Distributed networks of federated secure research data environments - enabling analytics across multiple platforms. Int J Popul Data Sci. 2020;5(5):1600.

69. Houses of the Oireachtas. Committee on the Future of Healthcare debate - Wednesday, 14 Sep 2016 [Internet]. 2016 [cited 22 Nov 2022]. Available from: https://www.oireachtas.ie/en/debates/debate/committee_on_the_future_of_healthcare/2016-09-14/3/

70. Fitzsimons M, Dunleavy B, O'Byrne P, Dunne M, Grimson J, Kalra D, et al. Assessing the quality of epilepsy care with an electronic patient record. Seizure. 2013;22(8):604–10.

71. European Commission. European Open Science Cloud (EOSC) [Internet]. [n.d.] [cited 23 Nov 2020]. Available from: https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en

72. Aarestrup FM, Albeyatti A, Armitage WJ, Auffray C, Augello L, Balling R, et al. Towards a European health research and innovation cloud (HRIC). Genome Med. 2020;12(1):18.

73. Robertson D, Giunchiglia F, Pavis S, Turra E, Bella G, Elliot E, et al. Healthcare data safe havens: towards a logical architecture and experiment automation. J Eng. 2016;2016(11):431–40.

74. Bennett Institute for Applied Data Science, University of Oxford. OpenPrescribing.net [Internet]. 2022 [cited 2 Jun 2022]. Available from: https://openprescribing.net/

75. openEHR Foundation. openEHR Specifications [Internet]. [n.d.] [cited 2 Aug 2022]. Available from: https://specifications.openehr.org/

76. Jarret M, Hills B, Zhao Y, Brown A, Randall S, Boyd J, et al. Evaluating PPRL vs clear text linkage with real-world data. Int J Popul Data Sci. 2020;5(5):1542.

77. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584:430–6.

78. Gaia-X European Association for Data and Cloud AISBL. Gaia-X [Internet]. [n.d.] [cited 23 Nov 2020]. Available from: https://gaia-x.eu/

79. European Commission. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions – Artificial Intelligence for Europe. COM(2018) 273 final. 2018.

80. European Commission. Coordinated Plan on Artificial Intelligence 2021 Review [Internet]. 2021 [cited 2 Nov 2021]. Available from: https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

81. Centre for Information Policy Leadership. Artificial Intelligence and Data Protection – How the GDPR Regulates AI [Internet]. 2020 [cited 2 Nov 2021]. Available from: https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf

82. High-Level Expert Group on AI. Ethics guidelines for trustworthy AI. 2019.

83. European Commission. On Artificial Intelligence – a European approach to excellence and trust. 2020.

84. European Commission. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – a European strategy for data. 2020.

85. European Commission, Directorate-General for Communications Networks, Content and Technology. Shaping Europe's Digital Future. 2020.

86. Government of Ireland. Ireland's National Submission to the Public Consultation on the EU White Paper on Artificial Intelligence. 2020.

87. European Commission. Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence [Internet]. 2021 [cited 3 Aug 2021]. Available from: https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682

88. European Commission. Digital Europe Programme [Internet]. [n.d.] [cited 10 Aug 2021]. Available from: https://digital-strategy.ec.europa.eu/en/activities/digital-programme

89. Government of Ireland. AI – Here for Good: A National Artificial Intelligence Strategy for Ireland. 2021.

90. Sumon Shahriar M. Application of Machine Learning to Streamline Clerical Review in Data Linkage. Int J Popul Data Sci. 2020;5(5):1571.

91. McCradden MD, Sarker T, Paprica PA. Conditionally positive: A qualitative study of public perceptions about using health data for artificial intelligence research. BMJ Open. 2020;10:e039798.

92. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017;46(5):1699–710.

93. De Lusignan S, Navarro R, Chan T, Parry G, Dent-Brown K, Kendrick T. Detecting referral and selection bias by the anonymous linkage of practice, hospital and clinic data using Secure and Private Record Linkage (SAPREL): Case study from the evaluation of the Improved Access to Psychological Therapy (IAPT) service. BMC Med Inform Decis Mak. 2011;11:61.

94. Randall S, Brown A, Ferrante A, Boyd J, Irvine K, Eitelhuber T, et al. Overcoming the Impasse 2: Assessing the Quality of Recent Australian Applications of a Privacy-Preserving Record Linkage Method (PPRL-BLOOM). Int J Popul Data Sci. 2020;5(5):1489.

95. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak. 2009;9:41.

96. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. J Biomed Inform. 2014;50:205–12.

97. Leeming G, Ainsworth J, Clifton DA. Blockchain in health care: hype, trust, and digital health. The Lancet. 2019;393(10190):2476–7.

98. Chukwu E, Garg L. A systematic review of blockchain in healthcare: Frameworks, prototypes, and implementations. IEEE Access. 2020;8:21196–214.

99. Mamo N, Martin GM, Desira M, Ellul B, Ebejer JP. Dwarna: a blockchain solution for dynamic consent in biobanking. Eur J Hum Genet. 2020;28:609–26.

100. Porsdam Mann S, Savulescu J, Ravaud P, Benchoufi M. Blockchain, consent and prosent for medical research. J Med Ethics. 2021;47(4):244–50.

101. Heston T. A Case Study in Blockchain Healthcare Innovation. Int J Curr Res. 2017;9(11):60587–8.

102. Panetta R, Cristofaro L. A closer look at the EU-funded My Health My Data project. Digital Health Legal. November 2017;10–1.

103. Rupasinghe T. Do you need my health data - just ask: using blockchain technology for collaborative patient-centric health care. Int J Popul Data Sci. 2020;5(5):1607.

104. My Health My Data. Why MHMD? [Internet]. 2016 [cited 23 Nov 2020]. Available from: http://www.myhealthmydata.eu/why-mhmd/

105. Sweeney L. k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowledge-Based Syst. 2002;10(5):557–70.

106. Surendra H, Mohan H. S. A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing.Int J Sci Technol Res. 2017;6(3):95-101.

107. Nowok B, Raab GM, Dibben C. synthpop: Bespoke creation of synthetic data in R. J Stat Softw. 2016;74(11):1-26.

108. Oderkirk J. Survey results: National health data infrastructure and governance. OECD Health Working Papers No. 127. 2021.

109. The Population Health Information Research Infrastructure. The PHIRI project [Internet]. [n.d.] [cited 9 Jun 2021]. Available from: https://www.phiri.eu/

110. EHAction. Towards the European Health Data Space Workshop [Internet]. 2020 [cited 30 Jun 2020]. Available from: http://ehaction.eu/towards-the-european-health-data-space-workshop/

111. Towards European Health Data Space. Joint Action Towards the European Health Data Space – TEHDAS [Internet]. 2021 [cited 9 Jun 2021]. Available from: https://tehdas.eu/

112. FinData. How to ensure efficient and secure use of health data beyond borders? France and Finland collaborating to find answers [Internet]. 2021 [cited 20 Jun 2022]. Available from: https://findata.fi/en/news/how-to-ensure-efficient-and-secure-use-of-health-data-beyond-borders-france-and-finland-collaborating-to-find-answers/

113. Irish Social Science Data Archive: Frequently Asked Questions [Internet]. [n.d.] [cited 22 Apr 2020]. Available from: https://www.ucd.ie/issda/help/faq/

114. Seoighe C, Bracken A, Buckley P, Doran P, Green R, Healy S, et al. The future of genomics in Ireland – focus on genomics for health. HRB Open Res. 2020;3:89.

115. European Commission. European '1+ Million Genomes' Initiative [Internet]. 2020 [cited 25 Nov 2020]. Available from: https://ec.europa.eu/digital-single-market/en/european-1-million-genomes-initiative

116. Health Information and Quality Authority. Guidance on a data quality framework for health and social care. 2018.

117. Health Information and Quality Authority. The need to reform Ireland's national health information system. 2021.

118. Health Information and Quality Authority. Catalogue of national health and social care data collections. 2017.

119. Central Statistics Office. Databases [Internet]. [n.d.] [cited 27 Jul 2021]. Available from: https://www.cso.ie/en/databases/

120. Government of Ireland. Public Service Data Catalogue [Internet]. [n.d.] [cited 27 Jul 2021]. Available from: https://datacatalogue.gov.ie/

121. European Commission. DCAT Application Profile for data portals in Europe [Internet]. [n.d.] [cited 27 Jul 2021]. Available from: https://ec.europa.eu/isa2/solutions/dcat-application-profile-data-portals-europe_en

122. Document, Discover and Interoperate (DDI) Alliance. Membership List [Internet]. [n.d.] [cited 4 Mar 2022]. Available from: https://ddialliance.org/ddi-membership

123. Towards European Health Data Space. Identification of relevant standards and data models for semantic harmonization. 2022.

124. Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) Foundation. FHIR [Internet]. [n.d.] [cited 27 Jul 2021]. Available from: https://www.hl7.org/fhir/

125. Health Information and Quality Authority. Guidance on Messaging Standards for Ireland. 2012.

126. Lane JCE, Weaver J, Kostka K, Duarte-Salles T, Abrahao MTF, Alghoul H, et al. Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. Lancet Rheumatol. 2020;2(11):e698–711.

127. Health Service Executive. Standardising Data for The Future: Dataset Specification Management Process. 2022.

128. Health Service Executive. HSE National Health and Social Care Data Dictionary [Internet]. [n.d.] [cited 8 May 2022]. Available from: https://datadictionary.hse.ie/

129. Donenfeld J. WireGuard. 2022. Available from: https://www.wireguard.com/

130. Red Hat. KeyCloak. 2022. Available from: https://www.keycloak.org/

131. Templ M, Kowarik A, Meindl B. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. J Stat Softw. 2015;67(4):1–36.

132. Statistics Netherlands. τ-ARGUS. 2020. Available from: https://research.cbs.nl/casc/tau.htm

133. Central Statistics Office. CSO Best Practice for Statistical Disclosure Control of Tabular Data [Internet]. [n.d.] [cited 8 May 2022]. Available from: https://www.cso.ie/en/media/csoie/aboutus-new/dataforresearchers/CSO_Guidance_on_Tabular_SDC.docx

134. Data Protection Commission. Data Protection Impact Assessments [Internet]. [n.d.] [cited 8 Jun 2021]. Available from: https://www.dataprotection.ie/en/organisations/know-your-obligations/data-protection-impact-assessments

135. Central Statistics Office. Resources for Researchers [Internet]. 2022 [cited 8 Jun 2021]. Available from: https://www.cso.ie/en/aboutus/lgdp/csodatapolicies/dataforresearchers/resourcesforresearchers/

136. Data Protection Commission. Guidance on Anonymisation and Pseudonymisation. 2012.

137. European Network and Information Security Agency. Pseudonymisation techniques and best practices [Internet]. 2019 [cited 6 Jul 2022]. Available from: https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices

138. Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques. 2014.

139. Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, Ford DV, et al. Residential Anonymous Linking Fields (RALFs): A novel information infrastructure to study the interaction between the environment and individuals' health. J Public Health (Oxf). 2009;31(4):582–8.

140. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, et al. GUILD: GUidance for Information about Linking Data sets. Journal of Public Health (Oxf). 2018;40(1):191–8.

141. Statistics New Zealand. Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project. 2014.

142. Schnell R, Rukasz D, Borgs C, Brumme S, Brogden WB, O'Brien T, Lacy S. Package 'PPRL' [Internet]. [n.d.] [cited 6 Jul 2022]. Available from: https://cran.r-project.org/web/packages/PPRL/index.html

143. Karr AF, Taylor MT, West SL, Setoguchi S, Kou TD, Gerhard T, et al. Comparing record linkage software programs and algorithms using real-world data. PLoS One. 2019;14(9):e0221459.

144. Mercure Avocats. Recent developments concerning the French Health Data Hub [Internet]. 2022 [cited 21 Jul 2022]. Available from: https://www.mercure-avocats.com/en/news/news-detail/recent-developments-concerning-the-french-health-data-hub

145. Finnish BioBank Cooperative. Fingenious [Internet]. [n.d.] [cited 21 Jul 2022]. Available from: https://site.fingenious.fi/en/

146. South Australian Genomics Centre. South Australian Genomics Centre [Internet]. [n.d.] [cited 25 Jul 2022]. Available from: https://www.sa-genomics.com.au/index.php

147. DataCebo. The Synthetic Data Vault [Internet]. [n.d.] [cited 21 Jul 2022]. Available from: https://sdv.dev/

148. Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. 2018. Available from: https://doi.org/10.48550/arxiv.1812.04948

149. Badawi A, Elgazzar K. Detecting Coronavirus from Chest X-rays Using Transfer Learning. COVID. 2021;1(1):403–15.

150. Jones KH, Mcnerney CL, Ford DV. Involving consumers in the work of a data linkage research unit. Int J Consum Stud. 2014;38(1):45–51.

151. Irvine K, Hall R, Taylor L. A profile of the centre for health record linkage. Int J Popul Data Sci. 2019;4(2):1142.

152. National Office for Research Ethics Committees. Enabling a trusted national ethics opinion [Internet]. 2020 [cited 1 Dec 2020]. Available from: https://www.nrecoffice.ie/

153. Health Research Charities Ireland. Using data for better health research. 2022.

154. Health Research Consent Declaration Committee. About us [Internet]. 2020 [cited 23 Jun 2020]. Available from: https://hrcdc.ie/about-us/

155. Paprica PA, de Melo MN, Schull MJ. Social licence and the general public's attitudes toward research based on linked administrative health data: a qualitative study. CMAJ Open. 2019;7(1):E40–6.

156. Towards European Health Data Space. Citizens' perception of and engagement with health data secondary use and sharing in Europe – a literature review. 2021.

157. Data Saves Lives. What is Data Saves Lives? [Internet]. 2020 [cited 2 Jul 2020]. Available from: https://datasaveslives.eu/aboutdsl

158. Irish Platform for Patient Organisations, Sciece and Industry. Citizen's Jury on Health Information [Internet]. 2021 [cited 8 Jan 2021]. Available from: https://www.ipposi.ie/our-work/policy/health-information/citizens-jury/

159. Health Information and Quality Authority. Findings from the National Public Engagement on Health Information. 2021.

160. European Commission. Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space. 2022.

161. Health Data Rearch UK. Our Advisory Groups [Internet]. [n.d.] [cited 15 May 2020]. Available from: https://www.hdruk.ac.uk/about/people/our-advisory-groups/

162. Secure Anonymised Information Linkage (SAIL) Databank. SAIL Databank 2020-2021 Annual Report. 2022.

163. Tingay KS, Bandyopadhyay A, Griffiths L, Akbari A, Brophy S, Bedford H, et al. Record linkage to enhance consented cohort and routinely collected health data from a UK birth cohort. Int J Popul Data Sci. 2019;4(1):579.

164. Downs J, Setakis E, Mostafa T, Hayes R, Hotopf M, Ford T, et al. Linking strategies and biases when matching cohorts to the National Pupil Database. Int J Popul Data Sci. 2017;1(1):369.

165. Godlee F. What can we salvage from care.data? BMJ. 2016;354:i3907.

Proof-of-Concept: Technical Prototype for Data Access, Storage, Sharing and Linkage (DASSL) to support research and innovation in Ireland

Case studies

ICHEC
Irish Centre for High-end Computing

Ollscoil na Gaillimhe
University of Galway

Introduction

Conceived by the Health Research Board (HRB) in 2016, the Data Access, Storage, Sharing and Linkage (DASSL) model aims to facilitate data access, sharing, storage, and linkage (1).

The Irish Centre for High-End Computing (ICHEC) – the national high-performance computing centre hosted by the University of Galway – received funding to develop a proof of concept (PoC) technical infrastructure to support the DASSL model and to provide recommendations for the roll-out of a national DASSL solution. The DASSL model and the PoC have been based on international best practice with the overall aim of protecting the privacy of individuals while supporting the use of our health and data resources for public benefit. The *Proof of concept: Technical prototype for Data Access, Storage, Sharing and Linkage (DASSL) to support research and innovation in Ireland* report provides the learnings from the PoC DASSL project in relation to the health data landscape in Ireland and internationally; the processes involved in the operation of a national DASSL solution; and the technical and resourcing requirements for operating this solution. Case studies were developed for the DASSL PoC to test the capabilities of the infrastructure while also demonstrating the potential of linking national health and related datasets in Ireland if a national DASSL solution and governance existed. While these case studies are discussed briefly in the above-mentioned report, further details are described in this case studies annex.

Synthetic data were generated to mimic the characteristics of the health and related datasets in Ireland. In some cases, the synthetic datasets were adapted to allow them to be linked and to demonstrate different types of linkage. Although some of these datasets do not currently exist in Ireland, the DASSL PoC project aimed to demonstrate how the datasets could be linked if they were developed.

Case study development

Five case studies were developed in order to test the infrastructure, and inclusion criteria were set based on the DASSL landscape analysis report, stakeholder engagement, and international examples in order to ensure a broad spectrum of research questions. This included dataset types, linkages, study designs, research question types, and analyses (Table 1). Synthetic data were then generated for each case study based on real national health and related datasets. The R package synthpop (2) and the Python package Synthetic Data Vault (SDV) (3) were used to generate the synthetic tabular data, while StyleGAN (or Stylized Generative Adversarial Network) (4) was used to generate synthetic images.

