A National Network of Safe Havens

Gao, Chuang; McGilchrist, Mark; Mumtaz, Shahzad; Hall, Christopher; Anderson, Lesley; Zurowski, John

DOI:
10.2196/preprints.31684

Publication date:
2021

Licence:
CC BY

Document Version
Early version, also known as pre-print

Link to publication in Discovery Research Portal

Citation for published version (APA):
Gao, C., McGilchrist, M., Mumtaz, S., Hall, C., Anderson, L., Zurowski, J., Gordon, S., Lumsden, J., Munro, V., Wozniak, A., Sibley, M., Banks, C., Duncan, C., Linksted, P., Hume, A., Stables, C., Mayor, C., Caldwell, J., Wilde, K., ... Jefferson, E. (2021). *A National Network of Safe Havens: A Scottish Perspective*. JMIR Publications Inc. https://doi.org/10.2196/preprints.31684

A National Network of Safe Havens: A Scottish Perspective

Chuang Gao, Mark McGilchrist, Shahzad Mumtaz, Christopher Hall, Lesley Ann Anderson, John Zurowski, Sharon Gordon, Joanne Lumsden, Vicky Munro, Artur Wozniak, Michael Sibley, Christopher Banks, Chris Duncan, Pamela Linksted, Alastair Hume, Catherine L Stables, Charlie Mayor, Jacqueline Caldwell, Katie Wilde, Christian Cole, Emily Jefferson

Submitted to: Journal of Medical Internet Research
on: July 09, 2021

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

A National Network of Safe Havens: A Scottish Perspective

Chuang Gao[1] PhD; Mark McGilchrist[1] PhD; Shahzad Mumtaz[1] PhD; Christopher Hall[1]; Lesley Ann Anderson[2] PhD; John Zurowski[3]; Sharon Gordon[4]; Joanne Lumsden[4]; Vicky Munro[4]; Artur Wozniak[4]; Michael Sibley[5]; Christopher Banks[5]; Chris Duncan[6]; Pamela Linksted[6]; Alastair Hume[7]; Catherine L Stables[8]; Charlie Mayor[9]; Jacqueline Caldwell[5]; Katie Wilde[4]; Christian Cole[1]; Emily Jefferson[1] PhD, BSc

[1]Health Informatics Centre Ninewells Hospital & Medical School University of Dundee Dundee GB
[2]Centre for Health Data Science University of Aberdeen Aberdeen GB
[3]Imaging Centre of Excellence Queen Elizabeth University Hospital Glasgow GB
[4]DaSH Aberdeen Centre for Health Data Science University of Aberdeen Aberdeen GB
[5]Electronic Data Research and Innovation Service Public Health Scotland Edinburgh GB
[6]Lothian Research Safe Haven Department of Public Health and Health Policy NHS Lothian Edinburgh GB
[7]EPCC University of Edinburgh Edinburgh GB
[8]DataLoch Usher Institute University of Edinburgh Edinburgh GB
[9]Glasgow Safe Haven Research and Development division of NHS Greater Glasgow and Clyde Glasgow GB

Corresponding Author:
Emily Jefferson PhD, BSc
Health Informatics Centre
Ninewells Hospital & Medical School
University of Dundee
Mail Box 15
Ninewells Hospital & Medical School
Dundee
GB

Abstract

For over a decade, Scotland has implemented and operationalised a system of Safe Havens providing secure analytics platforms for researchers to access linked, de-identified Electronic Health Records (EHRs) while managing the risk of unauthorised re-identification. In this paper a perspective is provided on the state-of-the-art Scottish Safe Haven Network, including its evolution, to define the key activities required to scale Scottish Safe Haven Network capability to facilitate research and healthcare improvement initiatives. A set of processes related to EHR data and their delivery in Scotland are discussed. An interview with each Safe Haven was conducted to understand their services in detail and the commonalities. The results present how Safe Havens in Scotland have protected privacy while facilitating the reuse of the EHR data. This study provides a common definition of a ‘Safe Haven’ and promotes a consistent understanding among the Scottish Safe Haven Network as well as the clinical and academic research community. We conclude by identifying areas where efficiencies across the network can be made to meet the needs of population-level studies at scale.

(JMIR Preprints 09/07/2021:31684)
DOI: https://doi.org/10.2196/preprints.31684

Original Manuscript

Viewpoint

A National Network of Safe Havens: A Scottish Perspective

Chuang Gao[1], Mark McGilchrist[1], Shahzad Mumtaz[1], Christopher Hall[1], Lesley Ann. Anderson[2], John Zurowski[3], Sharon Gordon[4], Joanne Lumsden[4], Vicky Munro[4], Artur Wozniak[4], Michael Sibley[5], Christopher Banks[5], Chris Duncan[6], Pamela Linksted[6], Alastair Hume[7], Catherine L. Stables[8], Charlie Mayor[9], Jacqueline Caldwell[5], Katie Wilde[4], Christian Cole[1], Emily Jefferson[1&3*]

[1] Health Informatics Centre, University of Dundee, Mail Box 15, Ninewells Hospital & Medical School, Dundee, DD1 9SY, UK; [2] Centre for Health Data Science, University of Aberdeen, King's College, Aberdeen, AB24 3FX, UK; [3] Imaging Centre of Excellence, Queen Elizabeth University Hospital, 1345 Govan Road, Glasgow, G51 4TF, UK; [4] DaSH, Aberdeen Centre for Health Data Science, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen, AB25 2ZD UK; [5] Electronic Data Research and Innovation Service, Public Health Scotland, Gyle Square, 1 South Gyle Crescent, Edinburgh EH12 9EB, UK; [6] Lothian Research Safe Haven, Department of Public Health and Health Policy NHS Lothian, Waverley Gate 2-4, Waterloo Place, Edinburgh, EH1 3EG, UK; [7] EPCC, University of Edinburgh, Bayes Centre, 47 Potterrow, Edinburgh EH8 9BT, UK; [8] DataLoch, Usher Institute, University of Edinburgh, NINE Edinburgh Bioquarter, 9 Little France Road, Edinburgh EH16 4UX, UK; [9] Glasgow Safe Haven, Research and Development division of NHS Greater Glasgow and Clyde, 1 Smithhills Street, Paisley, PA1 1EA, UK;

Abstract

For over a decade, Scotland has implemented and operationalised a system of Safe Havens providing secure analytics platforms for researchers to access linked, de-identified Electronic Health Records (EHRs) while managing the risk of unauthorised re-identification. In this paper a perspective is provided on the state-of-the-art Scottish Safe Haven Network, including its evolution, to define the key activities required to scale Scottish Safe Haven Network capability to facilitate research and healthcare improvement initiatives. A set of processes related to EHR data and their delivery in Scotland are discussed. An interview with each Safe Haven was conducted to understand their services in detail and the commonalities. The results present how Safe Havens in Scotland have protected privacy while facilitating the reuse of the EHR data. This study provides a common definition of a 'Safe Haven' and promotes a consistent understanding among the Scottish Safe Haven Network as well as the clinical and academic research community. We conclude by identifying areas where efficiencies across the network can be made to meet the needs of population-level studies at scale.

Keywords: Electronic Health Records; Safe Haven; Data Governance

1 Introduction

Electronic Health Records (EHRs) are routinely collected data generated when an individual receives care in a health care setting. EHRs typically contain records of medical history, diagnoses, medications, allergies, immunisations, and other treatments as well as laboratory results [1]. The records can be generated in different settings e.g. primary care facilities, such as clinics and health care centres; secondary care facilities, such as hospitals and emergency care centres. Although the primary purpose of EHRs is to improve the direct care of patients, they also have some other purposes that are termed 'secondary use' or 'reuse' [2]. Using EHR data in research is one such type of secondary use [3, 4].

Safe Havens are secure environments that have been widely used to support access to EHRs for research whilst protecting patient identity and privacy [5, 6]. The four Safe Havens collaborating as part of the UK-wide Farr Institute were described by Lea *et al* [5] and were found to have different processes, controls and environments. In Scotland, a network of five Safe Havens has been established to support EHR reuse and over the past decade, have enabled researchers to access data at scale [6].

The Scottish Network of Safe Havens has been highly successful in supporting research. Over the last 5 years, the network has supported over a thousand research studies. There are a small number research and innovation projects, e.g. iCAIRD [7] and Research Data Scotland [8], which are collaborations across Safe Havens. The majority of research projects, however, are delivered by a single Safe Haven. Each Safe Haven maintains and controls access to EHR data collected from their geographically local regions and therefore has detailed knowledge of these datasets. The exception in Scotland is the National Safe Haven (eDRIS) which holds national-level. Researchers generally only access either the breadth of the nationally held data, with high cohort coverage, which are collected at a Scottish level or the depth of the local clinical data which has more detailed information about persons/entities from the Regional Safe Havens.

Representatives from each Safe Haven within the network meet regularly, supported and chaired by the Scottish Government's Chief Scientist Office (CSO). The Safe Havens collaborate to develop and share best practices. The network is primarily funded on a cost-recovery basis by charging researchers for services, with some Safe Havens also receiving some core support from the National Health Serivce (NHS) Scotland Research and Development funds.

This study provides an analysis of the infrastructure, operations and features of each Safe Haven and assesses how these impact the interoperability and technical options to support multi-Safe Haven projects. We present how Safe Havens in Scotland have protected privacy, as well as facilitating the reuse of the EHR data.

1.1 What is a Safe Haven?

Safe Havens have evolved as a set of processes to support researchers accessing sensitive data in a streamlined and secure way maintaining patient confidentiality [5, 9, 10]. The term "Safe Haven" is a widely used term but can have a different meaning to different people and in different contexts. Barton *et al* [11] described in detail the origins and evolution of the term. A Safe Haven was defined as 'a repository in which useful but potentially sensitive data may be kept securely under governance and informatics systems that are fit-for-purpose and appropriately tailored to the nature of the data being maintained, and may be accessed and utilised by legitimate users undertaking work and research contributing to biomedicine, health and/or to the ongoing development of healthcare systems.'

