International Journal of Population Data Science

Journal Website: www.ijpds.org

A profile of the Grampian Data Safe Haven, a regional Scottish safe haven for health and population data research

Katherine K. O'Sullivan[1,*] and Katie J. Wilde[1]

Submission History

Submitted:	30/06/2022
Accepted:	24/1/2023
Published:	16/03/2023

[1]Grampian DaSH, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen AB25 2ZD

Abstract

There has been a recent emphasis to establish and codify large-scale or national Trusted Research Environments (TREs) in the United Kingdom, with a view to limit smaller, local TREs. The basis for this argument is that it avoids duplication of infrastructure, information governance, privacy risks, monopolies and will promote innovation, particularly with commercial partners. However, the work around establishing TREs in the UK largely ignores the long-established local TRE landscape in Scotland, and the way in which local TREs can actually improve data quality, solve technical architecture challenges, promote information governance and risk minimisation, and encourage innovation and collaboration (both academic and commercial).

This data centre profile focuses on the Grampian Data Safe Haven (DaSH),a secure, virtual healthcare data analysis and storage centre located in Aberdeen, Scotland. DaSH was co-established by the NHS Grampian Health Board and University of Aberdeen to allow for the secure processing and linking of health data for the Grampian and Scottish population when it is not practicable to obtain consent from individual patients. As an established trusted research environment now in its 10th operating year, DaSH technology ensures healthcare, social care data and other types of sensitive data, routinely collected and used without individual patient consent, are made accessible for both academic research and clinical service evaluation and improvements whilst protecting individuals' privacy at the local, national and international levels. DaSH has registered almost 600 projects and facilitated over 200 distinct research projects with data hosting, extraction, and novel linkages to completion. Ongoing innovation and collaboration between DaSH and the NHS Grampian Health Board continues to expand researcher access to new types of data and data linkages, introduce new technologies for advanced statistical research methods, and supports interdisciplinary research using population health and social care data for research, clinical and commercial advancements, and real-world practitioner applications.

The purpose of this paper is to present DaSH's data population, operating model, architecture and information technology, governance, legislation and management, privacy-by-design principles and data access, data linkage methods, data sources, noteworthy research outputs, and further developments in order to demonstrate the value of local TREs within the data management and access debate.

Keywords

data centre; safe haven; data linkage; data extraction; health informatics; population data; Trusted Research Environment

*Corresponding Author:
Email Address: katherine.osullivan@abdn.ac.uk (Katherine O'Sullivan)

https://doi.org/10.23889/ijpds.v4i2.1817

Background

Within the United Kingdom, recent recommendations for accessing and analysing health data safely and transparently has centred on creating a limited number of Trusted Research Environments (hereafter, TREs) as a key priority for secure data access and linking health and population data on a large scale [1–3]. The argument put forth in Professor Ben Goldacre's report on TREs and data management (hereafter, 'Goldacre review'), for example, is to:

> build a small number of secure analytics platforms . . . then make these the norm for all analysis of NHS [National Health Service] patient records by academics, NHS analysts, and innovators, wherever there is any privacy risk to patients, unless those patients have consented to their data flowing elsewhere. Every new TRE brings a risk of duplicated effort, duplicated information governance, duplicated privacy risks, monopolies on access or task, and obstructive divergence around data curation and similar activities: there should be as few TREs as possible, with a strong culture of openness and re-use around all code and platforms (p.10).

The argument to limit the number of TREs is obviously compelling, seemingly solving concerns and challenges around infrastructure, scalability, competition, governance, privacy and data integrity—all complex issues that researchers, programmers, information governance specialists and technical architects would welcome. Moreover, the report implies that these concerns and challenges are the root cause of innovation blockers, with the 'small number' of TREs the only viable pathway to promote commercial innovation (p.41).

However, the recommendations in the Goldacre report also fail to acknowledge the realities of health and social care data in the UK as it currently exists: each Health Board or Trust has its own governance policies and nuanced interpretation of legislation [4]. Each Health Board or Trust will have its own interpretation of data protection laws and how they relate to patient-level data and a different focus or concerns around information governance and risk appetite for data use, particularly health and social care data used without individual consent in large-scale research. Each Health Board or Trust (and historically, individual hospitals within Health Boards or Trusts) procured its own systems that have been configured to its own needs.[1] There will be local policies or practices in place around data capture or data entry. There are local specialisms, particularly where clinicians engage in research. Finally, each Health Board or Trust may have its own research and innovations team, which will largely focus on the local strategic and/or operational priorities that address local pressures or local interests [6].

The arguments for limiting health data accessible via national or a limited number of large-scale TREs in the UK as a way to eliminate 'obstructive divergence around data curation and similar activities' ultimately suffers from the reality that healthcare (and social care) data in the UK is inherently local – captured by local GPs, hospitals and Local Authorities. Any proposal therefore requires local knowledge and expertise to understand and contextualise this local data, as well as a framework that promotes collaborative, large-scale data analysis whilst ensuring patient privacy and data security at the local level. This approach was taken by the Scottish Government and formalised in its Scottish Charter for Safe Havens in 2015 [7], setting out the principles and standards for the four regional Scottish safe havens and one national Scottish safe haven.[2] The four regional safe havens link an NHS research and development node to a local university, ensuring best practice in unconsented health data security, research and innovation to meet the needs of both clinicians and researchers, offering independent research platforms but crucially, working collaboratively through a federated model to harmonise data, streamline governance and ensure reproduceable analytic pipelines (RAP) across local datasets [9].