Table 1 Inclusion criteria for case studies

Datasets	Record linkage	Study designs	Research questions	Analyses	Data management models
Administrative	Linkage types	Large-scale whole population	Virtual patient register/prevalence	Descriptive analysis	Centralised (i.e. pseudonymised data stored on an ongoing basis)
Patient registries	Individual persons	Cohort studies	Trends	Inferential analysis	Distributed (i.e. data gathered on a project-by-project basis)
Clinical records	Family and/or households	Purposive sampling	Identification of predisposing factors	Correlations (relationship between variables)	Hybrid
Longitudinal cohorts	Addresses/locations	Case control	Identification of outcomes	Regressions (cause and effect)	
Health surveys	Linkage methods	Methodological research	Cost evaluation	Analysis of variance	
Operational	Deterministic using unique identifiers	Two-group comparison	Data validation	Cost analysis	
Health-related social data	Probabilistic using names, addresses, dates of birth, etc.		Characteristics of individuals		
Genomics	Population spine		Policy/clinical guidelines evaluation		
Imaging	Existing		Predictive analytics		
	Purpose-built		Development of new tools, diagnostics, or treatments		
	None				

Case Study #1: Virtual patient registry (foetal valproate syndrome)

Case Study 1 (CS#1) demonstrates a large-scale, whole-population linkage study which identifies the prevalence of a condition of critical importance for health service planning (i.e. a virtual patient register). In this case, children with foetal valproate syndrome (FVS) were identified by linking mothers and children. Data on prescriptions, births, and disabilities were required in order to answer this research question. A distributed model was demonstrated in this case study, as well as probabilistic linking without the availability of a population spine.

CS#1 background

Sodium valproate is recommended as an effective treatment for epilepsy and bipolar disorder (5), but it is now known that if a woman becomes pregnant while taking sodium valproate, the baby is at an increased risk of serious birth defects and developmental disorders; 30–40% of children exposed to sodium valproate reportedly have serious developmental disorders (6) and approximately 11% have major congenital malformations (7). New restrictions on the use of sodium valproate in pregnancy include the Valproate Pregnancy Prevention Programme (PREVENT), which requires that those of childbearing potential no longer be prescribed sodium valproate unless other options are ineffective or not tolerated, and that patients be made fully aware of the risks and the need to avoid becoming pregnant and sign a PREVENT form (8). A Health Service Executive (HSE) review estimated that from 1975 to 2015 in Ireland, between 153 and 341 children experienced a major congenital malformation and up to 1,250 children experienced some form of neurodevelopmental delay (8) diagnosed as Fetal Valproate Syndrome (FVS). However, this analysis largely depended on international data and a range of assumptions and limitations, as individual Irish datasets could not be linked (8). Therefore, the aim of CS#1 was to test how a national DASSL solution could be used to evaluate the impact of sodium valproate on women and children and the impact of the introduction of the PREVENT Programme.

CS#1 datasets

Several datasets could be used to answer this research question. Synthetic versions of the following clinical records, administrative datasets, and patient registers were developed:

- The Primary Care Reimbursement Service (PCRS)
- The epilepsy electronic patient record (EPR)
- Central Statistics Office (CSO) Vital Statistics: Births (CSO Births)
- The National Perinatal Reporting System (NPRS), and
- The National Ability Supports System (NASS).

As this case study was demonstrating a distributed system which gathered the required data from the data controllers on a project-by-project basis, the specific data required would be requested from the data providers once the relevant permissions and approvals were received (Table 2). One file with the matching variables and a second file with the content data were uploaded to the DASSL platform.

Table 2 CS#1 Datasets, Data and Data Providers						
Dataset	**PCRS**	**Epilepsy EPR**	**NPRS**	**CSO Births**	**NASS Service User**	**NASS Services**
Data controllers/ providers	HSE	Beaumont Hospital, St James's Hospital, and St Vincent's University Hospital	Healthcare Pricing Office (HPO)	CSO	HRB and HSE	HRB and HSE
Cohort	Women aged 15–50 years prescribed sodium valproate (with or without folic acid and contraceptives)	Women aged 15–50 years prescribed sodium valproate	All births	All births	All included	All included
Time period (inclusive)	1975–2020	2012–2020	1975–2020	1975–2020	2002–2020	2002–2020

Dataset	PCRS	Epilepsy EPR	NPRS	CSO Births	NASS Service User	NASS Services
Matching variables	First name Surname Birth surname Date of birth (DOB) Personal Public Service Number (PPSN) Address Eircode Nationality General Medical Services (GMS)/ Long-Term Illness (LTI)/Drugs Payment Scheme (DPS) number	Medical record number (MRN) First name Surname Date of birth PPSN Address Nationality Ethnicity	Mother's forename Mother's surname Mother's birth surname Mother's DOB Mother's PPSN Mother's address Mother's county Mother's nationality Child's forename Child's surname Child's DOB Child's sex	Mother's forename Mother's surname Mother's birth surname Mother's DOB Mother's PPSN Mother's address Mother's county Mother's nationality Child's forename Child's surname Child's DOB Child's gender	Forename Surname Address Eircode Sex Date of birth	Forename Surname Address Eircode Sex Date of birth

Dataset	PCRS	Epilepsy EPR	NPRS	CSO Births	NASS Service User	NASS Services
Content variables	ATC code Date of claim	Age Epilepsy diagnosis Prior AEDS Current AEDs Current valproate Prevent reviewed	Child's DOB Child's sex Period of gestation Type of birth Multiple births Birthweight Type of death Cause of death Method of delivery Diseases and conditions in the infant	Child's DOB Child's gender Gestation period Birthweight Multiple type Main disease (foetus)	Sex Disability type Intellectual Autism Deaf or blind Developmental delay Hearing loss/ deafness Neurological Physical Learning disorder Speech/ language Visual Mental health	Sex Service level support Service nights/ week Service days/ week Service weeks/ year Service day sessions Service overnights Service frequency Service hours/ week Service enhancement

Dataset	PCRS	Epilepsy EPR	NPRS	CSO Births	NASS Service User	NASS Services
					Degree intellectual Diagnosis	Unmet service type Unmet service Unmet level support
Estimated size of population (people)	25,000 women	520 women	2,500,000 children 750,000 women	3,000,000 children 1,000,000 women	600,000 children	600,000 children
Estimated size of dataset (rows)	250,000	520	2,500,000 (children) 750,000 (women)	3,000,000 (children) 1,000,000 (women)	6,000,000	6,000,000

CS#1 linkage process

As probabilistic matching was required and there was no population spine, the PoC data linkage unit (DLU) used pairwise record linkage between two datasets at a time with clear guidance provided from the PoC Research Support Unit (RSU). The DLU carried out the following steps/actions:

- Data standardisation and cleaning for data field names (i.e. matching "DOB" field to "date of birth" field) and formats (e.g. changing all names to upper case, removing accents/apostrophes, etc.).
- Internal linkage of individuals within the PCRS was conducted to reduce the number of rows from about 250,000 to about 25,000 in order to reduce the linkage workload. A similar process was completed for CSO Births, NPRS, and the NASS Service (the epilepsy EPR and NASS Service User data were already at an individual level, so internal linkage was not required).
- The PCRS and epilepsy EPR were then linked using probabilistic matching.
- CSO Births was then linked to the PCRS and epilepsy EPR (the same step that was completed for NPRS) using a blocking strategy, which grouped records in smaller matching 'blocks' based on date of birth and county and evaluated them against each other.
- A blocking fields strategy was used due to the size of the CSO Births and NPRS datasets, with both the date of birth and county needing to be correct for each record in order for the rest of the data to be compared.
- Children from CSO Births and NPRS were then each matched using probabilistic methods with NASS Service Users.
- NASS Service Users was linked with NASS Services.
- Clerical review of datasets for matches and non-matches was conducted in order to alter the probabilistic algorithm where required, and to estimate the accuracy of linkage for the researcher linkage quality report.
- Tables describing the linkage between rows of each of these datasets were shared with the RSU along with a linkage quality report.

CS#1 data view preparation

The RSU received the content data files (Table 2) from each of the data providers and the linkage keys from the DLU. As the RSU only received the required data fields from the data providers, and due to the family linkage and the different time periods and population coverages of the different datasets, the following steps were undertaken by the RSU in order to prepare the data:

1. Where individuals from CSO Births and NPRS did not appear on the PCRS and/or epilepsy EPR, the data were removed (all of the data rows from the PCRS and epilepsy EPR were relevant to the research question).
2. Project-specific pseudo-identifiers were then given to these individuals (i.e. women).
3. Where individuals on NASS Service User and NASS Services did not appear on the remaining records from CSO Births and NPRS, these data were also removed, as they were not required.
4. Project-specific pseudo-identifiers were also given to these individuals (i.e. children).
5. Each of the pseudonymised datasets was then uploaded to the safe haven.