In Scotland, the main goal of a Safe Haven [5, 6] is to make data (consented and unconsented, predominantly unconsented data) available in a safe and secure setting for approved research undertaken by approved researchers whilst mitigating risks such as re-identification of patients and unauthorised use. The controls Safe Havens implement assure data controllers and the public that the data provided for analysis and research is securely managed and accessed. This reduces the risk to organisations, such as the NHS, increasing their confidence in providing data for research purposes. It also enhances public confidence in the use of their health data (NHS data) for research in the UK. In addition, the standardised security across many research projects reduces the burden on individual research groups to build, maintain, fund, and document such environments themselves.

Researchers often require data from different sources. Data that might need to be linked to healthcare data include social care, justice, education, research datasets or another NHS dataset. Data linkage is the process of linking any two or more datasets on an individual or entity level [12].

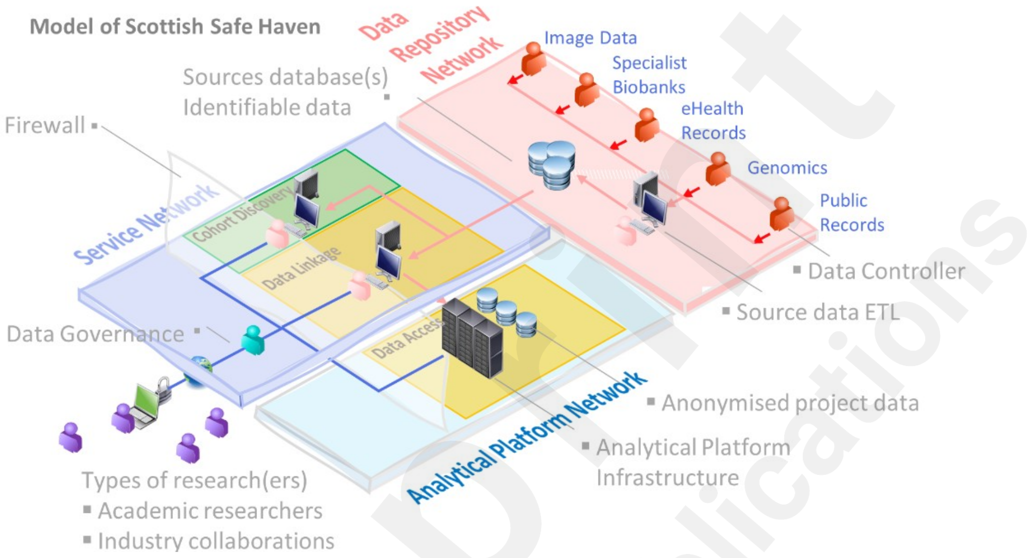


Figure 1: Model of Scottish Safe Havens. Researchers have access to the Safe Haven application process after data governance approvals. Safe Haven staff link and de-identify data and make them available in the analytical platform for researchers to analyse. Please refer to Table 1 for Service Network, Analytical Platform Network and Data Repository Network detail of each Scottish Safe Haven.

Figure 1 provides a model of how Scottish Safe Havens are structured. We have identified that Scottish Safe Havens mainly offer the following services:

A Data Processor and/or Data Repository Management: Secure handling and linking of data from multiple sources and possible hosting/managing of longitudinal data (detailed information about persons or entities, such as conditions, hospital admissions, prescription data etc). Scottish Safe Havens can also provide the function of a “Trusted Third Party” [6, 13]. They can support linkage of identifiable information where the roles of “indexer” and “linker” (see detail definition in Section 2.3) are separated so that no single organisation or individual has visibility of another organisation's identifiable data linked to their descriptive data [14, 15]. Safe Havens function as Data Processors [6] for any given dataset, and agree on terms with each Data Controller (Safe Havens can also be the Data Controller) to ensure that activities are centrally logged, monitored and audited [6].***Analytical Platform:*** An analytical platform is a highly secure, high-performance computing virtual and high-performance environment that enables researchers to securely analyse data without the row-level de-identified data leaving the environment (only aggregate level results can be exported). Strict governance and controls are implemented to ensure data security in the analytical platform. ***Research Support:*** The Safe Haven coordinators provide support to researchers navigating the data requirements, permissions landscape and provide a review mechanism to share the lessons from one project to the next. Some Scottish Safe Havens provide support for analysis.

Internal Safe Haven data scientists can help the research group with statistical analysis.

The term “Safe Haven” is defined here to mean the overarching service that combines the above services: A data processor and/or data repository, an analytics platform and research support.

The Scottish Safe Havens follow the ‘Five Safes’ principles of a Trusted Research Environment (TRE): Safe People, Safe Project, Safe Setting, Safe Data and Safe Output [16] as described within the Health Data Research (HDR) UK green paper [16].

	SAFE HAVEN				
Function	**eDRIS (national)**	**DaSH**	**Glasgow SH**	**HIC**	**Lothian/ DataLoch[a]**
General and Data Governance					
SafeHaven affiliation	PHS	UoA/NHS	NHS	UoD/NHS	NHS
Analytical Platform affiliation	UoE (EPCC)	UoA	UoG (RCB)	UoD	UoE (EPCC)
Network for SH services (cohort building and linkage)	NHSnet/ EPCC	NHSnet	NHSG/ NHSnet	NHS	NHSL/ NHSnet
Network for Analytical Platform	UoE/Janet	UoA/Janet	UoG/Janet	UoD/Janet & Secure Public Cloud	UoE/Janet
Data Repository Network	NHSnet/EPCC	NHSnet	NHSnet	NHSnet	NHSnet
Geographical Region [b]	Scotland	NHS Grampian	West of Scotland	NHS Tayside and Fife	Lothian/SE of Scotland
Population [c]	5.7 M	600 K	1.2 M	850 K	900 K
Active projects in 2020	>600	>40	>100	>100	>20
Controller(s)	PHS + NRS + Scottish Government	Original data sources	Original data sources	Original data sources	Original data sources
Processor(s)	eDRIS	DaSH	Glasgow SH	HIC	Lothian/ DataLoch
Governance Committee	HSC PBPP and Stats PBPP	NNPAC	Privacy Advisory Committee	HIC Governance Committee	Data Access Committee
Data Discovery / Metadata					
Feasibility	Manual/NDC	Manual/Local docs	Manual/Local docs/TriNetX[d]	Manual/ using RDMP [13] automation	Manual/ Local docs
Metadata provided with project extracts	No	Yes/Standard (Workflow)	Yes/Bespoke	Yes/Standard (RDMP)	Yes/Bespoke
Phenotype/cohort development	ICD code from user	By User	Locally stored algorithms/User	By User	By User/ Caliber Library
Data Linkage and de-de-identification					
Indexer	External (PHS for health data)	Internal	Internal	Internal (RDMP)	Internal
De-identification method	Workflow [17]	SQL procedure	Database Views (Usually SQL)	Workflow (RDMP)	SQL procedure
CHI seeding	NSS/CHILI	CHILI/ Internal	Internal	Internal	CHILI
Analytical Platform					
Archival	NHSnet & UoE (EPCC)	UoA	NHSnet &UoG (RCB)	NHSnet & UoD & Secure Public Cloud	NHSnet & UoE (EPCC)
Project data-content standards	As source	As source/ICD	As source	As source	As source
Project data format standards	CSV	SPSS/Stata/ CSV	CSV	CSV/Database	CSV
Data Repository					
Data repository No.	85+	1	1	1	One each

Data repository ownership	No	Yes	Yes	Yes	Yes
Source data metadata	National Data Catalogue	Internal shared files	Internal shared files	RDMP	Data dictionaries
Metadata publicly available	Yes	No	No	Yes	Yes[e]
Number of datasets available	85	17	200+	163	12
Source data ETL	Data management team PHS	Internal (SQL & Python)	BI in NHS Glasgow	Internal (RDMP)	Internal (SQL & Python)
Repository uses CDM	No (proprietary)	No (proprietary)	No (proprietary)	No (proprietary)	No (proprietary)

Table 1: A summary table of Safe Haven propterties

Note: PHS: Public Health Scotland; UoA: University of Aberdeen; NHS: National Health Service; UoD: University of Dundee; UoE: University of Edinburgh; EPCC: Edinburgh Parallel Computing Centre; UoG: Univeristy of Glasgow; RCB: Robertson Centre For Biostatistics; PBPP: Public Benefit And Privacy Panel; HSC PBPP: Public Benefit and Privacy Panel for Health and Social Care; Stats PBPP: Statistics Public Benefit and Privacy Panel; NNPAC: North Node Privacy Advisory Committee; NDC: National Data Catalogue; RDMP: Research Data Management Platform; ICD: International Statistical Classification of Diseases and Related Health Problems; NSS: National Services Scotland; CHILI: CHI Linkage Team; BI: Business Intelligence and Informatics;

[a] When this work was done Lothian Safe Haven and DataLoch were separate (though closely partnered). Since 1st of April 2021, they merged into Lothian/DataLoch Safe Haven.

[b] Regional Safe Havens have governance to request regional Health Board data. For example, Glasgow Safe Haven can request West of Scotland Health Board data.

[c] Safe Havens have access to historic records for deceased patients, which can increase the accessible data.

[d] TriNetX is a health research network tool that connects to assist drug discovery by helping pharmaceutical companies access clinical data. Glasgow Safe Haven has a TriNetX node. For data mapped into TriNetX tool, their study feasibility can be done using TriNetX.