This paper presents a case study of the Grampian Data Safe Haven (DaSH), one of the four regional safe havens in Scotland, as an example of how local TREs are already working to the principles set out in the Goldacre review. DaSH was established in 2012 as a joint endeavour between the NHS Grampian Health Board (covering the region of Aberdeen City, Aberdeenshire and Moray, Scotland) and the University of Aberdeen to improve the security of extraction and linkage of routinely collected, unconsented NHS patient data for research purposes. DaSH guarantees the safe, secure and trustworthy processing of individual patients' sensitive data for health and social care research for public benefit as an accredited Scottish safe haven, accredited by the Scottish Government and independent ISO27001 accreditation. DaSH allows researchers to contribute to the public benefits of heath research through analysis of local, pseudonymised (e.g. de-identified), patient-level data whilst ensuring individual patient privacy and confidentiality are safeguarded by rigorous principles of data security, patient confidentiality and patient anonymity.

Founded on the principle of improving health and social care outcomes by providing a safe and secure environment enabling cutting-edge research, DaSH has become a centre of innovation in the north-east of Scotland, maintaining a balance between the legal and ethical requirements for the safe handling of confidential healthcare data, the public concerns over data sharing and linkage of personal, private data, and the public benefit of research conducted by university researchers, NHS clinicians as well as private (commercial) organisations [10–14]. DaSH brings together heath and population data from the NHS Grampian region and the University of Aberdeen's own holdings of large cohort datasets, providing robust longitudinal population data for researchers to investigate not only medical and clinical research, but also economic, educational and social factors in early life

[1] The National Health Service in the United Kingdom has a complex organisation across the four nations. Scotland and Wales are divided into regional Health Boards England is divided into NHS Trusts, and Northern Ireland has Health and Social Care Trusts [5].

[2] The four regional Safe Havens are: (1) Grampian Data Safe Haven (NHS Grampian and University of Aberdeen); (2) Glasgow Safe Haven (NHS Greater Glasgow and Clyde and University of Glasgow); (3) DataLoch (NHS Lothian and University of Edinburgh); (4) Health Informatics Centre (HIC) – NHS Tayside and University of Dundee; the national Safe haven is the electronic Data Research and Innovation (eDRIS). The precursor to the Charter for Safe Havens in Scotland was the ScottisH Informatics Programme (SHIP), founded in 2011 [8].

that affect outcomes and health later in life.[3] In recent years, DaSH has expanded its standard linkage support of text-based datasets to partner with commercial organisations developing and training artificial intelligence algorithms for use within the NHS. These projects not only involve AI to assist with reading text files, including clinical free text, but also employ the use of artificial intelligence diagnostics for medical imaging data. Whilst DaSH continues to produce critical core linkages to researchers, it is also expanding and evolving to provide greater extraction, linkage and analysis services that accelerates greater breadth and depth of medical research, improvements to clinical settings by empowering clinicians and practitioners to have access to and support with analytics, and to facilitate interdisciplinary research utilising linked data across healthcare and community settings.

In this paper, we provide a profile of the Grampian Data Safe Haven, its population setting, current operating model, architecture and information technology, data governance, linkage methods, data sources, impact, and planned future developments. The aim is to demonstrate the importance of local, smaller scale TREs within debates around data management and their contributions to innovations across digital infrastructure, information governance, data collection and curation, data harmonisation, and clinical, research and commercial innovation – and how well-established, local TREs have an important place within proposed frameworks of how unconsented health, social care and population data are stored, processed and accessed for research purposes.

Approach

Population setting

DaSH holds permissions to extract and link health and social care datasets from the NHS Grampian Health Board and University of Aberdeen [19]. The population of the NHS Grampian Health Board covers a half-million people in North East Scotland (approximately 10% of the Scottish population) residing in Aberdeen City, Aberdeenshire and Moray [20]. Aberdeen City is the third-most populated city in Scotland, though the remainder of Aberdeenshire and Moray is rural and agricultural [21, 22]. Because the data held by NHS Grampian is longitudinal, DaSH hold health records for nearly 1.4 million people.

Although the population of Grampian is relatively small compared to other parts of the UK, the attractiveness to researchers who use data from this region for research is that it reflects the UK as a whole, covering urban, suburban and rural communities, relative diversity across ethnicities and varying degrees of deprivation [23, 24]. The population split across these different socio-economic, geographic and ethnic groups enables researchers to analyse data from a multi-faceted approach, as opposed to TREs which may hold data that skews towards particular geographic, socio-economic or ethnic groups. Access to this population provides researchers with robust (but not overwhelming) cohorts on which to base their research, particularly as researchers are increasingly focused on wider health determinants and patient circumstances in their analyses.