If it were deemed inappropriate for a researcher to get access to the month and year of birth and prescription claims, the RSU would need to process the data further and remove rows of data subjects who were exposed to sodium valproate within 10 months of their birth.

CS#1 data analysis

These data could largely be analysed using descriptive statistics in any statistical package to answer several research questions, including:

1. How many women of child-bearing age were taking sodium valproate from 1975 to 2020?
2. Did these women have any prescription birth control or folic acid recorded?
3. How many women who were prescribed sodium valproate had a live birth in the following 10 months?
4. From 2002 onwards, how many of those children had a recorded disability? If so, what types of disabilities? What services did/do they require?
5. Has the introduction of the PREVENT Programme reduced the prescription of sodium valproate? Are those who have been prescribed sodium valproate since 2018 receiving the required information?

CS#1 lessons learned

Several lessons were learned during the generation of the synthetic data, linking of the available variables, creation of the data view, and analysis of the data. These lessons learned are outlined below.

Data utility, quality, and fit for purpose

- The PCRS only captures public prescription data (i.e. gathered using GMS, LTI, and/or DPS numbers), so sodium valproate and folic acid would be included for people with epilepsy with an LTI card, but contraceptives and people taking sodium valproate for other conditions would be missing if they are not on the GMS Scheme or the DPS.
- Data recorded by CSO Births and NPRS can differ, as the CSO collects the register of all births in Ireland while NPRS uses perinatal records from public hospitals.
- The NASS was developed by combining the National Intellectual Disability Database (NIDD) and the National Physical and Sensory Disability Database (NPSDD) in 2018, but the data cover different time periods for intellectual disabilities (1995–present) and physical and sensory disabilities (2002–present), which should be considered in analysis.
- While the epilepsy EPR has been available for over a decade, the data field for the PREVENT form was only added in 2020 and is not yet widely completed enough in order to allow use for data analysis.

Record linkage

- An individual could appear on the PCRS with a different GMS, LTI, and/or DPS number; therefore, these identifiers should be linked back to other personal identifiers (e.g. the PPSN).
- Probabilistic matching was required, as there was no consistent unique identifier across all datasets.
- A blocking strategy was needed for linking the large CSO Births and NPRS datasets.
- Matching of each dataset in a pairwise manner was required, as no population spine was used and the identifiers of mothers and children were being linked to different datasets.
- The NASS is jointly controlled by the HRB and HSE, and only the HSE currently has access to the matching variables to support linkage of the NASS with other datasets.
- The national NPRS dataset does not collect personally identifiable data that would enable linkage.
- Linking of family members required this information to be collected explicitly within the CSO Births and NPRS datasets, but would not facilitate linkage with other family members.

Data view preparation

- Datasets were gathered for this specific project, so the data providers extracted and shared only the necessary, relevant information, which required more work on their behalf.
- As the research question required access to dates (e.g. of childbirth, appointments, and claims), if there were a data protection concern, only month and year could be shared; or, alternatively, the RSU could process the data further.
- Less processing was required by the DLU and RSU, as the data controllers only provided relevant data fields.

Data analysis and interpretation of findings

- Whole-population analysis is facilitated by national-level administrative datasets.
- Prescriptions do not equate to ingestion of medication at that time, but it was assumed that this was the case for this research question, as no other information was available.
- Combining individual-level data (i.e. the NASS) with observation-level data (i.e. the PCRS, CSO Births) can be challenging for the RSU and/or researcher.
- NPRS and CSO Births can be used to expand on the information available in each of these datasets alone.
- Data quality and utility issues in relation to population and time coverage and data completion (discussed above) would need to be shared with the researcher analysing the data.
- Mapping of coding systems in a consistent manner would be helpful, but many datasets have their own specific codes (e.g. the NASS) which would need to be understood by the researchers.
- Many different statistical packages could be used for this analysis, so this depends on researcher preference.
- Usual incidence of a condition (or disability) should be considered in analyses such as this.
- Steps to develop virtual patient registers differ depending on the condition, drug, or intervention of interest, and experts in the field are required in order to determine what constitutes a person of interest (e.g. in the case of those with diabetes, blood testing for glucose levels or the prescription of the drug metformin for diabetes treatment).
- The methodology used in this case study for virtual patient registers can result in false positives and false negatives, as well as potentially being limited in scope (e.g. testing for gestational diabetes only).
- Reidentification of individuals could be deemed important in this type of study (i.e. virtual patient register) if used for service delivery or compensation.

Case Study #2: Identification of social risk factors (mental health and addiction)

The second case study (CS#2) demonstrated the linkage of health data with related social data. In addition to linking individuals, address location was linked to a dataset on social deprivation. CS#2 demonstrated probabilistic linkage with a purposively built population spine. A hybrid model was also demonstrated, with some datasets gathered on a regular basis and stored in a pseudonymised manner, while others were gathered for this specific project. CS#2 focused on mental health and addiction, as this is a critical priority area for Sláintecare and the Department of Health.

CS#2 background

The highest rates of self-harm in Ireland are observed in young people aged 15–24 years (9), and the age of onset of self-harming has been decreasing over a 10-year period from 2007 to 2016 (10). Incidences of self-harm may be associated with social inequalities in particular socioeconomic factors (9), mental health disorders (10), and childhood adversities such as abuse; experiences of deprivation or poverty; and family dysfunction, such as parental separation and familial substance abuse (11). Additionally, 14.4% of new cases of drug treatment in Ireland in 2020 were aged 17 years or under (12), and Irish teenagers were ranked as having among the highest rates of binge drinking in the world in a recent study (13). Social risk factors in childhood can be identified using linked data research in order to inform targeted early interventions. Therefore, the aim of this case study was to demonstrate the use of the DASSL PoC to explore any potential risk factors in childhood for self-harm, suicide, psychiatric conditions, and alcohol and drug issues later in life by linking health and social data.

CS#2 datasets

In order to answer this research question, a synthetic version of a longitudinal cohort was linked with the following national statistics and patient registers:

- The Growing Up in Ireland (GUI) 1998 Cohort
- The National Self-Harm Registry Ireland (NSHRI)
- CSO Vital Statistics: deaths (CSO Mortality)
- The National Psychiatric Inpatient Reporting System (NPIRS)
- The National Drug Treatment Reporting System (NDTRS), and
- The National Drug-Related Deaths Index (NDRDI), and
- The Social Deprivation Index.

As this case study demonstrated a hybrid model, only the (synthetic) data fields required in order to answer the research question were requested from some of the data providers that were sharing the data for this specific project (i.e. GUI, the NSHRI, CSO Mortality), whereas the entire dataset was shared from the NPIRS, NDTRS, and NDRDI on a regular basis, and was updated and maintained (Table 3).

Table 3 CS#2 Datasets, Data and Data Providers

Dataset	GUI 1998 Cohort	NSHRI	CSO Mortality	NPIRS	NDTRS	NDRDI
Data controllers/ providers	Economic and Social Research Institute	National Suicide Research Foundation	CSO	HRB	HRB	HRB
Cohorts	All individuals born in 1998	All individual	All individuals	All individuals	All individuals	All individuals
Time period (inclusive)	2008	2008–2021	2008–2021	2000–2021	2000–2021	2005–2021
Matching variables	First name Surname Date of birth Address Gender	First name Surname Address Gender Date of birth Age	Forename Forename 2 Forename 3 Surname Birth surname DOB Date of death (DOD)	Individual health identifier (IHI) Address Gender Date of birth Ethnicity Country of birth	Forename Surname County Address Eircode Small area Electoral division IHI Gender Transgender Date of birth Country of birth Ethnic/cultural background	Forename Surname Gender Marital status County DOB DOD Country of birth Nationality Ethnicity

Dataset	GUI 1998 Cohort	NSHRI	CSO Mortality	NPIRS	NDTRS	NDRDI
Content variables	Electoral division Parent death Close family member death Death of close friend Divorce/separation of parents Foster home Drugs/alcoholism in family Mental disorder in family Parent in prison Bullying Inadequate dressing Too tired to participate Without lunch Hungry Lack cleanliness Late Homework incomplete Depression score of primary caregiver Depression score of secondary caregiver	Methods	Gender Age at death Cause of death 1a Cause of death 1b Cause of death 1c Cause of death 2	Marital status Gender Age Primary diagnosis Secondary diagnosis No fixed abode Date of discharge Reason for discharge Order of admission Legal status Employment status Private health insurance Medical card Length of stay Hospital type	Age Self-defined sexual orientation Living where Living with whom Education: highest level Age left primary or secondary school for the first time Employment status Main reason for referral Drug (please specify) Other (please specify) Source of referral Assessment outcome Where client is suitable for treatment Number of times started alcohol or drug treatment in this centre this year (January–December)	Age Lives where Occupation Employment status Ever in prison Current in prison Date of release Days released Prison Prison release Alcohol dependency Drug dependency Drug used Drug route Current injecting Previous history of overdose Ever treated Drug treated Current treated Alcohol detox Opiate detox Benzodiazepine detox Other detox Methadone