[e] COVID 19 data dictionary is on DataLoch website.

1.2 Scottish Federated Network of Safe Havens

The network of five Safe Havens operating in Scotland is accredited by the Scottish Government and each Safe Haven adheres to the Scottish Safe Haven Charter [6] Each offers the three services, described in Section 1.1, with different data access procedures (subject to necessary local governance approvals) applied to different data sources and with different standard operating procedures (SOPs). There are four Regional Safe Havens and one National Safe Haven. There is a Regional Safe Haven for each Research and Development (R&D) node of the NHS supported by the Scottish Government's Chief Scientists Office [18]. They are provisioned by partnerships between the NHS Boards within each R&D node and a leading University from the region. Whether the primary contact organisation for a Safe Haven is an NHS board or a University differs between regional Safe Havens (Table 1, rows 'SH affiliation' and 'Analytical Platform Affiliation' show the composition of the partnerships for each Scottish Safe Haven). eDRIS [19], part of Public Health Scotland (PHS) [20], commissions Edinburgh Parallel Computing Centre (EPCC) [21], University of Edinburgh to provide the National Safe Haven. Grampian Data Safe Haven (DaSH) [22], a collaboration between the University of Aberdeen and NHS Grampian, is the Safe Haven for the Grampian region encompassing Aberdeen City, Aberdeenshire and Moray. The Health Informatics Centre (HIC) [23] at the University of Dundee is covering the Tayside and Fife regions. The Glasgow [24] and Lothian/ DataLoch Safe Havens [25, 26] are led by the NHS, covering the west of Scotland, the Edinburgh and the South East region, working in collaboration with Glasgow and Edinburgh Universities, respectively.

1.3 Scottish NHS Data Sources

Scotland has a single healthcare provider (NHS Scotland) and world-leading national health linked data assets from birth to death. In high-level summary, the National Safe Haven has direct access to health administrative data with high cohort coverage collected at a Scottish National level, and the Regional Safe Havens have direct access to more detailed health data from clinical systems.

Regional Safe Havens can work closely with local data custodians, which gives them easy access to additional data sources that are not routinely held, for example, other health data, educational data or police data. Access to these other sources of data may require additional time, due to different access processes and governance approvals.

The Research Data Scotland initiative [27] has been set up to streamline and support access to linked health and administrative datasets across the country.

National level NHS data:

PHS collect national level NHS [28] and adminstractive data to provide health information, health intelligence, statistical services and advice to support the NHS in progressing quality improvement in health and care, facilitate robust planning, and decision making. These datasets can be accessed in the National Safe Haven. Each Health Board across Scotland provides a regular update of a subset of their identifiable administrative data to PHS. This is standardised by PHS to create homogenous data within several National databases. Such data includes Scottish Morbidity Records (SMR) and Community-dispensed prescriptions. SMR data covers several different datasets such as SMR00 (Hospital Outpatient), SMR01 (Acute Stay Hospital Admissions), SMR02 (Maternity), SMR04 (Psychiatric Returns), SMR06 (Cancer Registry), SMR11 (Neonatal) and SMR25 (Substance Misuse). National Records Scotland records births, marriages and deaths.

Prescription data is collected nationally in two different ways. Through the ePharmacy system [29], prescriptions written by General practitioners (GPs), are captured directly in the system. The long-standing Data Capture Validation and Pricing (DCVP) Paid system [30] is used in Scotland to capture dispensing data which determines remuneration for community pharmacies. The ‘watermarked’ prescriptions go from GP to the patient then to a pharmacy, and are then collected and transferred monthly to DCVP/PHS for automated processing [31].

Regional level NHS data:

The Regional Safe Havens all receive a subset of the data from the National standardised datasets (SMR, Prescribing data, etc.) from PHS which includes only the patients who are resident / received healthcare within the relevant boards. They also have access to the deeply phenotyped data which is captured within local clinical systems but not collected at a national level. For example, they have access to microbiology data, virology data, laboratory test data, stroke data, echocardiology data etc. The type and level of local data available differs between Safe Havens. Individuals in Scotland are assigned a Community Health Index (CHI) number [32] when they first interact with the Health Service which is retained within their EHRs as much as possible throughout their health history. Regional Safe Havens use CHIs to link datasets to the Nationally captured records for the population within their region.

Research data:

In addition to unconsented access to routinely collected administrative/clinical records, the National and Regional Safe Havens can also host or manage researcher-collected consented datasets from many sources such as clinical trials, patient questionnaires etc. Compared with routinely collected EHRs, the research data often covers a narrower spectrum but provides more detailed information about the individual. Participants in research cohorts are volunteers who have consented to data access rules approved by ethics at the outset of the study. For example, Generation Scotland (GS) [33] is a resource of human biological samples and data which are available for medical research to create more effective treatments based on gene knowledge for the health, social and economic benefit of Scotland and its people. Another example is the SHARE [34] cohort which has consented over 280,000 individuals recruited to allow genotyping of any remaining blood samples after routine tests and applied to research upon their health data [34].

Regional Safe Havens also host disease-specific study data [35-40]. The data within these studies can be collected from a range of sources clinical data, patient surveys and routinely collected EHR data. Some disease registries where originally created at a regional level but were then rolled out Nationally. For example, the SCI-Diabetes [41] disease registry was formed by curating and linking routinely collected data from the Tayside Region. It was later developed into a nation-wide resource which now also collects patient reported out-come data. The data collected in SCI-Diabetes is used in clinical care. Extracts from SCI-Diabetes can also be linked on a study-by-study basis for research studies by the regional or the National Safe Havens.

Safe Havens can link research data to routinely collected administrative/clinical records and provide access to the combined data (in a de-identified form) within a analytical platform for analysis.

2 Scottish Safe Havens

Each Scottish Safe Haven has their data repository hosted on the NHS network (as shown in the 'Data Repository Network' row in Table 1), except for the National Safe Haven which also hosts some data within EPCC on a secure University environment. Safe Havens have data-sharing agreements with multiple data controllers and regularly receive new data from them.

All Safe Havens have committed to an approach to data access based around analytical platforms. Each Safe Haven has either established or has access to a analytical platform.

There are differences between Safe Havens in how they achieve the three main services described in Section 1.1. Table 1 summarises each Scottish Safe Haven. In the following section, we discuss the Safe Havens' in detail and their common deployment features.

2.1 Data governance and workflow

The governance approval step looks into aspects of the project such as ethics, peer review, funding source, public benefit and adherence to the 'Five Safes' described in section 1. The governance approvals process varies from Safe Haven to Safe Haven. Even for the same Safe Haven, different projects may require different governance approval processes in order to satisfy the different Data Controllers. However, the governance process for a standard de-identified project where data is accessed in an analytical platform is relatively streamlined. Each Safe Haven has a delegated governance authority/committee, as shown in Table 1, which includes representatives from the sponsor, ethics, lay members and NHS board to streamline the governance process and is relatively fast. For the DaSH Safe Haven, projects with local researchers using local data can obtain governance approval through their NNPAC process. The HIC Safe Haven's'standard projects' (where de-identified data is analysed by approved academic researchers within the analytical platform and the activity is funded by a peer-reviewed research grant) are covered by a blanket governance approval. The list of the standard projects supported is provided to the relevant governance committee for information rather than a request for approval for each one prior to the research commencing. For Lothian/DataLoch, projects involving de-identified data within an analytical platform can obtain governance approval through their local Data Access Committee process, which includes delegated Caldicott review.

Most Safe Haven responses to project requests adhere to a standard set of processes e.g. de-identified linked data is provided within a analytical platform for academic research. In exceptional circumstances, some projects require a different model and such exceptions need to be justified to obtain governance approvals. Example exceptions include: (a) the prepared project data would not be placed in a analytical platform; (b) the project data has some identifiable information.

To work on identifiable EHR data within a data repository, Safe Haven staff members are either NHS employees or have honorary NHS contracts. Safe Havens all have rule-based segregation of the team, specifying those with and without access to identifiable data. Only a handful of people in each Safe Haven can access the NHS network and see identifiable data. Other data sources (e.g. administrative data generated by the government or research data generated by research institutions) can be linked to EHRs. The linkage is performed by the Safe Haven data linkage team, and the linked data sets are then hosted in the analytical platform for the approved researchers/investigators to access. At each stage there is an oversight step to ensure all procedures were correctly followed and no unintentional identifiable data is released.

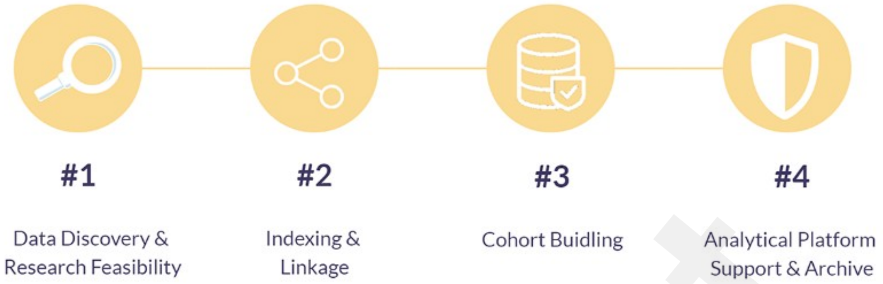

#1
Data Discovery & Research Feasibility
#2
Indexing & Linkage
#3
Cohort Buidling
#4
Analytical Platform Support & Archive

Figure 2: Safe Haven Project Workflow describes the stages a Safe Haven takes to support a typical project. (1) data discovery and research feasibility, users will initialise the application on the data governance aspects; (2) (Optionally) index and link a research dataset or administrative/clinical dataset for hosting at a given analytical platform; (3) cohort building the selected/agreed data from Safe Haven datasets; (4) transfer of extracted data to analytical platform after the data governance has been checked; user analyses analytical platform dataset. The project dataset is archived at the end of the project.