Researchers accessing DaSH have the ability to work on both Grampian-specific population research, as well as hosted national and international datasets. In cross-population analysis, researchers link to Grampian-specific data for comparisons between other UK populations or other international populations. Increasingly, clinical and research projects are working with DaSH to provide scalable models to national and international projects. The PIONEER Big Data Platform [25] for example, will first analyse Grampian cohort data from a number of local datasets using DaSH's local infrastructure solutions using a reproduceable analytical pipeline (RAP) before applying this model to Scotland-wide data as well as other European data. This in turn will link to the project's multi-national consortium Observational Medical Outcomes Partnership (OMOP) to improve prostate cancer outcomes.

Operating model

The Grampian DaSH design and operation is unique to other regional Scottish safe havens. As a partnership between NHS Grampian and the University of Aberdeen, the design of DaSH operating across both NHS and University network reinforces the focus on collaboration and partnership between the two entities and the equal importance of health analytics across clinical and academic settings. DaSH staff hold dual appoints as employees of both the University of Aberdeen and NHS Grampian Health Board and have direct access to both NHS Grampian and University of Aberdeen networks for data extraction. Because they have access to patient-identifiable data and work across clinical and research settings, DaSH staff must go through extensive information security and information governance training, including the Medical Research Council's Information Research, GDPR and Confidentiality module, National Institute of Health and Care Research's Good Clinical Practice module, and the University of Aberdeen's Good Research Practice module, in addition to all NHS Grampian and University of Aberdeen information security and information governance modules.

The dual appointment across university and NHS allows for DaSH data analysts to move between both environments to provide capacity management as well as knowledge sharing to improve clinical service delivery. For example, during the Covid-19 pandemic saw the Grampian DaSH data analysts produce the Grampian region Covid shielding list whilst NHS data analysts concentrated on hospital capacity management and regional case and vaccination tracking. In collaboration with academic colleagues, DaSH have begun a project on service delivery enhancements within the NHS clinical setting to improve children and young people's mental health in the clinical setting. With DaSH staff embedded in the research team, their analysis using pseudonymised prescribing and hospital data showed significant potential to assist clinical teams in monitoring and improving children and young people's mental health [26]. DaSH is facilitating the implementation of the code (developed to analyse the static, pseudonymised data) into the live NHS environment, which

[3] The University of Aberdeen's datasets offer a wealth of longitudinal studies; for example: University of Aberdeen Children of the 1950s; Aberdeen Maternity and Neonatal Databank (AMND), covering maternity, pregnancy and birth records in Aberdeen from 1949-2018; Parkinsonism Incidence in North-East (PINE) study: and Parkinson's Incidence Cohorts Collaboration (PICC), amongst others [15–18].

will be used to build real-time dashboards for clinical teams to enhance monitoring and treatment of patients. Similarly, DaSH analysts are increasingly working with NHS clinicians to produce reporting data within the NHS environment, particularly around non-hospital prescribing data and linking to live hospital data, using methodologies grounded in academic research in order to improve clinical operations and reporting.

Finally, the partnership model enables collaboration between NHS Grampian and University of Aberdeen information governance and research ethics teams. All projects, staff employed by and researchers using the DaSH facility comply with applicable information governance and ethical research requirements, safeguarding and harmonising appropriate levels of data sharing. Most importantly, patient confidentiality and public benefit are maintained across both NHS and University environments; this will be discussed further below in the section 'Governance, Legislation and Management'.

There are variations in operating models taken by the other regional safe havens in Scotland. An approach taken by one regional safe haven was to embed their staff within the NHS; NHS staff perform all data extraction and linkage, but with the safe haven's outputs tending towards clinical support. The second approach taken was to embed the safe haven almost exclusively within their university environment with only limited access to non-pseudonymised patient-level data; the major focus is academic research with limited clinical involvement. In spite of these different operating models, each of the four regional safe havens provide important analytic platforms for the safe access of healthcare data. We believe, however, that the Grampian model is particularly successful because of the joint partnership between NHS Grampian and University of Aberdeen, where there is equal emphasis on both clinical and academic research and innovation.

Governance, legislation and management

As a partnership between NHS Grampian and University of Aberdeen and a nationally accredited Scottish safe haven, DaSH must operate within the principles and technical standards for operation as set out in the Charter for Safe Havens [27]. Individual researchers need to obtain sponsorship and ethics approvals for their particular projects, but any review will be proportionate and consider the governance arrangements for DaSH. Approvals are required from the data custodian, academic institution and NHS R&D, as well as NHS Grampian's Caldicott Guardian, an individual that provides oversight of the arrangements for the use and sharing of clinical information [28]. For non-Grampian data hosted within DaSH, additional approvals may be required from the Health and Social Care PBPP, Local Authority, Data Controller, or other Scottish health boards. In some instances, a Data Sharing Agreement may also be required depending on the nature of the data. DaSH use an NHS Grampian Data Protection Impact Assessment (DPIA) to determine whether a project falls under the generic DPIA agreed by NHS Grampian and University of Aberdeen. When a project involves a commercial partner or falls outside of the scope of the generic DPIA, an individual DPIA or Data Sharing Agreement is undertaken between the two data controllers. Throughout this process, DaSH Research Coordinators provide advice and support to researchers with obtaining the required approvals and governance agreements for DaSH projects.