Dataset	GUI 1998 Cohort	NSHRI	CSO Mortality	NPIRS	NDTRS	NDRDI
	Frequency of alcohol consumption Making ends meet Without heating Social welfare Unemployment support Single parent support Child support Disability support Retirement payment Income quintile Accommodation Bedrooms Household members				Main drug Drug 2 Age first used any drug Exit details	Buprenorphine Medication free Counselling Psychiatric treatment Education awareness programmes Aftercare Prescribed medications Mental illness Mental class Inpatient Poisoning class Coroner verdict Toxicology Poisoning Substances involved Hepatitis B Hepatitis C Human immunodeficiency virus (HIV)
Estimated size of population (people)	8,500	189	168	235,987 (about 15,381 per year)	224,538 (about 16,000 per year)	13,362 (about 786 per year)
Estimated size of dataset (rows)	8,500	225	168	338,602	352,004	13,362

CS#2 linkage process

The DLU developed and maintained a population spine, which included the following data fields: first name, middle name 1, middle name 2, middle name 3, surname, birth surname, other prior surname, mother's birth surname, date of birth, address line 1, address line 2, address line 3, county, Eircode, nationality, other address line 1, other address line 2, other address line 3, other county, other Eircode, sex, PPSN, and IHI. The matching variables fields of the datasets controlled by the HRB (i.e. the NDTRS, NPIRS, and NDRDI) were linked on a yearly basis, while the others (i.e. the GUI 1998 Cohort and the NSHRI) were sent to the DLU for this specific project. This involved mainly probabilistic linkage methods using matching variables. The DLU then took the following steps in order to link the data files:

1. Data were standardised and cleaned prior to each step.
2. Each year, all NDTRS, NPIRS, and NDRDI personal identifiers were each individually linked to the population spine, and the PPSN was hashed and then shared with the RSU. (Datasets could not normally be linked with one another with this centralised model, as the data may be received on different days and only the data for each time period were included.)
3. The GUI 1998 Cohort, CSO Mortality, and the NSHRI were then compared with the population spine and also matched to one another, with CSO Mortality and the NSHRI compared directly with the GUI 1998 Cohort; the linkage key was then shared with the RSU.
4. Clerical review of matches and non-matches was completed and a linkage report was also shared with the RSU.

Alternatively, the DLU could have treated the GUI 1998 Cohort, CSO Mortality, and the NSHRI the same as the previous datasets and compared them with the population spine, and simply provided the hashed PPSNs rather than the linkage key. This would have reduced the guidance and steps required by the DLU; however, if the DLU had also matched the datasets, then the RSU could have immediately removed the unnecessary data related to individuals from CSO Mortality and the NSHRI.

CS#2 data view preparation

The RSU received the content data files for the NDTRS, NPIRS, and NDRDI on a yearly basis and updated the pseudonymised data storage each time using the hashed PPSN generated with the addition of 'salt' (i.e. extra piece of random data) that reduces the risk of cross-linking of the hashed PPSNs beyond this case study. The other datasets were then received by the RSU for the specific project from the data providers. The RSU then took the following steps in order to prepare the datasets for the researchers:

1. The RSU accessed the NDTRS, NPIRS, and NDRDI pseudonymised files and extracted data related to individuals born in 1998.
2. The pseudo-identifiers from the stored datasets and from the NSHRI and CSO Mortality were individually compared with those on the GUI 1998 Cohort.
3. A single dataset was created for all individuals appearing on the GUI 1998 Cohort.
4. Data from the NDTRS, NPIRS, and NDRDI that were not required for the research question were removed (only relevant data had been received for the GUI 1998 Cohort, the NSHRI, and CSO Mortality).
5. Electoral division was replaced with the corresponding Social Deprivation Index (1–10).
6. Project-specific identifiers were then given to each individual.

For the PoC, the RSU received the electoral divisions for each individual, which allowed the RSU to map individuals on the Social Deprivation Index, which includes a ranking from 1 to 10 mapped to each electoral division area. Alternatively, the Social Deprivation Index could have been treated the same as the other datasets, but as it comprises open data that are available online and each electoral division relates to an average of nearly 1,500 people, it was considered less identifiable and the PoC RSU was permitted to see it.

CS#2 data analysis

This research question could be answered using multivariate logistic regression while controlling for sociodemographics, including gender, Social Deprivation Index, and household income. Research questions which could be asked of these data include:

1. What proportion of children have experienced social adversities and/or self-harm, suicide, psychiatric admission, drug/alcohol treatment, or drug-related death?
2. Was the number of childhood adversities related to that person experiencing a drug- or alcohol-related issue, a psychiatric admission, self-harm, or suicide?
3. Was the type of childhood adversity linked to a drug- or alcohol-related issue, a psychiatric admission, self-harm, or suicide?
4. Was the type of outcome experienced by the person related to the number of childhood adversities?

CS#2 lessons learned

In addition to the lessons learned in CS#1, CS#2 provided insights into different datasets and the use of a population spine and a hybrid model. These insights are outlined below.

Data utility, quality, and fit for purpose

- Subjectively reported health information from the GUI 1998 Cohort can be validated by combining this dataset with health datasets.
- The NSHRI and the NDTRS only collect data from all publicly funded services, while the NPIRS collects data from both public and private services.
- The NDRDI gathers data directly from the coroner's office as well as the CSO, the Central Treatment List, and the Hospital In-Patient Enquiry (HIPE), and therefore the data from this dataset should largely contain the same individuals as CSO Mortality, but the NDRDI dataset contains additional metadata on deaths.
- The accuracy of address data can be improved where data entry systems include a database of addresses (e.g. the NDRDI).
- Free text data entry for drugs (compared with the use of codes only) could result in spelling errors and make it more difficult to aggregate data.

Record linkage

- The HRB version of the NPIRS only collects date of birth, gender, address, and nationality; in order to support record linkage, full names or a unique identifier would also need to be available from the service providers (the IHI was added to the synthetic version of the NPIRS in CS#2).
- Similarly, the NSHRI currently records only the initial letters of an individual's name in an encrypted form and encoded dates of birth and addresses for the purpose of identifying duplicates, but in reality, more information would need to be added in order to support linkage with other datasets.
- A mix of identifiers within these datasets required a combination of deterministic and probabilistic matching as well as a population spine.
- A population spine was required in order to facilitate the storage, maintenance, and subsequent linkage of pseudonymised datasets.
- The quality of personal identifiers can be improved where a database of addresses is used within the data entry system (e.g. the NDRDI).

Data view preparation

- Replacement of the location information with the relevant information (i.e. Social Deprivation Index) anonymises this information, and it may make more sense for the RSU to perform this task as opposed to the DLU.
- The creation of a single file for the researcher requires highlighting which data fields belong to which dataset and, as always, the data dictionary for each dataset is required, as both the coding systems and data field names may also be unfamiliar and non-intuitive to the researcher.
- Cleaning and standardisation of data were required in order to ensure a consistent use of codes for 'unknown', 'refused to answer', and 'missing', and this work could be completed by the RSU or left to the researcher.

Data analysis and interpretation of findings

- While the research data file was restricted to those individuals present in the GUI 1998 Cohort, the putative researcher may have requested access to all individuals born in 1998 from the other datasets in order to see how the results compare with the overall population, and ethical and information governance over this level of access would need to be considered.

- Datasets such as the NPIRS have used different versions of coding systems and terminologies (i.e. the International Classification of Diseases (ICD)), and where both an older and newer version are being aggregated for analysis, mapping of the different versions may be required during analysis.
- Data limitations need to be acknowledged, such as some datasets only collecting information on public services attended, and individuals from the GUI 1998 Cohort being lost to follow-up due to moving to another country (as opposed to not requiring health services as an adult).

Case Study #3: Long-term outcomes and costs of healthcare initiative (hip fractures)

CS#3 demonstrates the linkage of very large datasets across both primary and secondary healthcare services in order to identify the long-term outcomes of patients and evaluate a new healthcare initiative. The IHI was applied to most of the datasets used for this case study in order to demonstrate linkage of the IHI where errors could occur, as well as the linkage of the IHI with names, addresses, dates of birth, etc. A population spine was required for this type of linkage, and the IHI register was used for this. Additionally, this case study demonstrated a centralised model where each of the datasets was updated on a regular basis and stored within the data hub.