The project workflow for a data request is consistent across the Safe Haven network as shown in Figure 2. In the first step, the Safe Haven team runs research feasibility queries to identify data needed for the research topic. Once funding and governance is in place, data linkage is conducted (as required). Data extracted from the NHS network is de-identified, validated and assessed for disclosure before being released into the analytical platform. Details of linking and de-identification are given in Section 2.3 and Section 2.4. Section 2.5 provides discussions on the analytical platform support and data support. The archiving procedure of each Safe Haven and the infrastructures of each Safe Havens' data repository are discussed in Section 2.6.

2.2 Data discovery and metadata

Research feasibility analysis and data discovery remain a manual process involving discussions between researchers and the Safe Haven teams. During the project planning stage, researchers contact the relevant Safe Haven by email/phone call. Data discovery and research feasibility is done by document exchange and/or a face to face/virtual meeting. Safe Havens in Scotland require a meeting to capture the requirements of each study as well as guiding the governance process. Research feasibility is conducted by the Safe Haven by generating aggregate numbers for cohort/sub-cohort sizes based on the requirements defined by the researcher(s). For example, how many people are there in the data with diabetes who are over the age of 65 and who regularly have a prescription for insulin?

Researchers normally specify a phenotype or public phenotype algorithms to identify the correct cohort for their study. As shown in Table 1, no common standard procedure exists among the five Safe Havens to capture and reuse phenotype algorithms. However, DataLoch also uses the CALIBER phenotype library [42] while the Glasgow Safe Haven uses a suite of local matrix files storage phenotype algorithms, based on standard or published methods, which have been quality checked by clinicians. eDRIS, as they mainly work on national datasets using the ICD standard [43], usually agree what the ICD codes are with the researcher and conduct the cohort building using the codes as well as any date or other constraints given by the researcher. The remaining Safe Havens – HIC and DaSH - normally rely on the researcher to define the cohort themselves, where researchers have the choice of phenotype definition, for example, CALIBER phenotypes or ICD codes. Cohort identification is sometimes an iterative process between researchers and the Safe Haven team where a data constraint is applied, the impact on cohort size is observed and the constraint adjusted to optimise the cohort.

At the national-level PHS produces a National Data Catalogue (NDC) [44] as a single definitive resource of information on Scottish Health and Social Care datasets to assist cohort discovery.

Metadata provides the semantics associated with the Safe Haven datasets. There is, limited visibility of the metadata and data provenance available from Regional Safe Havens. The majority of the Safe Havens list the names of their easily accessible databases online [20, 22, 24-26, 45, 46] and provide researchers with a brief overview of the most commonly accessed datasets. Both HIC and eDRIS add their metadata and datasets to the HDR Innovation Gateway [47]. There is no common structure in EHR data storage across the healthcare system in Scotland. Since only a limited number of data scientists/analyst have experience in handling NHS data, this lack of visibility of metadata and data provenance can lead to a gap in understanding by data scientists/analyst about what data is available. Some Safe Havens will only release detailed metadata once they have an initial understanding of the project’s needs. There are multiple initiatives both internal [7, 48, 49] and external [8, 47] looking to improve the metadata visibility within Scottish Network of Safe Havens.

Research projects benefit from having clinical investigators who are familiar with NHS data or data scientists/analysts who have previous experience of working on Safe Haven projects within their project team. Such individuals can help to identify what data are available and to advise and support the data scientists/analysts working on the project. The majority Safe Haven projects generally require a suitable sponsor with relevant expertise to take responsibility for the initiation and management of the project and to support the project as an ethical safeguard.

All five Safe Havens provide research projects with metadata at the field level once a project is funded, approved and data extracts provided for analysis. However, feasibility discussions will generally take place at the ‘conceptual’ level. For example, a cohort definition may involve a fasting glucose constraint. The Safe Haven team will confirm that such a constraint is possible without disclosing the precise fields. To avoid bias and to get researchers to articulate what they need, and what is available, this ‘conceptual’ level feasibility can be quite limiting. The same could be true of the data extract requested, delivering BMI rather than height and weight for example. However, data extraction and delivery is generally at the field level and field-level metadata is be provided to ensure researchers can perform their analyses.

Table 1 shows that eDRIS provides metadata details on its website metadata (like the Cribsheet [50] on SMR) which can be used by researchers to define the fields they need when applying for data access. The researchers using the Regional Safe Havens can also utilise this metadata information for the nationally standardised datasets which the Regional Safe Havens hold for the subset of their region e.g. SMR data. None of the Safe Havens provide non-field-based metadata, such as through an ontology.

The eDRIS Safe Haven does not provide bespoke metadata to the user when delivering the project data. DaSH and HIC Safe Havens have a standard workflow and delivery format for supplying project specific metadata to each project. In the Glasgow and Lothian/Dataloch Safe Havens, projects are provided with all available metadata and provenance information. No standard format is used, and so the information included varies from project to project.

Safe Havens have different approaches to storing metadata about datasets in their data repositories. For eDRIS, PHS updates and maintains the National Data Catalogue which contains all the metadata for national datasets. The HIC Safe Haven uses an in-house, open-source software tool called the Research Data Management Platform (RDMP) [13] for importing data into their servers. The RDMP generates consistently formatted metadata for imported datasets. Lothian/Dataloch Safe Haven provides ‘data dictionaries’, which include metadata, for all the datasets in their Relational Database Management System (RDBMS). The Glasgow Safe Haven and the DaSH Safe Haven have internal document spaces to host the metadata and provenance provided by various data sources, which get manually entered and updated by staff.The lack of standard procedures in the Scottish Safe Haven

Network, has resulted in the available metadata varying between datasets. Highly processed datasets which have gone through Extract, Transform and Load (ETL) procedures have field parameters and rules imposed on them. These datasets have rich metadata associated with them. The majority of clinical data, however, is inherently of variable quality, with poor coverage, inconsistent and missing fields. The dataset metadata does not typically inform the user of the data variability or quality issues in the original data.

2.3 Data linkage

To answer many research questions, data linkage is required to enrich information about a defined cohort. Some Safe Haven projects involve linking NHS data with non-NHS data. Figure 3 illustrates the Indexing and Linking services in the Scotland Safe Haven Network.

According to the Guiding Principles of Data Linkage [51], an Indexer is an “Individual (or body) who receives personal data from one or more Data Controllers and determines which records in each dataset relate to the same individual (or entity). The indexer creates a unique reference for each individual (or entity) and a corresponding key to allow the data from the different sources to be joined.” Thus, an Indexing service [51] returns a unique identifier for each individual given an input dataset of identifying information e.g. name, address, date of birth, and other operational identifiers (such as CHI number). The relationship between identifiers associated with multiple datasets is maintained by the Indexing service. The Indexing service does not have visibility of the descriptive data pertaining to any individual (also termed payload data) e.g. an individual’s hospital admission information.

A Linker/Linking Service is an “Individual (or body) who receives datasets from data controllers and links them together using a key created by the indexer.” [51]. In this way, only output identifiers from the indexer service are exposed to the linker; only the linking service and the researchers see the linked data [14].

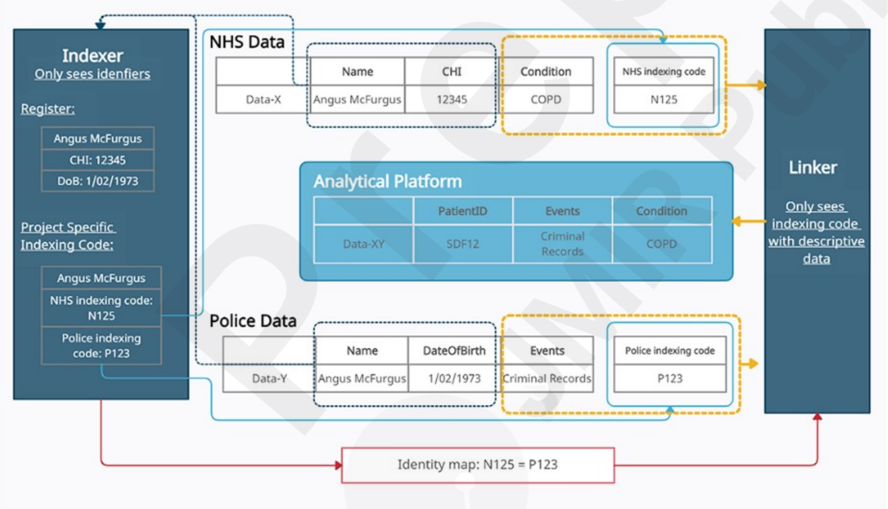


Figure 3: Data indexing and linking services in Scotland.

In Scotland, the NHS maintains the CHI. This is a patient identifier which concatenates a unique number, the person’s data of birth and their sex. CHI numbers are allocated at birth or on the first contact with the NHS in Scotland [32, 52]. Linking health data to other data where both datasets already contain CHI number is straightforward. All four Regional Safe Havens do this when preparing data for placement in a analytical platform using either software tools (HIC use RDMP [13] for example) or RDBMS user interfaces.

The National Safe Haven, eDRIS, established a data indexing and linkage procedure [17], the input identifiers are personal identifiers and the output identifier is an anonymised ID. eDRIS only receives data from providers with anonymised IDs and acts as Linker, placing the integrated data into a secure environment.