One of the most significant benefits provided to researchers and clinicians offered by DaSH as a local TRE are the internal improvements related to information governance. Many clinicians and researchers have vocalised frustrations about challenges in applying for permissions, and particularly the length of time it can take for national permissions panels to review projects [29]. As a way of improving the ease of access to local data, DaSH Research Coordinators worked together with NHS Grampian Information Governance staff to provide a streamlined permissions pathway, the North Node Privacy Advisory Committee [30], which aims to approve projects that meet specific criteria in approximately three weeks. Applicable projects require a local clinician or researcher accessing local datasets approved by the NHS Grampian Caldicott Guardian. The single application form enables researchers that is shared with sponsor, NHSG Caldicott Guardian, Ethics and NHS Research & Development teams, as appropriate.

The Scottish safe havens were established to ensure that unconsented patient healthcare data could be processed and linked, and used to support research where it is not practicable to obtain individual patient consent. DaSH is bound by data protection legislation including the UK Data Protection Act [31], the EU General Data Protection Regulation (GDPR) [32] and the Common Law Duty of Confidentiality [33]. These set out the lawful processing of personal identifiable data, including special categories of data for which there are enhanced protections. The work that DaSH performs falls under the legal condition of performance of a public task under UK DPA and GDPR ('processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller' [34]. DaSH undertakes ISO 27001 certification to ensure information security and information security management systems (ISMS) meet international standards demonstrating adherence to the legislation and trustworthiness to the public. In instances where DaSH links research data collected with consent to data collected without consent, all of the data is pseudonymised to ensure that linkage between meets the legal and ethical requirements for the use of data without consent.

The DaSH Steering Committee oversees the governance and management of the facility and consists of representatives from NHS Grampian and University of Aberdeen Directorate of Digital and Information Services, information security, governance, research ethics, academics, clinicians, and a member of the public. If there are major changes to the structure, infrastructure or use of DaSH for research purposes, such as collaboration with industry or commercial partners, the DaSH Steering Committee is consulted. The Steering Committee Chair is responsible for ensuring oversight, continual improvement and development of the ISMS and for ongoing compliance within DaSH. The DaSH Steering Committee reports to both the University of Aberdeen's Clinical Research Steering Group and Information Governance Group. The benefits of the local trusted research environment mean that governance and management of DaSH are driven

from both within the two organisations as well as the local needs of the community.

All DaSH staff (Research Coordinators and Analysts) are required to undertake Information Governance training accepted by the national safe haven as well as NHS Grampian and University of Aberdeen Information Governance training. Additionally, DaSH staff must undertake Good Clinical Research Practice training and Research Ethics training. This enhanced governance requirement for DaSH staff ensures that they fully understand and work to protect individuals' personal data throughout their work. Standard operating procedures are maintained in the Information Security Management System and are routinely updated to ensure working practices reflect all governance, ethical and legal requirements.

Researchers accessing DaSH also have to fulfil certain requirements in order to be able to access data: they must be an approved researcher with appropriate research experience and also must hold a valid Information Governance Training Certificate approved by the national Scottish safe haven (eDRIS) which has a fixed validity period. Researchers must renew their training prior to their certificate's expiry to continue access within the project's permissions period. Commercial organisations would need a researcher on their project team. DaSH staff monitor who has accessed the facility and keep audit logs.

Architecture and information technology

DaSH is hosted, managed and provisioned across three 'environments', one within the NHS Grampian network and two within the University of Aberdeen network. The two environments inside the University environment are segmented into one platform for DaSH analyst staff for data preparation, and one for approved, authorised researchers to conduct their analysis on pseudonymised data, as per the infrastructure below:

The NHS Grampian DaSH platform exists within the NHS Grampian eHealth platform on a virtual SQL server. All patient identifiable data—whether internally extracted from NHS Grampian databases or externally provided by researchers, other Scottish health boards, the national Safe Haven, or external collaborators—is stored within the NHS Grampian platform and pseudonymised within the NHS Grampian network. The data extraction and linkage performed by the trained DaSH analysts is checked twice to ensure correct extraction, linkage and pseudonymisation before signoff for transfer to the DaSH analyst platform in the University Safe Haven.

The University Safe Haven platform exists across two virtual servers, one for DaSH staff and one for approved researchers. The DaSH analyst virtual SQL server receives the securely transferred, pseudonymised data from the NHS network, and data linkage and an additional pseudonymisation is undertaken by trained DaSH analyst staff to ensure patient confidentiality and for any additional data processing required (e.g. sensitive data fields) or requested (by researchers). This second pseudonymisation is carried out using different coding so that individual patient's data cannot be linked between different projects in which they could appear in the cohort. Once the data has been checked and signed off for transfer within the DaSH analyst platform, the data is released to the DaSH researcher server.