CS#3 background

In 2019, 3,701 cases of hip fractures were captured by the Irish Hip Fracture Database (IHFD) across 16 hospitals in Ireland (14). Hip fractures are a major public health problem which can lead to disability, reduced quality of life, and higher mortality in those aged 65 years or over. In order to improve the management of hip fractures and thus patient outcomes, the IHFD introduced the Best Practice Tariff (BPT) in 2018. The BPT is a performance incentive for hospitals operating on patients with hip fractures (aged 60 years or over) which pays hospitals €1,000 per case that meets the eight standards of care of the BPT.

However, despite the large financial investment in the BPT in Ireland, it is unknown what impact it has had on the long-term outcomes of patients with hip fracture, as the longitudinal outcomes of these patients on discharge from the orthopaedic acute setting remain unknown. Therefore, the aim of this case study was to demonstrate how the impact of the introduction of the BPT could be evaluated by linking datasets using the DASSL model.

CS#3 datasets

In order to answer this research question, a synthetic version of a national clinical audit was linked with vital statistics from the CSO, an administrative dataset, and clinical records (Table 4). Staffing levels at each of 16 hypothetical hospitals were also linked in order to assess any impact on differences between hospitals. The linked datasets were:

- The IHFD
- HIPE
- General practitioner (GP) EPR
- CSO Vital Statistics: deaths (CSO Mortality)
- Healthlink, and
- Hospital staffing levels.

Table 4 CS#3 Datasets, Data and Data Providers

Dataset	IHFD	HIPE	CSO Mortality	GP EPR	Healthlink	Hospital staffing levels
Data controllers/ providers	National Office of Clinical Audit (NOCA)	HPO	CSO	GPs	HSE	HSE
Cohorts	All included	All individuals	All individuals	All individuals	All individuals	All hospitals
Time period (inclusive)	2015–2020	2010–2020	2010–2020	2010–2020	2010–2020	2010–2020
Matching variables	IHI Gender Hospital	IHI Full name Eircode Sex Date of birth Hospital	Forename Forename 2 Forename 3 Surname Date of birth Address County Country Nationality Country of birth DOD Gender	IHI Forename Middle name Surname Gender DOB Address PPSN	IHI	Hospital code

Dataset	IHFD	HIPE	CSO Mortality	GP EPR	Healthlink	Hospital staffing levels
Content variables	Gender Age Date and time of trauma Type of trauma Date and time of arrival Admission via emergency department (ED) Date and time of arrival in ED Directly to theatre from ED Date and time left ED Date and time seen by orthopaedic team Type of ward Pre-fracture walking (indoor, outdoor, and shopping) Mobility score Abbreviated Mental Test (AMT) score	Sex Age Marital status Residence Admission date Discharge date Length of stay Day case Admission source Admission type Elective type Waiting list Transfer in or out Emergency admission Procedures Procedure dates Discharge codes Date of transfer Major diagnostic code Diagnosis-related grouping Specialty	Age Gender Marital status Place of death Cause of death Occupation code Principal Economic status (PES) code Nomenclature of Territorial Units for Statistics (NUTS) 3 Area of residence Duration	Date Gender Age Patient type Consultation type Codes (International Classification of Diseases, Tenth Revision (ICD-10) and International Classification of Primary Care (ICPC)) Drugs Immunisations eReferrals Investigations	Message type Facility code Class of patient Financial class Admit source Ambulatory status Disposition Discharge location Provider role Diagnosis Observation/ test Observation/ result Flags indicating abnormalities Observation result status Patient death Allergy Admission reason	Orthopaedic consultants (full-time equivalents (FTEs))

Dataset	IHFD	HIPE	CSO Mortality	GP EPR	Healthlink	Hospital staffing levels
	Delirium assessment	Critical care bed days				
	Side of fracture	Intensive Therapy Unit (ITU) days				
	Type of fracture					
	Pathological	Continuous ventilation support				
	History of fragility fracture					
	Pre-operative medical assessment					
	Assessed by geriatrician					
	Nutritional risk assessment					
	Nerve block before theatre					
	Operation					
	American Society of Anesthesiologists (ASA) grade					
	Type of anaesthesia					
	Surgeon grade					
	Consultant orthopaedic surgeon					
	Anaesthetist grade					
	Consultant anaesthetist					

Dataset	IHFD	HIPE	CSO Mortality	GP EPR	Healthlink	Hospital staffing levels
	Date and time of primary surgery Reason if delay >48 hours Mobilised on day of or day after Physiotherapy assessment Cumulative Ambulatory Score Reoperation within 30 days Pressure ulcers Specialist falls assessment Bone protection medication Multidisciplinary team rehabilitation assessment Discharge location					
Estimated size of population (people or hospitals)	15,234	5,000,000	341,000	7,000,000	2,500,000	16
Estimated size of dataset (rows)	21,206 (about 3,500 per year)	18,700,000 (about 1,700,000 per year)	341,000 (about 31,000 per year)	49,000,000 (about 4,500,000 per year)	12,000,000 (about 1,091,000 per year)	16

CS#3 linkage process

This case study demonstrates the consistent use of a unique identifier across health datasets and the relatively straightforward linkage process. However, it also demonstrates minor errors in the IHI in some cases, as well as issues with linkage with the CSO Mortality register, which does not include the IHI. Hospitals were also linked in this case study. The IHI register was used as the population spine in this case. The following steps were followed for linkage by the DLU in this case study:

- The datasets as outlined in Table 4 were standardised and cleaned as required once received from data providers.
- The IHFD, HIPE, the GP EPR, and Healthlink were received yearly (or monthly) from the respective data providers and linked to the IHI register using probabilistic methods in order to overcome any potential errors or missing IHIs.
- IHIs were hashed.
- The hospital codes on the IHFD and the staffing numbers file were hashed.
- Pseudo-identifiers (i.e. hashed IHIs with 'salt') and hashed hospital codes were shared with the RSU on a regular basis.

In this case study, each dataset was matched to the population spine (datasets were not matched with one another) and the DLU required no guidance from the RSU due to the centralised model.

CS#3 data view preparation

The RSU received monthly updates of Healthlink and the GP EPR, and yearly updates of the IHFD, HIPE, hospital staffing levels, and CSO Mortality. The RSU received the corresponding pseudo-identifiers from the DLU and updated its stored versions of the datasets. Further pseudonymisation was applied to the pseudo-identifiers while they were in storage in order to further protect them (i.e. additional 'salt' was added to hashed IHIs). The staffing levels at hospitals were also updated by using the hashed hospital codes. The RSU took the following steps to prepare the data for this research project:

1. The data requested from the IHFD were extracted from the data storage.
2. Where a pseudo-identifier from the IHFD also appeared on HIPE, the GP EPR, Healthlink, and CSO Mortality, and the data were recorded after the IHFD presentation, these relevant data were also extracted and combined with the IHFD data.
3. The hashed hospital codes on the IHFD were replaced with staffing level numbers.
4. Aggregated age groups and gender were left on the dataset.
5. Project-specific identifiers were provided for each data subject in the researcher data view.

Alternatively, the identifiable hospital name or code could be shared directly with the RSU (as with electoral divisions in CS#2) if this was deemed appropriate. However, unless specifically approved by governance boards, the hospital code should not be shared with the researcher in any circumstance.

CS#3 data analysis

The analysis on this researcher data view to answer the above research question could include some regression analysis and an economic analysis. The following research questions could be asked of the researcher data view created:

1. What were the long-term outcomes of people post-hip fracture (e.g. further healthcare utilisation, death) pre- and post-2018 and the implementation of the BPT?
2. How did the patient outcomes compare pre- and post-2018?
3. Was the BPT cost-effective?

CS#3 lessons learned

A number of lessons were learned from this case study due to the size of the datasets and the use of the IHI as a population spine, as well as in relation to the specific datasets. These lessons learned are outlined below.

Data utility, quality, and fit for purpose

- HIPE data are entered by trained data coders, whereas many other datasets each have individual service providers entering data, which can impact on data reliability and validity.
- HIPE and the IHFD capture data from public hospitals only.
- CSO Mortality covers all deaths in Ireland.
- The CSO Mortality dataset includes useful information from the Coroner that are protected under the Statistics Act, 1993. A lawful basis for sharing CSO Mortality with a national DASSL solution external to the CSO would need to be considered for this valuable dataset to be made available for linkage and research.
- GP Electronic Patient Record (EPR) data cover both public and private patients, but the data used in this case study are not currently made available centrally. The capability to pull data from the four most commonly-used GP EPR systems would need to be developed in collaboration with the vendors of these systems.
- The four most commonly-used GP EPR systems produce different data formats and fields which would need to be mapped to one another if these data are to be centralised for reuse.
- Making the centralisation of data from GP systems either an opt-in or opt-out system for patients and/or GPs would impact on the population coverage of the data.
- Healthcare providers who use Healthlink could adopt several different standardised terminologies, coding systems, or local coding systems.
- The use of two different coding systems (one for diagnosis and one for reason for consultation) within GP systems in Ireland, along with the expected implementation of the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), will require mapping of terminologies and coding systems in order to facilitate data aggregation and analysis.
- Coded data are not always entered by the GP, resulting in only free text being available, which is valuable for important contextual information regarding the code but is not easily processible by computers.
- Healthlink does not currently store the data in a centralised database on an ongoing basis, but these data may be retrievable from different GP systems and hospital systems which will require considerable effort and investment.