CHI Seeding

When linking to a source dataset that does not have CHI numbers but features other identifiers, the indexing team will use ‘Probabilistic Matching’ against the Population Spine. Related to the CHI, the Population Spine [17] contains the personal identifiers of all individuals in Scotland who have been in contact with NHS Scotland. This process of matching source datasets to the Population Spine is known as ‘CHI seeding’. The recent seeding of regional social care systems with CHI is an example of this. CHI seeding is also important for historical data analysis for EHRs prior to the introduction of CHI indexing. In Scotland, two teams provide national-level CHI seeding using probability-matching: the National Records Scotland (NRS) Indexing Team and the Public Health Scotland (PHS) CHI Linkage Team (CHILI). When a research project only needs NHS data indexing would be carried out by CHILI.

Both eDRIS and Lothian Safe Havens rely on NRS/CHILI for CHI seeding. DaSH Safe Haven provides CHI seeding through the NRS Scotland indexing team; they will only do it themselves when they have specific personal identifiers available such as patient name and the dataset consists of only a small amount of local Grampian patient records (approximately 500 people). Glasgow and HIC Safe Havens have a more established probabilistic matching routine developed, and normally do CHI seeding themselves. HIC has worked with local authorities to CHI seed their non-health data to be able to link it to health data.

2.4 Data de-identification

De-identification is undertaken before a Safe Haven provides the data to a analytical platform for the researcher to access. De-identification replaces information that could identify an individual in a data set with a study identifier (ID) for that individual and specific to that study, or dilutes the identifier to remove its individual nature. In linked, de-identified datasets the study ID is the same across data sources enabling researchers to link these data sources and understand which data corresponds to the same individual within that study, but without knowing their identity. This also means that de-identified IDs are unique to that project so the same individual will have different IDs in different projects.

In general, Safe Havens apply consistent rules to identifiable data fields. Customising de-identification rules based on the bespoke project requirements, governance approvals and the variety of datasets can be accommodated. The treatment of identifiers depends on the project’s specific justification following data minimisation principles [51]. For example, a date of birth can be processed to the 1st of the month or can be replaced with ‘age-at’, or it can be removed if it is not considered necessary for the analysis. A postcode can be replaced with a deprivation score, a SIMD rank [53], or it can be removed from the data. For biometric data, where, for example, weight and height of the individual are included, Safe Havens often put such values into ranges. Each Safe Haven follows Standard Operation Procedures (SOP) for reproducibility, consistency and error reduction.

All Safe Havens indicated that they find it challenging to de-identify clinical reports and other documents containing free text, which often contain personal identifiers such as phone numbers and names. Safe Havens often exclude entire fields from research extracts when they are not confident that such fields are safe to release. iCAIRD [7] is using hidden in plain sight techniques for identifiable data on images. eDRIS has developed algorithms to remove Personal Identifying Information (PII) from the Dose Instructions in PIS (these can also extract structured information e.g. dose unit and frequency). As part of the Scottish Medical Imaging (SMI) service [54] and PICTURES [55] there is work in progress to de-identify and create metadata from the text written by radiologists on their findings. This uses Natural Language Processing (NLP) and the CogStack framework [56].

2.5 Data formats in the analytical platform

In general, Safe Havens make few changes to source data provided to researchers, these changes being limited to the process of de-identification. For example, there is no attempt to harmonise data through the transformation of diagnosis codes or drug codes, where significant versioning occurs in longitudinal data. Some Safe Havens do add derived data to datasets. Within HIC for example, these data derivations and transforms can be applied either within the Safe Haven, or at the point of extracting a de-identified research dataset. This is done using RDMP, an open-source solution that allows custom coding, or researcher created statistics package code to be executed in a repeatable and reproducible manner. When requested by the researcher, DaSH Safe Haven can provide Charlson Comorbidity Index [57] and Tonelli codes [58] alongside ICD codes. While data standards are not applied at data extraction and delivered to a analytical platform, standards are enforced for nationally captured datasets. A team in PHS works with the health boards and system suppliers to ensure the use of standards. e.g. SMR data must be structured in an agreed way and use agreed coding systems for content.

Safe Havens make their best efforts to accommodate the requirements of projects. However, software available in most analytical platforms is limited (Microsoft Office packages, SPSS [59], Stata [60], SAS [61], R [62] etc), and so the output data formats are also limited to R, Excel, SPSS files or Stata files. The exception is the recently launched HIC hybrid cloud-based, scalable analytical platform. This also includes the capability for software development, machine learning and artificial intelligence development including, for example, python [63], MATLAB [64], and a suite of tools within Jupyter Notebooks (Sagemaker instance) [65] such as TensorFlow. The environment is also being enhanced to support multi-omic data [66] analysis through pipelines, utilising tools such as Plink [67] and Nexflow [68] with resource scheduling through AWS Batch [69]. The analytical platform provides GPUs and high-performance computing capabilities.

For larger projects, where the number of rows is too high to manage in other formats, the HIC Safe Haven provides the data in an RDBMS in the analytical platform for use by researchers.

Researchers rely upon Safe Havens to archive the raw data and derived data products from their analysis since they're not permitted to export any of that data from an analytical platform. A research project may be archived for a period of between 5 and 30 years depending on regulations, researcher or funder requirements. Archival typically takes place using the analytical platform infrastructure. There can be significant costs for storing and securing large amounts of data and a policy for long-term archival is being jointly developed by the Scottish Safe Havens.

2.6 Data repository infrastructure

Safe Havens have their source EHRs on the NHS network which is transferred to the service network (where the cohort building and linkage takes place, please refer to Figure 1) when required. They create cohorts and associated data on the NHS infrastructure before the data goes through the Safe Haven functions of linkage, de-identification and transfer to analytical platform(s) for researchers to access. The exception is eDRIS who have some datasets managed securely on a university environment by EPCC.

The infrastructure and the ETL process for those data repositories vary between Safe Havens (Figure 4).

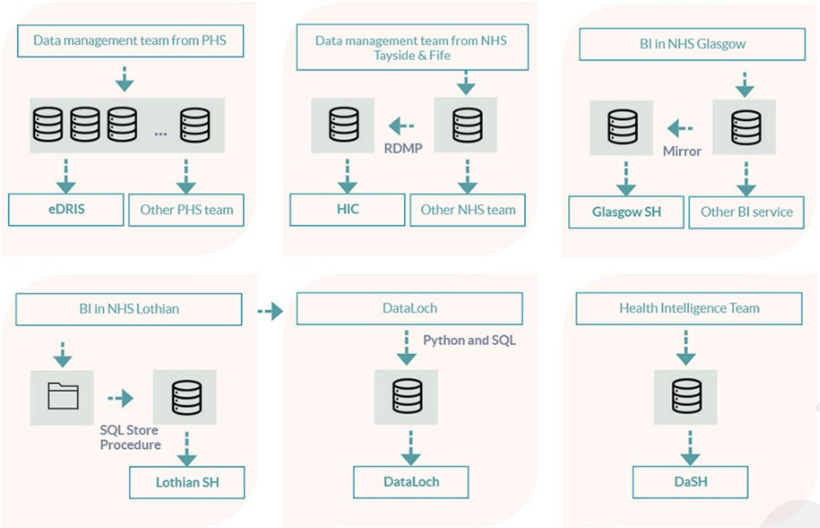


Figure 4: Safe Havens' data repository networks. Upper row from left to right: eDRIS, HIC and Glasgow Safe Haven. The lower row from left to right: Lothian, DataLoch and DaSH Safe Haven.

As shown in Table 1, 'Data Repository No.' row, eDRIS has access to 85 national NHS datasets; these are updated and maintained by PHS. There are datasets that eDRIS cannot access routinely but, for a known cohort, they can request data from other teams within PHS. The data management team within PHS perform quality assurance after ETL using R or SPSS (in cases of legacy data). As well as providing data to eDRIS, the data are also used to run hundreds of different reports and publications by other teams within PHS.

HIC's data repository infrastructure and NHS Tayside/Fife data are co-located within the same data centre. HIC run University of Dundee owned and managed servers connected to the NHS Tayside network and receive regular feeds of data from NHS Tayside clinical systems and from PHS (covering consented cohorts of research data and for the patients within the Tayside and Fife regions). The RDMP tool takes data from the feeds and performs ETL to clean and transform the data which are then stored within structured databases.

Glasgow Safe Haven's data repository mirrors some data from the routine data systems that are maintained by Business Intelligence and Informatics (BI) in NHS Glasgow. For custom NHS data or data collected for research projects, e.g. some SMR, PIS, all audit data, device data, trial data etc, Glasgow Safe Haven staff will conduct the ETL themselves.

Lothian/DataLoch data repositories residing on the NHS Lothian IT infrastructure use stored Python/ SQL to load data updates from PHS and data feeds via Business Intelligence and Informatics (BI) within NHS Lothian for copies of data from local clinical systems.

NHS Grampian's Health Intelligence Team updates the DaSH Safe Haven repository monthly. Both Lothian/DataLoch and DaSH Safe Haven deal with changing data formats by separating new and old data.

3 Discussion

Scotland has many strengths regarding enabling EHRs for reuse. There is a single National Health Service where patients are allocated CHI numbers which can be used to link their whole patient history. The Scottish Network of Safe Havens have similar architectures, adhere to the Scottish Safe Haven Charter [6], are accredited by the Scottish Government and ISO27001 [70, 71] is the common information security standard. Each Regional Safe Haven has a rich and deep data source from their local health boards, and the National Safe Haven has the breadth of a whole-population view and close links to other health and social care data sources. Scotland has many strengths regarding enabling EHRs for reuse. There is a single National Health Service where patients are allocated CHI numbers which can be used to link their whole patient history.

All the Safe Havens make use of two networks: 1) a analytical platform set up within university managed networks and 2) data repositories set up mainly on NHS networks. The existing operational separation of source data repository, linkage infrastructure and analytical platform, provides a solid foundation for increasing collaborative work across national and multi-Safe Haven projects.