Once logged in to either the DaSH analyst environment or DaSH researcher environment, neither DaSH analysts nor approved researchers have a connection to the internet to ensure the data cannot be passively released. Within the DaSH analyst platform, analysts only have access to vetted, approved coding languages and statistical packages. Within the DaSH researcher platform, researchers cannot download or print the data or results of any data analyses; likewise they only have access to pre-agreed datasets and statistical packages that they use to run the analysis. Any bespoke code that researchers have written themselves that can be run within a statistical package is reviewed prior to upload. Researchers are provided access to the safe haven via VPN to the virtual environment; DaSH also offer a physical 'safe room' monitored by DaSH staff if they prefer to access the data on campus. Once the researcher is ready to publish their findings, their outputs are disclosure checked before any outputs are released. Only aggregated results are allowed and are disclosure checked for a minimum of five records [25]. If researchers request the release of in their analysis with the values of N/A or a null result, they need permission from the DaSH Director and Clinical Lead to ensure these values would not disclose a patient's identity if released.[4]

Privacy and data access

DaSH's infrastructure was designed to meet and exceed all required technical conditions as set out within the Scottish Charter for Safe Havens based on the principle of privacy by design. Whilst maintaining patient privacy and confidentiality as its core principle, DaSH also recognising the importance of ongoing advancements in information systems, technology, software and analytical methodologies. The DaSH privacy by design model as set out in the Scottish Charter for Safe Havens ensures the following privacy and security measures are in place and ensures restricted data access.

Physical controls

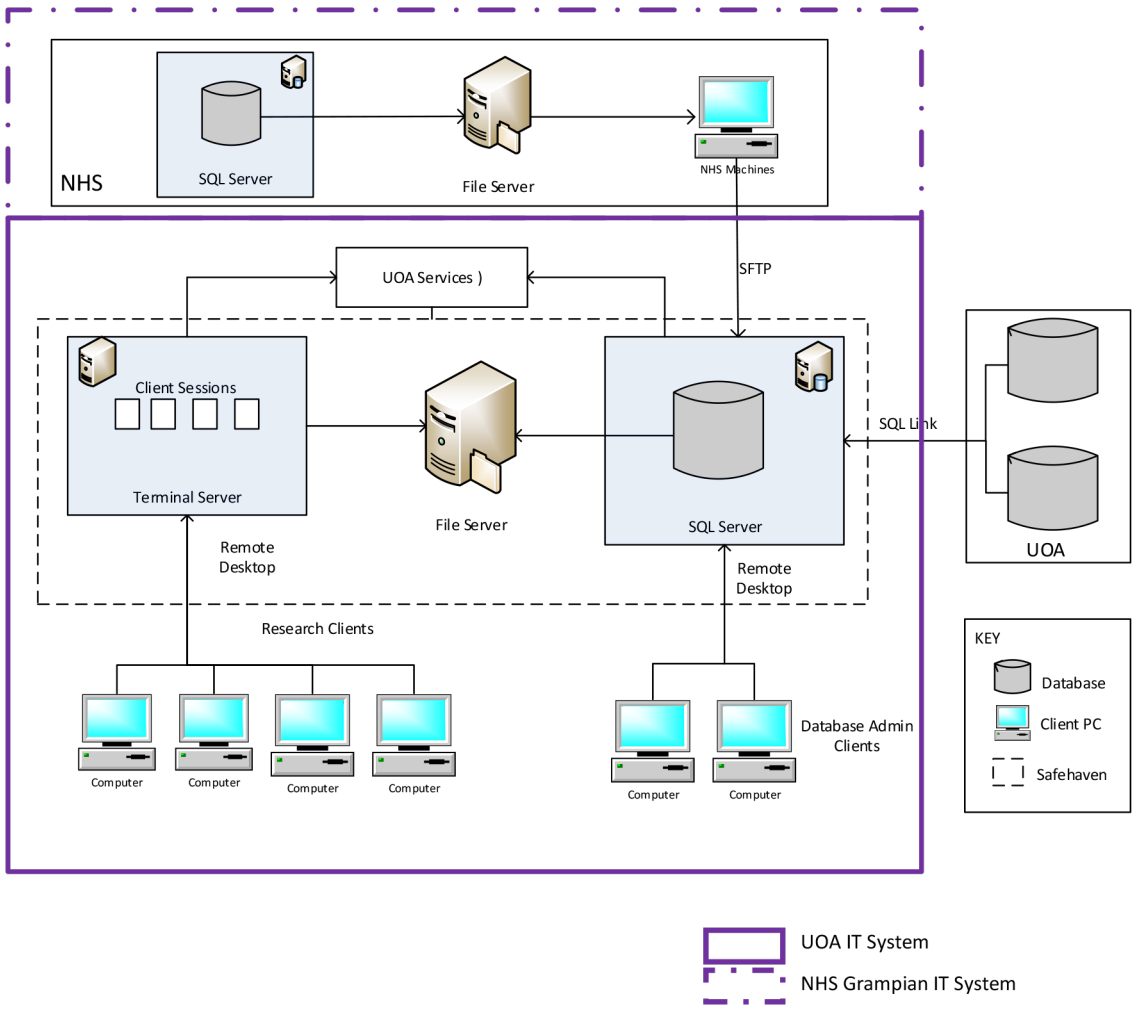
DaSH staff machines with access to the virtual environment are located within an access-controlled office within an access-controlled research centre on the University of Aberdeen Campus. The University of Aberdeen data centre that houses the DaSH infrastructure is within an access-controlled building. NHS Grampian infrastructure is similarly arranged with physical and environmental controls.