Record linkage

- The IHFD and the national HIPE dataset do not collect personally identifiable data (although HIPE datasets at individual hospitals do), so the data in these datasets cannot be linked; however, for the purpose of this case study, the IHI was applied to these datasets as well as to the GP EPR and Healthlink datasets.
- Linking of the IHI with the names, addresses, and DOBs on CSO Mortality required a population spine.
- Errors can occur if unique identifiers are manually entered, requiring the use of other personally identifiable information in those datasets for record linkage.
- Even with yearly updates, HIPE, the GP EPR, and Healthlink were so large that blocking of data variables was required (e.g. year of birth).

Data view preparation

- Gathering the approved data from the stored data required more time and effort on behalf of the RSU compared with CS#1 and CS#2, which demonstrated the use of centralised and hybrid models and smaller datasets.
- In order to avoid the provision of exact dates contained within records to researchers, the RSU can implement additional data processing to protect privacy where required and dependent upon the research analysis.

Data analysis and interpretation of findings

- Cost evaluations require costing information for inpatient and GP visits, which the researcher may ask the RSU to import into the safe haven on their behalf; this would require the researcher to operate as a data provider.
- Using coding standards alone to determine if a person presents to healthcare providers for the same hip issue can be challenging without access to free text fields.
- If only purposive samples of the GP EPR and Healthlink are available due to consent procedures and use of Healthlink functions, then only the data subjects present on these datasets should be included in the analysis of healthcare utilisation.

Case Study #4: Predisposing genetic factors (cancer)

CS#4 demonstrates the combination and analysis of genomics data linked with tabular data. For the purpose of this demonstration, the PPSN was applied to the synthetic datasets, and only the data required for the research question were gathered (i.e. it used the distributed model). This case study was conceived to demonstrate the significant difference in the nature of analyses involving genomic data, which would involve considerably larger compute and data storage resources compared with the other case studies. Furthermore, genomic studies tend to require bespoke software that will necessitate extensive configuration of the safe haven system used for the project.

CS#4 background

Genomics can provide information on predisposition to certain types of cancer. Colorectal cancer is diagnosed in more than 2,500 people in Ireland each year (11). It has been shown that some rare genetic mutations are related to the occurrence of colorectal cancer, and particularly to the early onset of cancer (in those aged under 40 years) (12). This case study examined the identification of different genetic mutations for correlation with the manifestation of colorectal cancer in different age groups. The research aim of this case study was to emulate the identification of novel gene mutations (in the APB gene) that correspond to incidences of late-onset colorectal cancer, relative to other known mutations of the gene that typically lead to early-onset cancer (in those aged under 40 years).

CS#4 datasets

This case study aimed to demonstrate the use of genomics within the DASSL PoC, but there is no actual national genomics dataset in Ireland to base this on. A synthetic version was therefore created based on the literature and expertise in genomics. Additionally, both synthetic datasets in this case study used the PPSN to allow testing of this type of linkage. The following datasets were artificially generated:

- An artificial national genomics dataset (genomics), and
- The National Cancer Registry Ireland (NCRI).

Table 5 CS#4 Datasets, Data and Data Providers		
Dataset	**Genomics**	**NCRI**
Data controllers/ providers	Not applicable	NCRI
Cohorts	Sample of population	Individuals with colorectal cancer
Time period (inclusive)	2016–2020	2016–2020
Matching variables	PPSN	PPSN
Content variables	Genetic sequences with mutations	Gender Age group Cancer type ICD-10-O-3 ICD-10-Site Topography Tumour behaviour (ICD-O-3) ICCC group Malignancy Microscopic verification Smoking status Deprivation index Sum stage Grade Death certificate Autopsy Mortality Year of death Survival (months) Surgery Chemotherapy Hormone therapy Medical oncology treatment Radiotherapy
Estimated size of population (people)	1,000,000	12,500
Estimated size of dataset (rows)	1,000,000	12,500

CS#4 linkage process

The PoC DLU received the file of PPSNs from the data providers, and conducted the following steps in order to link the two datasets:

1. The PPSNs were deterministically linked across datasets.
2. The PPSNs were hashed.
3. The linkage key and hashed PPSNs were shared with the RSU.

No data standardisation, cleaning, population spine, or clerical review were required.

CS#4 data view preparation

The PoC RSU received the content data and linkage key, and took the following steps:

1. The RSU removed individuals who did not appear on both datasets.
2. The RSU created a single purposive sample with the content data.

CS#4 data analysis

A genomics software application suite, such as Genome Analysis Toolkit (GATK) (15), along with the R package for statistical analysis, may be used to identify gene mutations and conduct statistical analyses incorporating cancer patient data from the NCRI. Potential research questions that can be posed include the association of mutations with particular disease patterns (e.g. are there novel genetic mutations in the APB gene that correlate with late-onset colorectal cancer?) and patient outcomes.

(16, 17) CS#4 lessons learned

Additional learnings were derived from CS#4 in relation to the use of genetic data and the use of the PPSN across all datasets. These lessons learned are outlined below.

Data utility, quality, and fit for purpose

- The NCRI collects data on all cancer diagnoses in Ireland under legislation (S.I. No. 19/1991 - The National Cancer Registry Board (Establishment) Order, 1991).
- No national genomics dataset currently exists in Ireland, but it is expected that this would be available in the future, although it would only represent a portion of the population.

Record linkage

- PPSNs can be matched using exact matching, but if an individual's PPSN changed or errors in entry occurred, false negatives would occur, reducing the quality of the linked data.
- A population spine and clerical review are not required if only linking datasets which consistently collect the same unique identifiers.

Data view preparation

- It is relatively trivial for the RSU to combine two datasets which already contain only the relevant and approved research data.

Data analysis and interpretation of findings

- If a future national genomics data provider (e.g. a biobank) in Ireland is established independent of a national DASSL solution, the data may not be allowed to leave the genomics data provider due to the particular sensitivity of genomics data; however, the linkage key and pseudonymised NCRI content data may be shared with the genomics data provider.
- Notably, some biobanks (e.g. in Finland) have their own safe haven environments for conducting genomics analyses (16, 17). These biobanks may employ data trusts to conduct records linkage with other datasets (external to the biobanks).
- Linking a national level dataset with a sample population of genomics data ensured a relatively large sample, but if a smaller dataset (e.g. The Irish Longitudinal Study on Ageing (TILDA)) was linked with the sample of genomics data, the size of the population available for analysis would further decrease.
- Genomic analyses on the raw data (should such data be permitted to be analysed in the safe haven) tend to require bespoke software that would necessitate extensive configuration of the safe haven system used for the project.
- Genomics datasets are typically larger in size compared with tabular statistical datasets, requiring greater computing power.

Case Study #5: Image interpretation using machine learning (COVID-19)

CS#5 demonstrates the linkage of tabular data with medical images and the application of machine learning to linked data. CS#5 tested the ability to diagnose COVID-19 using X-rays by training a model on a purposive sample of lung images with and without COVID-19, and with and without a COVID-19 vaccine at the time that the image was taken.

CS#5 background

A polymerase chain reaction (PCR) test is commonly conducted by healthcare professionals in a laboratory-based environment in order to determine if a person has COVID-19. However, timely assessments of disease progression in patients with COVID-19 is very important in providing personalised treatment. Radiomics have been shown to provide good predictive performance in determining the diagnosis, progress, and outcome of patients with COVID-19 (18). As several COVID-19 vaccines have also now been widely distributed, the impact of having or not having the vaccine on the lung X-rays is also of interest. The aim of this case study was to demonstrate the development and testing of an algorithmic model to identify the diagnosis and prognosis of patients with COVID-19 and determine whether receiving one or more vaccine doses impacted on this algorithm's ability to diagnose COVID-19.

CS#5 datasets

Synthetic X-ray images and tabular COVID-19 data were generated for this case study (Table 6). As this case study aimed to demonstrate a situation where all datasets used the IHI, the IHI was applied to each synthetic dataset and only the required data were gathered from data providers or were synthetically generated. Four datasets were used/generated:

- X-ray images
- COVID Care Tracker (CCT)
- COVID-19 Vaccine Database (COVAX), and
- Computerised Infectious Disease Reporting (CIDR).