There are some barriers, as highlighted Section 2, to making multi-Safe Haven projects as streamlined as possible. Addressing them in coordinated manner would pave the way to achieving a federated system of Safe Havens in Scotland. These opportunities for improvement include:

- *Data visibility:* The depth of the Scottish Data, which is hosted by Regional Safe Havens (described in Section 1.3), is not widely utilised by the wider community. These datasets are unique to each Regional Safe Haven and are difficult to bring up to a consistent national level. Interactions with researchers for feasibility, generating aggregate numbers, scoping projects, providing quotes for work can be resource intensive. Many data requests to Regional Safe Havens are from frequent users who know well the specific data structures and terminologies used by each Safe Haven. In making the regional data more visible and accessible, researchers will be more able to run projects using data from multiple Safe Havens.
- *Data standards and common data models:* As shown in Table 1, the Safe Havens accept data which uses any of number of standards. Due to the processing efficiency, the “create and destroy” model mandated by the Safe Haven Charter and the fact that researchers normally prefer to have the original data, there has been little attempt to harmonise extracted data for placement in analytical platforms. If the use of common data models such as OMOP [72] and i2b2 star schema [73] were used, either for data repositories or analytical platforms, the burden on multi-Safe Haven projects would be reduced and operational access to data would be faster and more predictable.
- *Governance:* In the Safe Haven Network, access to linked data is fragmented with researchers and healthcare providers having to work with Safe Havens to obtain local, regional or national data controller’s approval. Data governance in general is much easier at a local level. At Scottish national level, application forms for submission to the Public Benefit and Privacy Panel for Health and Social Care (HSC-PBPP) and Statistics Public Benefit and Privacy Panel (Stats PBPP) are normally required. This is a complex process and can take significant time for review and approval.

With the experience and knowledge gained from supporting projects requiring diverse local and regional datasets [74, 75]; and to build capability for a federated network, we propose that the following aspects of the network be addressed in future research:

- The establishment of a shared method for cataloguing and managing metadata would facilitate data discovery and research feasibility.
- To facilitate cross Safe Haven data governance, standardising the application interface specifications to Safe Havens would permit easier cross-access of Safe Havens by researchers.
- Healthcare delivery is explicitly devolved to local structures via Health and Social Care Partnerships in Scotland and associated legislation. With functions devolved to individual Health Boards, the linking of regional available data will require greater collaboration across the organisations and appropriate benefit shares.

4 Conclusions

The Safe Haven network in Scotland has supported over a thousand projects in the past 5 years, underpinning world class research outputs. It not only brought grant research, jobs and funding to Scotland but also enables international health research with many countries such as Brazil, India etc.

This paper reports an operational assessment of each of four Regional Safe Havens and the National Safe Haven. We compared a set of functions and services related to data forming part of EHRs in Scotland. We have described the operation of Scottish Safe Haven data services and functions and their technical implementation from the following points of view: (1) data governance and workflow; (2) data discovery and metadata; (3) data linkage; (4) data de-identification; (5) analytical platforms; and (6) data repository infrastructure. The results obtained should assist the Scottish Safe Havens to scale operations to larger cohorts and more diverse data, reduce timescales and operate more cost-effectively. More importantly, this work identified the responsibilities and work needed for each Scottish Safe Haven to contribute to the building of a national federated data-sharing platform. While this paper has focused on experiences across Scotland, the findings will be of interest nationally/internationally to inform understanding of the challenges that exist for the re-use of EHR data in clinical and other kinds of research.

Abbreviations

CHI	community health index
COPD	chronic obstructive pulmonary disease
DaSH	Grampian data Safe Haven
eDRIS	Scottish National Safe Haven
EHRs	electronic health records
ETL	extract, transform, load
GP	general practitioner
HIC	Health Informatics Centre
ICD	international statistical classification of diseases and related health problems
NDC	National Data Catalogue
RDBMS	relational database management system
NHS	national health service
PHS	Public Health Scotland
R&D	research and development
SMR	Scottish Morbidity Records
TRE	trusted research environment

References

1. WorldHealthOrganization, From innovation to implementation: Ehealth in the who european region. 2016: World Health Organization. Regional Office for Europe.
2. Geissbuhler A, Safran C, Buchan I, et al., Trustworthy reuse of health data: A transnational perspective. International journal of medical informatics, 2013. **82**(1): p. 1-9.
3. Doiron D, Raina P, and Fortier I, Linking canadian population health data: Maximizing the potential of cohort and administrative data. Canadian journal of public health, 2013. **104**(3): p. e258-e261.
4. Charities A o M R. A matter of life and death: How your health information can make a difference. https://www.amrc.org.uk/Handlers/Download.ashx?IDMF=652f0e60-315b-4258-8e81-fe1fdf8b65dc.
5. Lea N C, Nicholls J, Dobbs C, et al., Data safe havens and trust: Toward a common understanding of trusted research platforms for governing secure and ethical health research. JMIR medical informatics, 2016. **4**(2): p. e22.
6. Charter for safe havens in scotland: Handling unconsented data from national health service patient records to support research and statistics. https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-

support-research-statistics/.
7. iCAIRD. Homepage. https://icaird.com/.
8. Research-Data-Scotland. Homepage. https://www.researchdata.scot/.
9. Caldicott F, Information: To share or not to share? The information governance review, N.D.G.f.H.a.S. Care, Editor. 2013.
10. Thomas R and Walport M, Data sharing review report 2008.
11. Burton P R, Murtagh M J, Boyd A, et al., Data safe havens in health research and healthcare. Bioinformatics, 2015. **31**(20): p. 3241-3248.
12. ScottishGovernment. What is data linkage for research? https://www2.gov.scot/Topics/Statistics/datalinkageframework/Whatdatalinkageis.
13. Nind T, Galloway J, McAllister G, 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): p. giy060.
14. Jefferson E R and Trucco E, The challenges of assembling, maintaining and making available large data sets of clinical data for research, in Computational retinal image analysis. 2019, Elsevier. p. 429-444.
15. Joined-up data for better decisions: Guiding principles for data linkage. https://www.gov.scot/publications/joined-up-data-better-decisions-guiding-principles-data-linkage/.
16. HDRUK. Trusted research environments (tre) a strategy to build public trust and meet changing health data science needs. https://ukhealthdata.org/wp-content/uploads/2020/07/200723-Alliance-Board_Paper-E_TRE-Green-Paper.pdf.
17. EDRIS. Data linkage. https://www.isdscotland.org/Products-and-Services/eDRIS/FAQ-eDRIS/#e1.
18. ChiefScientistOffice. Homepage. https://www.cso.scot.nhs.uk/.
19. eDRIS. Products and services. https://www.isdscotland.org/Products-and-Services/eDRIS/.
20. PHS. Covid-19 pandemic response. https://www.isdscotland.org/Products-and-Services/eDRIS/.
21. EPCC. https://www.epcc.ed.ac.uk/.
22. DaSH. Grampian population platform: Datasets permissioned for access for research purposes via dash. https://www.abdn.ac.uk/iahs/facilities/grampian-population-platform-1475.php.
23. HIC. Health informatics centre - trusted research environment. https://www.dundee.ac.uk/hic.
24. NHS-Greater-Glasgow-and-Clyde. Glasgow safe haven services. https://www.nhsggc.org.uk/about-us/professional-support-sites/glasgow-safe-haven/services/.
25. Lothian-Research-Safe-Haven. Datasets. http://accord.scot/researcher-access-research-data-nrs-safe-haven/datasets.
26. DATALOCH. Covid-19 collaborative. https://www.ed.ac.uk/usher/dataloch/covid-19-collaborative.
27. RDS. Homepage. https://www.researchdata.scot/.
28. NationalDataCatalogue. National datasets. https://www.ndc.scot.nhs.uk/National-Datasets/.
29. INPS. Epharmacy user guide (scotland). http://www.inps.co.uk/sites/default/files/ePharmacy%20User%20Guide.pdf.
30. InformationServiceDivision. Glossary of terms: Practice level prescribing data. https://www.isdscotland.org/health-topics/prescribing-and-medicines/_docs/Open_Data_Glossary_of_Terms.pdf?1.
31. Alvarez-Madrazo S, McTaggart S, Nangle C, et al., Data resource profile: The scottish national prescribing information system (pis). International journal of epidemiology, 2016.

45(3): p. 714.