Infrastructure and operational controls

The DaSH infrastructure (Figure 1) on both the NHS Grampian and University networks are tightly controlled, with only patient-identifiable data stored on NHS Grampian and pseudonymised data stored within the University of Aberdeen environment. Transfer of personally identifiable data must use the NHS's secure file transfer software using NHS staff

[4]Occasionally, researchers request to release values of not applicable or null values. In these instances, DaSH Director and/or Clinical Lead will assess any risk of patient identifiability and whether these values can be retained in outputs.

Figure 1: The Grampian DaSH Infrastructure



accounts. Transfer of pseudonymised data from the NHS network must use the same secure file transfer software to transfer data to the DaSH analyst platform in the University of Aberdeen network.

Data transfers from the NHS Grampian platform to the DaSH analyst platform undergo two data checks by different staff members as well as a signoff procedure to ensure only approved variables are released and that appropriate pseudonymisation has been undertaken correctly. Duplicate patients are also screened. For the payload data (e.g. all data related to the cohort) released to researchers, a check and signoff protocol is in place to ensure that all values match against the transferred values, the datasets contain the approved variables and a further pseudonymisation has been conducted.

Data is released to researchers only after a verification process: all permissions must be in place, each member of the research team must sign an Investigator Declaration agreeing to the operating procedures of the safe haven and all researchers have attend a pre-access call with DaSH Research Coordinators, which covers their legal and ethical responsibilities when using the safe haven. Release of researcher analysis outputs from the DaSH facility is only done after disclosure checks to ensure small numbers (<5) are removed. Requested outputs are transferred via end-to-end encrypted software.

There is no internet access within the safe haven environment and only approved, vetted statistical software and packages are made available to DaSH analysts and researchers. If a non-approved package is required, staff and researchers must submit a request, which is reviewed by the DaSH Director and Information Security Manager in the first instance. There are a number of system access levels and user authentication controls including: multi-factor authentication, password policy controls, banned password list, controlled group access. The University's VPN and remote desktop also require a username and password. The safe haven undergoes monthly vulnerability scans and regular penetration testing and security assessments.

Staff access controls

DaSH staff are only provided access to the Safe Haven environments after all information governance, research ethics and clinical research training has been undertaken. Training certificates are stored and monitored, and staff are required to renew their training within a set period of time to ensure ongoing compliance with information governance policies and updated legislation and requirements. Staff can only access NHS Grampian or University of Aberdeen networks via specific machines and accounts. University of Aberdeen IT staff supporting the infrastructure and databases within

the university safe haven environments must also undergo enhanced information governance training and must maintain ongoing training at set intervals.

Researcher access controls

Access controls for researchers are at multiple levels. Researchers must be named on individual project permissions, and they must undertake enhanced information governance training and must maintain training at set intervals during their project's duration. Their access to the safe haven is only provided after all documentation is completed and a pre-access call is undertaken. They are only provided access at a system level to their specific project folder(s).

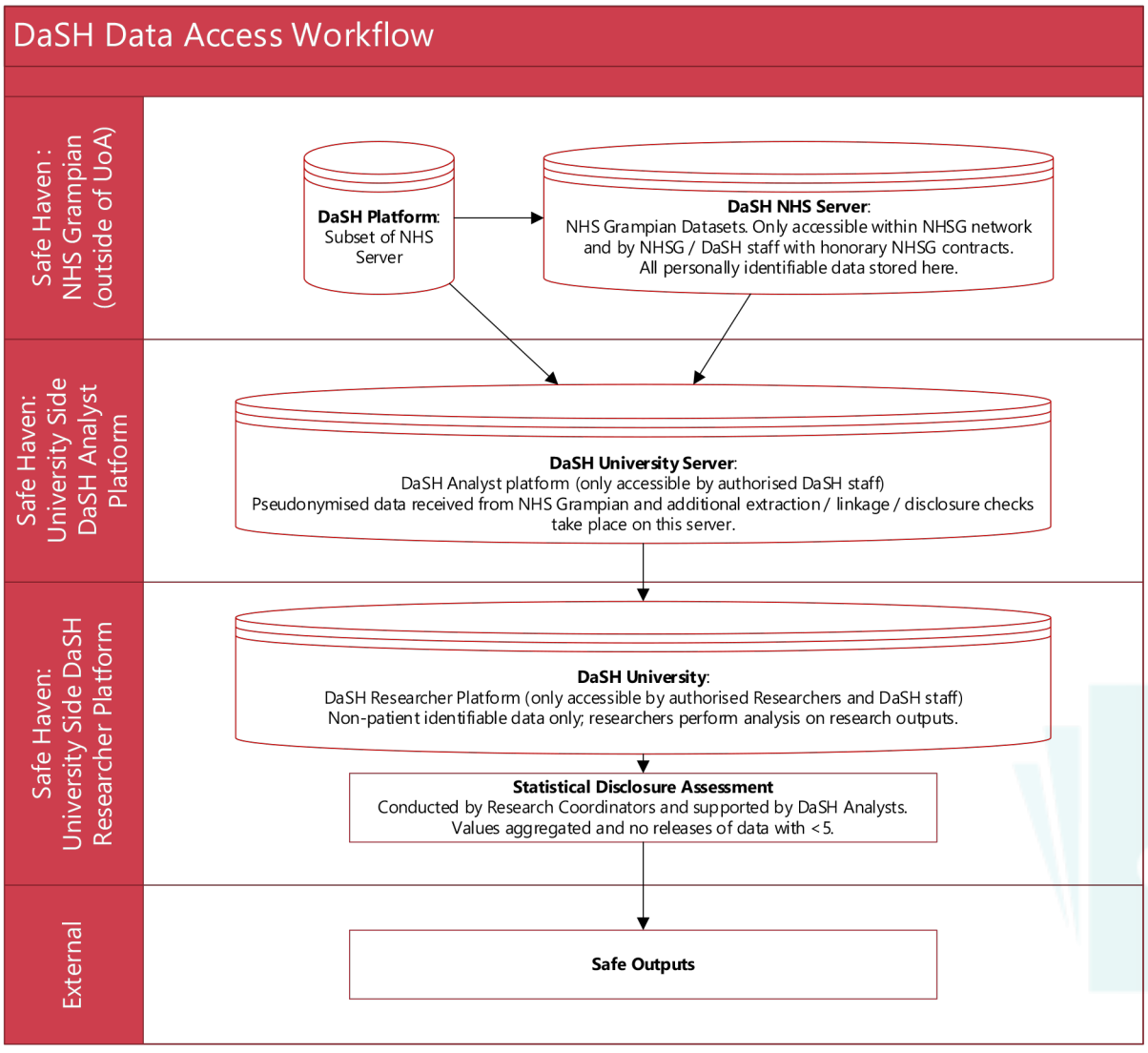
Privacy and data access is visualised as shown in Figure 2.