Table 6 CS#5 Datasets, Data and Data Providers				
Dataset	**X-rays**	**CCT**	**COVAX**	**CIDR**
Data controllers/ providers	Not applicable	HSE	HSE	Health Protection Surveillance Centre (HPSC)
Cohorts	Random sample	All individuals	All individuals	All individuals with COVID-19 infection
Time period (inclusive)	2020–2022	2020–2022	2020–2022	2020–2022
Matching variables	IHI	IHI	IHI	IHI
Content variables	X-ray Date	Gender Age range Date COVID-19 result	Gender Age Date Vaccination status Has booster Vaccine product	Gender Age Date Disease
Estimated size of population (people)	4,500	1,250,000	3,800,000	1,250,000
Estimated size of dataset (rows)	4,500	1,500,000	7,800,000	1,500,000

CS#5 linkage process

For the purpose of this case study, it was assumed that the IHI was seeded across each of these datasets accurately. Therefore, a relatively simple deterministic linkage process could be undertaken. The PoC DLU took the following steps:

1. Deterministically matched the IHIs from the National Integrated Medical Imaging System (NIMIS), the CCT, COVAX, and CIDR
2. Hashed the IHIs, and
3. Shared the hashed IHIs and linkage key with the RSU.

CS#5 data view preparation

The RSU received the purposively sampled images from NIMIS (N=4500) along with all of the CCT, COVAX, and CIDR content data. As the researcher was only interested in the data related to the X-ray images, the RSU took the following steps to prepare the data view:

1. Matched the pseudo-identifiers from the X-ray images to the CCT, COVAX, and CIDR
2. Removed data from the CCT, COVAX, and CIDR that were not relevant to the research question
3. Placed images that were taken within 2 weeks of a positive COVID-19 test in the 'COVID-19-positive' folder, and those where either COVID-19 was not confirmed on the CCT or CIDR, or that were not taken within the 2-week time frame, in the 'COVID-19-negative' folder
4. Created subfolders within the 'COVID-19-positive' and 'COVID-19-negative' folders in order to segregate those who had received a COVID-19 vaccine at least 2 weeks prior to infection, and
5. Of the individuals appearing on NIMIS, included 1,000 with COVID-19 (250 of whom were vaccinated) and 1,000 without COVID-19 (250 of whom were vaccinated) in the data view, and removed all other data, as they were not required by the researcher.

Alternatively, all the images could have been shared in a single folder with the researcher, with the corresponding information in relation to COVID-19 infection and vaccination supplied in a table. Sharing of dates with the researchers would have reduced the workload of the RSU, which had to analyse these to determine if each image was taken within the specified time frame. However, this could be considered identifiable information not appropriate for the external researcher to see.

CS#5 data analysis

A binary classifier could be developed by training it on 80% of the images (50% of which were from individuals who had COVID-19). The remaining 20% of the images could then be used to test the model to see if it could determine if someone has COVID-19 or not. During the PoC, the requirements for the development of this artificial intelligence (AI) model were the Python programming environment; installation of a Jupyter Notebook through Anaconda; and Python libraries such as TensorFlow and/or PyTorch, along with Matplotlib, scikit-learn, and NumPy. The analyses also required at least 500 gigabytes of storage.

CS#5 lessons learned

The inclusion of the recently developed COVID-19-related datasets, as well as images, provided some learnings during this PoC. These are outlined below.

Data utility, quality, and fit for purpose

- NIMIS captures images gathered from most public hospitals in Ireland, which would make it a valuable source of diverse images from many different people and taken by different machines that could be used for analytics and machine learning with the potential to improve diagnoses and healthcare delivery.
- Images on NIMIS all use the common Digital Imaging and Communications in Medicine (DICOM) standard.
- Both the CCT and COVAX had a very comprehensive view of the entire population that contracted COVID-19 or received a COVID-19 vaccination, as a significant proportion of the population participated in the COVID-19 testing and vaccination programmes; however, there could be cases missed, for example those who were vaccinated in another country.
- CIDR has a legislative basis for collecting information on COVID-19 as an infectious disease, and therefore provides a comprehensive view of the Irish population.

- The CIDR and CCT datasets may overlap, containing information about the same people who have received a COVID-19 vaccine, but this overlap can be used to cross-validate data.
- As COVAX largely used PPSNs, and the IHI was seeded in the CCT, these datasets should in reality allow for high-quality linkage.

Record linkage

- The DLU's role is largely simplified where every dataset accurately uses the IHI, but if a social dataset without the IHI were to be linked, then a population spine would be required.

Data view preparation

- If personally identifiable data were embedded within the images, these would need to be removed (potentially by the data provider) prior to sharing the images with the RSU.
- For this case study, folders of images were created, but a table of information in relation to each image could have been shared with the researcher, which would have reduced the RSU's workload.

Data analysis and interpretation of findings

- A purposive sample of images could be used for this analysis, as opposed to a whole population analysis as in CS#1.
- Researchers could require a number of different tools for this type of analysis, and these would need to be requested from the RSU along with the specific software packages and libraries for the safe haven.
- Should the researchers request to export a trained AI model from the safe haven, checking this type of output would be significantly different in nature to assessing traditional data tables, as it can be more difficult to determine if there is any potentially identifiable information within an AI model.

Conclusion

The case studies demonstrated many of the benefits, risks, and requirements of a national DASSL solution and the national health and related datasets in Ireland. While the technical infrastructure will support the operation of a DASSL model, including the secure and safe access, sharing, storage, and linkage of health and related datasets, the quality of the datasets and the ability to link important datasets are critical. There are many different use cases for these types of data in the Irish context, and these are only some of the case studies that were used to test the DASSL PoC.

Bibliography

1. Moran R. Proposals for an Enabling Data Environment for Health and Related Research in Ireland. Dublin: Health Research Board; 2016.
2. Nowok B, Raab GM, Dibben C. synthpop: Bespoke creation of synthetic data in R. J Stat Softw. 2016;74(11):1-26.
3. DataCebo. The Synthetic Data Vault [Internet]. [n.d.] [cited 21 Jul 2022]. Available from: https://sdv.dev/
4. Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. 2018. Available from: https://doi.org/10.48550/arxiv.1812.04948
5. National Institute for Health and Care Excellence. Epilepsies in children, young people and adults. London: National Institute for Health and Care Excellence; 2019.
6. Meador KJ. Effects of In Utero Antiepileptic Drug Exposure. Epilepsy Curr. 2008;8(6):143–7.
7. Meador KJ, Baker GA, Browning N, Cohen MJ, Bromley RL, Clayton-Smith J, et al. Fetal antiepileptic drug exposure and cognitive outcomes at age 6 years (NEAD study): a prospective observational study. Lancet Neurol. 2013;12(3):244–52.
8. Health Service Executive. Health Service Executive Valproate Response Project. 2019.
9. National Self-Harm Registry Ireland. National Self-Harm Registry Ireland – Annual Report 2018. 2018.
10. Griffin E, McMahon E, McNicholas F, Corcoran P, Perry IJ, Arensman E. Increasing rates of self-harm among children, adolescents and young adults: a 10-year national registry study 2007–2016. Soc Psychiatry Psychiatr Epidemiol. 2018;53(7):663–71.
11. Carbone JT, Jackson DB, Holzer KJ, Vaughn MG. Childhood adversity, suicidality, and non-suicidal self-injury among children and adolescents admitted to emergency departments. Ann Epidemiol. 2021;60:21–7.
12. Kelleher C, Carew AM, Lyons S. HRB Bulletin – National Drug Treatment Reporting System – 2014-2020 Drug Treatment Data. Dublin: Health Research Board; 2021.

13. Azzopardi PS, Hearps SJC, Francis KL, Kennedy EC, Mokdad AH, Kassebaum NJ, et al. Progress in adolescent health and wellbeing: tracking 12 headline indicators for 195 countries and territories, 1990-2016. Lancet. 2019;393:1101–18.

14. National Office of Clinical Audit. Irish Hip Fracture Database National Report 2019. 2019.

15. van der Auwera G, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. 1st ed. O'Reilly Media. 2020.

16. Finnish BioBank Cooperative. Fingenious [Internet]. [n.d.] [cited 21 Jul 2022]. Available from: https://site.fingenious.fi/en/

17. South Australian Genomics Centre. South Australian Genomics Centre [Internet]. [n.d.] [cited 25 Jul 2022]. Available from: https://www.sa-genomics.com.au/index.php

18. Wang D, Huang C, Bao S, Fan T, Sun Z, Wang Y, et al. Study on the prognosis predictive model of COVID-19 patients based on CT radiomics. Sci Rep. 2021;11:11591.

Notes

Notes