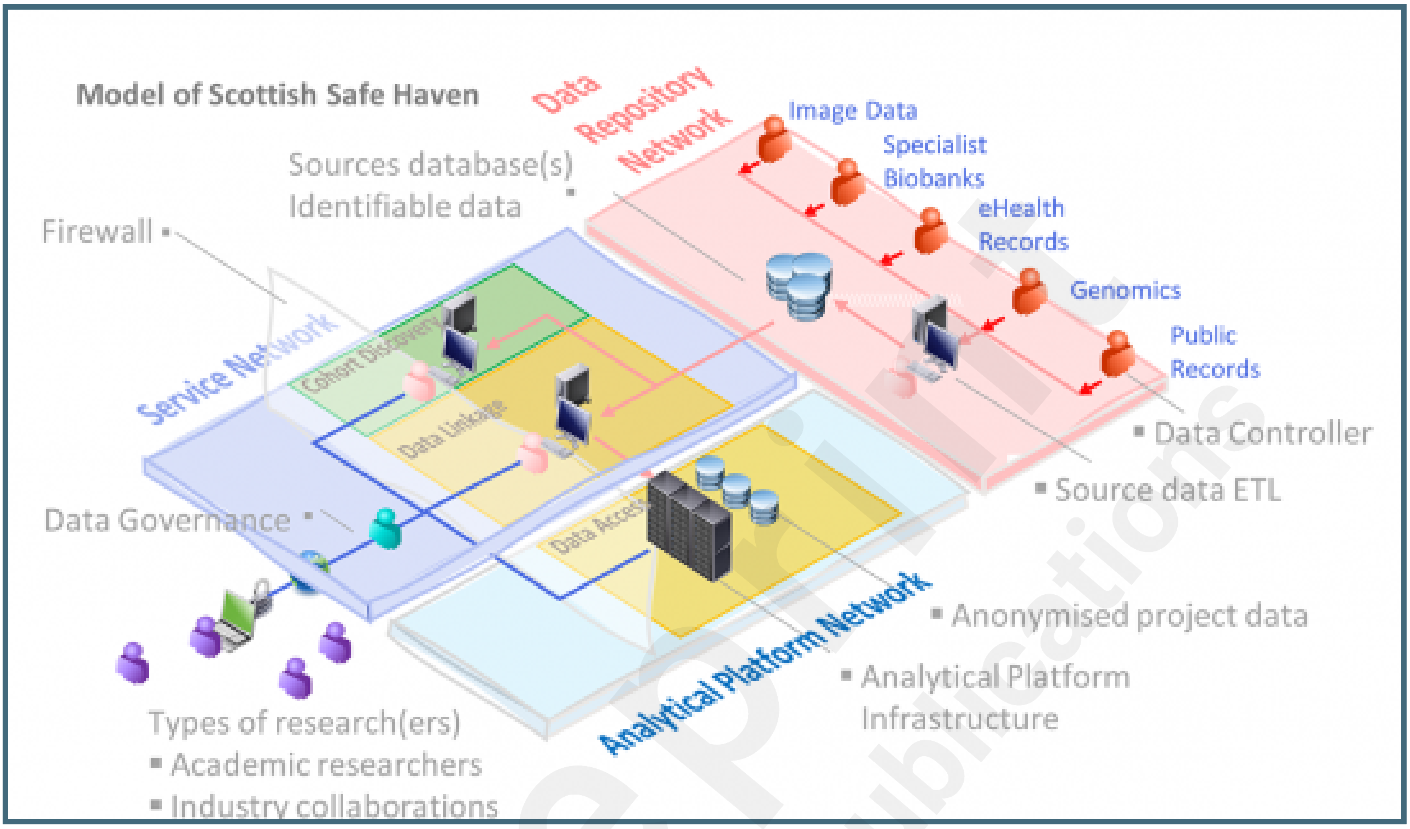
32. ISD. Chi number. http://www.ndcdev.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?Search=C&ID=128&Title=CHI%20Number#:~:text=The%20CHI%20number%20is%20a,and%20an%20arithmetical%20check%20digit.
33. GenerationScotland. Generation scotland home. https://www.ed.ac.uk/generation-scotland.
34. McKinstry B, Sullivan F M, Vasishta S, et al., Cohort profile: The scottish research register share. A register of people interested in research participation linked to nhs data sets. BMJ open, 2017. **7**(2).
35. EMBARC. The european bronchiectasis registry. https://www.bronchiectasis.eu/registry.
36. GoDARTS. Diabetes audit and research in tayside scotland. https://godarts.org/.
37. Childsmile. Childsmile – improving the oral health of children in scotland. http://www.child-smile.org.uk/.
38. Sawhney S, Marks A, Fluck N, et al., Intermediate and long-term outcomes of survivors of acute kidney injury episodes: A large population-based cohort study. American Journal of Kidney Diseases, 2017. **69**(1): p. 18-28.
39. Ayorinde A A, Wilde K, Lemon J, et al., Data resource profile: The aberdeen maternity and neonatal databank (amnd). International journal of epidemiology, 2016. **45**(2): p. 389-394.
40. AberdeenBirthCohorts. Children of the 1950s. https://www.abdn.ac.uk/birth-cohorts/1950s/.
41. SCI-Diabetes. Sci-diabetes home. https://www.sci-diabetes.scot.nhs.uk/.
42. Denaxas S, Gonzalez-Izquierdo A, Direk K, et al., Uk phenomics platform for developing and validating electronic health record phenotypes: Caliber. Journal of the American Medical Informatics Association, 2019. **26**(12): p. 1545-1559.
43. WHO. International statistical classification of diseases and related health problems (icd). https://www.who.int/standards/classifications/classification-of-diseases.
44. NationalDataCatalogue. National data catalogue. https://www.ndc.scot.nhs.uk/.
45. PHS. The national data catalogue (ndc). https://www.ndc.scot.nhs.uk/.
46. HIC. Dataset inventory. https://www.dundee.ac.uk/hic/datalinkageservice/datasetinventory/.
47. HDRUK, Specification for phase 2: Technology partnership. 2019.
48. Jefferson E, The safe haven metadata (shema) project. 2020, Chief Scientist Office: Scotland.
49. CentreforHealthDataScience. Wellcome trust data provenance. https://www.abdn.ac.uk/achds/research/data-provenance-172.php.
50. PHS. Smr crib sheets. https://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Crib-Sheets/.
51. ScottishGovernment, Joined-up data for better decisions: Guiding principles for data linkage.
52. ISD. Dictionary: Chi number. https://www.ndc.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?Search=C&ID=128&Title=CHI%20Number.
53. ScottishGovernment. Scottish index of multiple deprivation 2020. https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/#:~:text=SIMD%20ranks%20data%20zones%20from,deprived%20data%20zones%20in%20Scotland.
54. PHS. Scottish medical imaging (smi) service. https://beta.isdscotland.org/products-and-services/scottish-medical-imaging-smi-service/.
55. Nind T, Sutherland J, McAllister G, et al., An extensible big data software architecture managing a research resource of real-world clinical radiology data linked to other health data from the whole scottish population. GigaScience, 2020. **9**(10): p. giaa095.
56. Jackson R, Kartoglu I, Stringer C, et al., Cogstack-experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital. BMC medical informatics and decision making, 2018. **18**(1): p. 1-13.
57. Sundararajan V, Henderson T, Perry C, et al., New icd-10 version of the charlson comorbidity index predicted in-hospital mortality. Journal of clinical epidemiology, 2004. **57**(12): p. 1288-1294.

58. Tonelli M, Wiebe N, Fortin M, et al., Methods for identifying 30 chronic conditions: Application to administrative data. BMC medical informatics and decision making, 2016. **15**(1): p. 1-11.
59. IBM. Ibm spss statistics. https://www.ibm.com/uk-en/products/spss-statistics.
60. STATA. Home page. https://www.stata.com/.
61. SAS. Home page. https://www.sas.com/en_gb/home.html.
62. TheRFoundation. The r project for statistical computing. https://www.r-project.org/.
63. Python. Homepage. https://www.python.org/.
64. MATLAB. Homepage. https://www.mathworks.com/products/matlab.html.
65. Amazon. Use amazon sagemaker notebook instances. https://docs.aws.amazon.com/sagemaker/latest/dg/nbi.html.
66. Palsson B and Zengler K, The challenges of integrating multi-omic data sets. Nature chemical biology, 2010. **6**(11): p. 787-789.
67. plink. Homepage. https://zzz.bwh.harvard.edu/plink/.
68. nextflow. Homepage. https://www.nextflow.io/.
69. Amazon. Aws batch. https://aws.amazon.com/batch/.
70. Calder A and Watkins S G, Information security risk management for iso27001/iso27002. 2010: It Governance Ltd.
71. International-Organization-for-Standardization, Iso/iec 27001: 2013: Information technology--security techniques--information security management systems--requirements. 2013: International Organization for Standardization.
72. OHDSI. Omop common data model. https://www.ohdsi.org/data-standardization/the-common-data-model/.
73. Murphy S N, Weber G, Mendis M, et al., Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association, 2010. **17**(2): p. 124-130.
74. Shah A S, Anand A, Strachan F E, et al., High-sensitivity troponin in the evaluation of patients with suspected acute coronary syndrome: A stepped-wedge, cluster-randomised controlled trial. The Lancet, 2018. **392**(10151): p. 919-928.
75. Shah A S, Griffiths M, Lee K K, et al., High sensitivity cardiac troponin and the under-diagnosis of myocardial infarction in women: Prospective cohort study. bmj, 2015. **350**: p. g7873.