Data linkage

Across most DaSH projects, the linkage of datasets is by means of the Community Health Index (CHI) number issued to an individual in Scotland on first registration within the NHS Scotland system. The CHI acts as a unique identifier and is the most accurate way of creating linkages between different health datasets, since the numbers in different datasets can be matched. There are inevitably data inconsistencies where temporary CHI numbers are recorded (for example, where a patient has not been issued a CHI number through the local GP or it cannot be found at the time of the event recording); however regular updates to the CHI register are performed to capture the permanent CHI where possible. Link tables to CHI and UID are stored only within the NHS Grampian environment.

Within DaSH, there are projects which require linkage without the ability to match on CHI. These projects often include social services data or non-Scottish data where CHI may not be captured. For these projects, probabilistic methodologies are used, attempting to match data based on variables set out within project permissions (first name, surname, postcode, date of birth, etc.). A probability score is provided based against the matched variables to researchers.

Figure 2: DaSH data access workflow



In complex projects, basic cohort statistics are provided to clinician-researchers in line with project permissions and output release processes to ensure that the cohort numbers reflect the clinical presentation. The benefit of this additional output is that it acts as a validation mechanism between healthcare professionals' clinical experience and the data analysts' cohort coding, preventing patients from being included in a cohort erroneously (or conversely, excluding patients that should be in the cohort) prior to full dataset extraction, linkage and data release.

Data sources

DaSH has NHS Grampian Caldicott approval in place for access to over 30 NHS Grampian datasets. Additionally, DaSH holds generic ethics approval to seven NHS Grampian and University of Aberdeen datasets. The importance of these local datasets includes their extensive longitudinal population data: two of these datasets have over 70 years of longitudinal health and social data, extensively curated by NHS clinician and University of Aberdeen researchers.

DaSH regularly expands its generic dataset access approvals as further datasets are requested by researchers or there is a likely research interest in datasets if made available to researchers. Additionally, datasets that DaSH does not have Caldicott/generic approvals for can be extracted and linked within NHS Grampian on a project-by-project basis with appropriate approvals.

The following is a summary of 'core' datasets most frequently requested by researchers; information on all datasets held by DaSH can be found at: https://www.abdn.ac.uk/iahs/facilities/dash-datasets.php