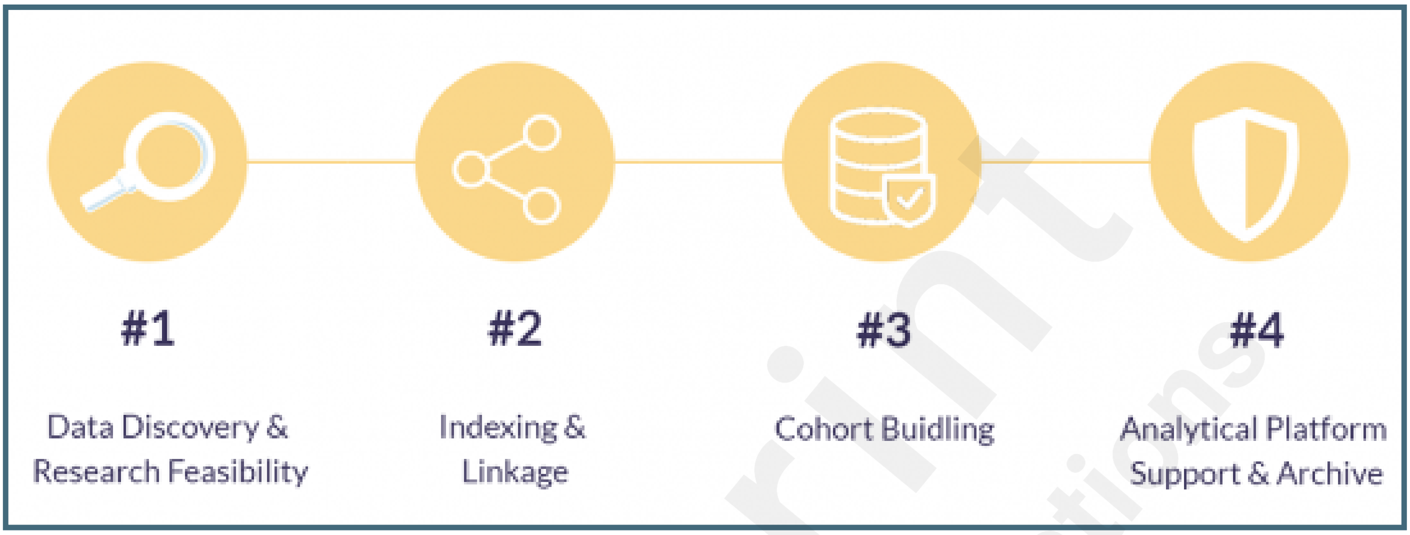
Supplementary Files

Figures

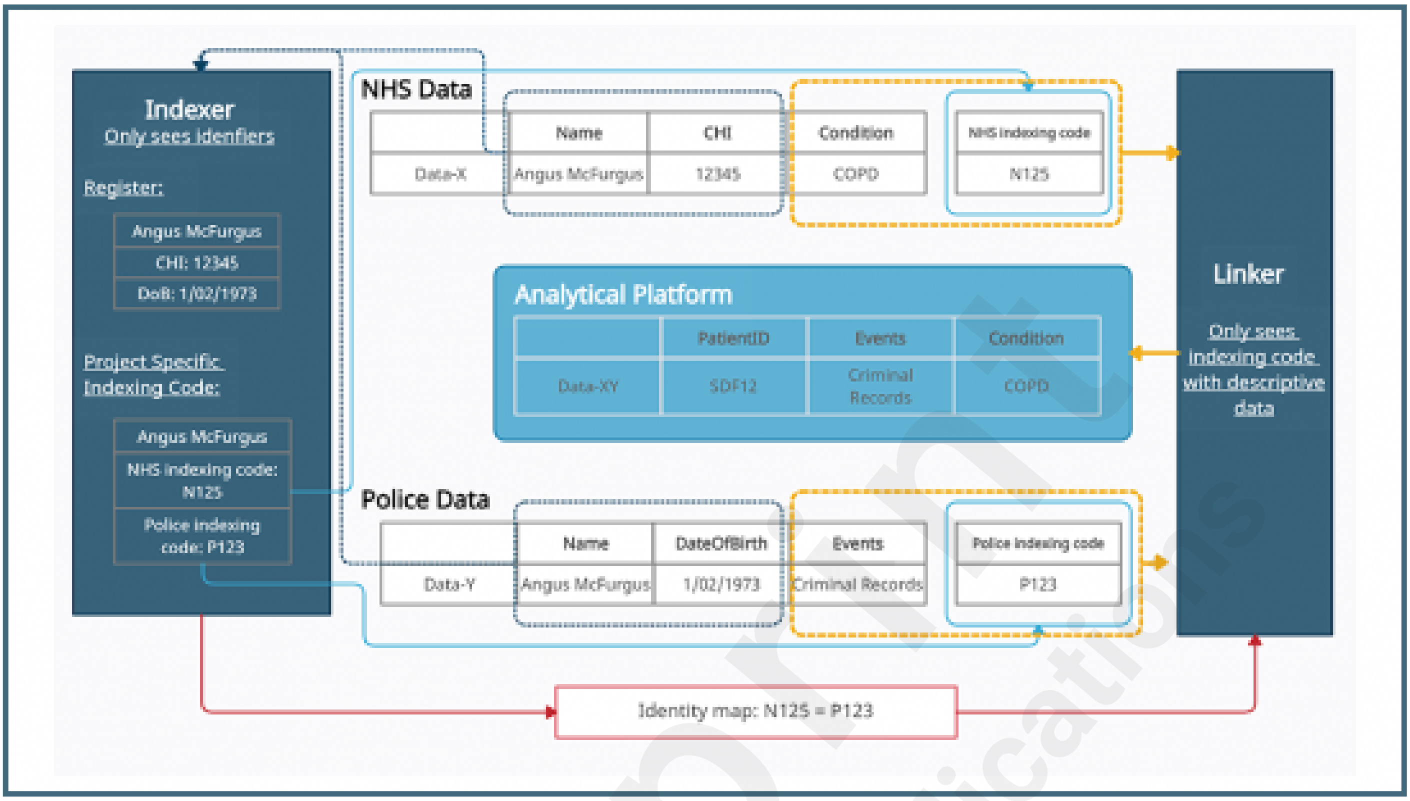
Model of Scottish Safe Havens. Researchers have access to the Safe Haven application process after data governance approvals. Safe Haven staff link and de-identify data and make them available in the analytical platform for researchers to analyse. Please refer to Table 1 for Service Network, Analytical Platform Network and Data Repository Network detail of each Scottish Safe Haven.



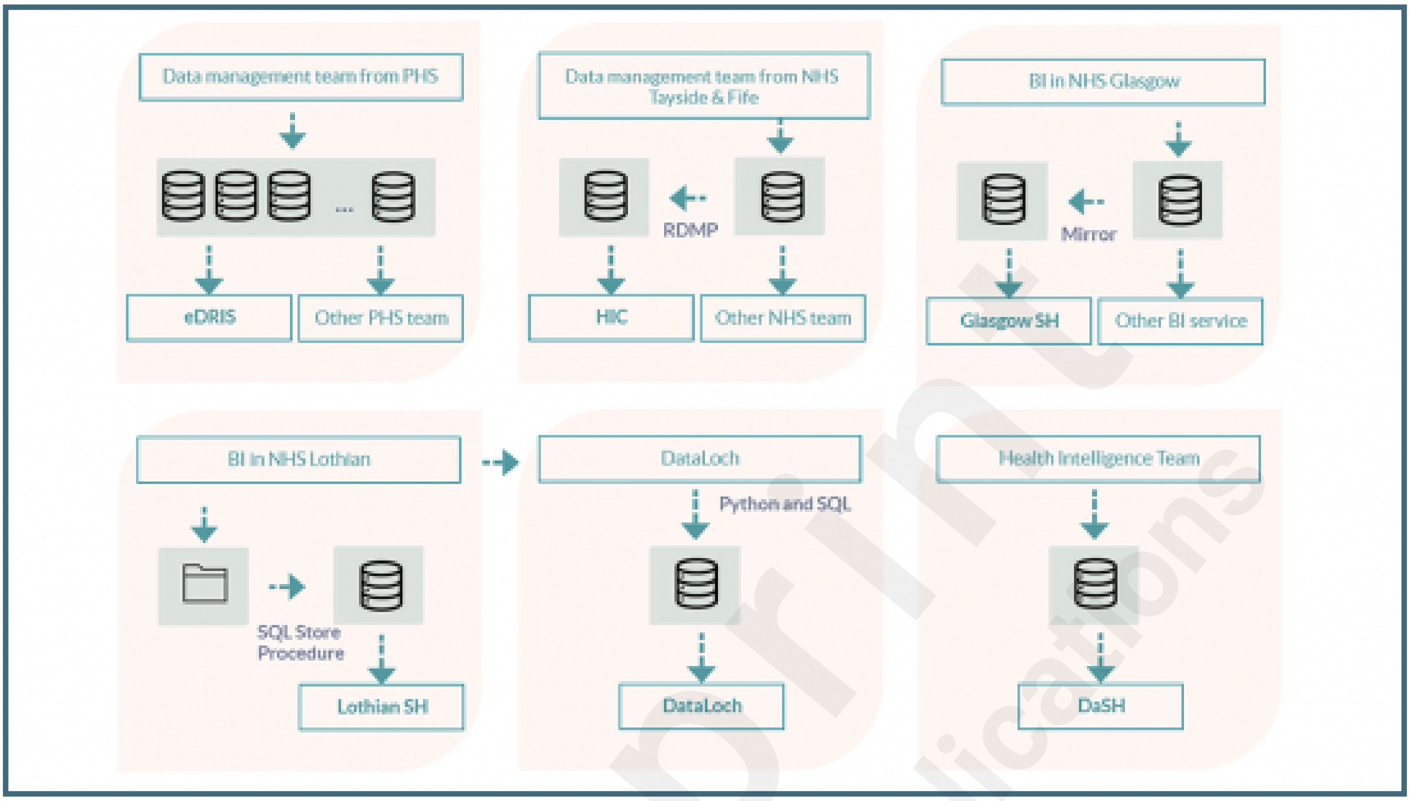
Safe Haven Project Workflow describes the stages a Safe Haven takes to support a typical project. (1) data discovery and research feasibility, users will initialise the application on the data governance aspects; (2) (Optionally) index and link a research dataset or administrative/clinical dataset for hosting at a given analytical platform; (3) cohort building the selected/agreed data from Safe Haven datasets; (4) transfer of extracted data to analytical platform after the data governance has been checked; user analyses analytical platform dataset. The project dataset is archived at the end of the project.


#1
Data Discovery & Research Feasibility
#2
Indexing & Linkage
#3
Cohort Buidling
#4
Analytical Platform Support & Archive

Data indexing and linking services in Scotland.



Safe Havens’ data repository networks. Upper row from left to right: eDRIS, HIC and Glasgow Safe Haven. The lower row from left to right: Lothian, DataLoch and DaSH Safe Haven.


Data management team from PHS
...
eDRIS
Other PHS team
Data management team from NHS Tayside & Fife
RDMP
HIC
Other NHS team
BI in NHS Glasgow
Mirror
Glasgow SH
Other BI service
BI in NHS Lothian
SQL Store Procedure
Lothian SH
DataLoch
Python and SQL
DataLoch
Health Intelligence Team
DaSH