Core datasets	Features (with approximate volumes)
Grampian SMR00 (Outpatient attendance)	The Grampian segment of the national SMR00 dataset. Outpatient attendance records. 10.4 million records from 1997 to present; updated monthly.
Grampian SMR01 (General/Acute Inpatient and Day Case)	The Grampian segment of the national SMR01 dataset. General / acute Inpatient and Day Case records. 3.2 million records from 1981 to present; updated monthly.
Grampian SMR02 (Maternity Inpatient and Day Case)	The Grampian segment of the national SMR02 dataset. Maternity Inpatient and Day Case records. 300,000 patient records from 1975 to present; updated monthly.
Grampian SMR04 (Mental Health Inpatient and Day Case)	The Grampian segment of the national SMR04 dataset. Mental Health Inpatient and Day Case. 59,000 patient records from 1981 to present; updated monthly.
Grampian SMR06 (Cancer Registry)	The Grampian segment of the national SMR06 dataset. Cancer Registry records. 175,000 patient records from 1981 to present; updated monthly.
Grampian NRS Deaths	The Grampian segment of the NRS statistics of deaths. 234,000 patient records from 1981 to present; updated monthly.
Grampian Prescribing Information System (PIS)	The Grampian segment of the national PIS dataset. Monthly prescribing data for prescriptions prescribed and dispensed in the community. Patient and drug level details for prescribed, dispensed and paid prescriptions.111 million records from January 2009 to present; updated monthly.
Grampian TrakCare	TrakCare is NHS Grampian's live hospital system data. Covers from the period of 2016 to present and updated every two months, with counts against the following: Inpatient Admissions:1.47 million records Inpatient Ward Changes:3 million records Inpatient Specialty Changes: 2.8 million records Inpatient Care Provider Changes: 2.67 million records Emergency Department Attendances: 957,000 records Outpatient Activity: 1.35 million records Waiting List: 6.37 million records
Grampian APEX Laboratory – Biochemistry	Grampian Apex Laboratory data for Biochemistry samples. 313 million records from 1996 to present; updated monthly.
Grampian Apex Laboratory – Haematology (including Immunology)	Grampian Apex Laboratory data for Haematology and Immunology samples. 262 million records from 1996 to present; updated monthly.
Aberdeen Maternity and Neonatal Databank (AMND)	Information on approximately 200,000 pregnancies/deliveries from 1949-2018and links all the obstetric and fertility-related events occurring to women from the Grampian region.
Aberdeen Children of the 1950s (ACONF)	A population-based resource for the study of biological and social influences on health across the life-course and between generations. Comprising 12,150 individuals born in Aberdeenbetween 1950 and 1956, comprised of participants of the Aberdeen Child Development Survey in December 1962. ACONF have been successful in ascertaining the current vital status and whereabouts of 98.5% of the 12,150 subjects (6276 males, 5874 females) with full baseline data.

Other datasets that are gaining increased interest are Grampian Picture Archiving and Communication System (PACS) Imaging and Report data; Grampian Covid-19 testing, vaccination and shielding data, including care home addresses; Grampian Badgernet Maternity database, and Grampian Child and Adolescent Mental Health Services (CAMHS).

Noteworthy outputs and service delivery improvements

DaSH has registered over 600 projects, has helped support over 200 distinct projects with hosting, data extraction and/or data linkage to completion [35], and has over 150 researchers accessing the safe haven per year. Three examples provided shows the breadth and depth of the research supported.

Example 1: Large-scale multinational cohort analysis to advance healthcare research

University of Aberdeen, NHS Grampian, University of Dundee, Aarhus University, Aarhus University Hospital, and University of Calgary examined incidence of advanced kidney injury and advanced kidney disease across seven million residents in Canada, Denmark, Tayside (UK) and Grampian to measure the incidence and prognosis outcomes [36]. Their study confirmed the use of detection algorithms to distinguish patients based on laboratory testing as opposed to earlier studies which showed variability in incidence and prognosis. Researchers also noted the importance of applying consistent analytical methods across cohort datasets to improve the comparability of kidney research in different settings. The latter finding further substantiates the importance of trusted research environments such as DaSH in providing and hosting data supports the analysis of large-scale, multinational research.

Example 2: Artificial Intelligence algorithm training for use in clinical settings

Researchers from the University of Aberdeen and NHS Grampian have identified that active clinical referral triage in hip and knee osteoarthritis has significant challenges, including waiting times for surgeries and difficulties of non-specialists in determining patients that might benefit from surgical interventions [37]. The aim of this project is to develop AI machine learning to select suitable patients correctly and rapidly for surgery. The aim of the research is to define the criteria for the machine learning using extracted and linked routine medical records and images to determine the demographic, clinical and imaging characteristics of selected patients who would benefit hip and knee arthroplasty, and to develop a validated predictive model using AI in a clinical setting. This project is now underway and builds on co-current iCAIRD projects across a number of specialisations. The suite of work produced within iCAIRD research has won numerous awards its programme of digital research in the design and training of AI for clinical applications [38]. NHS Grampian and the University of Aberdeen, as partners with iCAIRD, have helped facilitate the extraction and linkage of complex datasets, ensuring patient confidentiality and privacy is critical in developing machine learning diagnostics.

Example 3: Understanding trends in child mental health support and treatment

University of Aberdeen and NHS Grampian researchers (in partnership with The Health Foundation) have recently investigated the incidence of and inequalities between child mental illness treatment and support within the Grampian child and adolescent population by examining differences in treatment and service by age, sex, and socioeconomic factors [39]. DaSH provided support in novel linkages between prescribing, hospital attendance, referrals, and acceptance at outpatient treatment to allow researchers to understand trends in recent treatment and support and suggests that there may be an increasing prevalence of poor mental health treatment and specialised support for particular groups of young people. This emerging research shows the importance of novel linkages supported by DaSH in identifying health and social care challenges, and could be of use to clinicians, local authorities, and other educational and social care institutions in growing services through data-driven research. In phase 2 of this research, this novel linkage will be linked to Aberdeen City Council Local Authority data to improve the identification and triaging of at-risk children, aiming to improve early intervention and long-term outcomes in vulnerable populations.

Service delivery improvements

As project complexity grows within DaSH and with researchers wanting to access larger amounts of complex data, particularly imaging data, there is an increasing requirement for complex statistical software requirements, all within an internet-free environment. This not only makes installing, updating and supporting infrastructure and software incredibly challenging, but there are risks to using open-source programming languages related to security. Acknowledging this challenge, DaSH has worked to improve and innovate its existing infrastructure and software portfolio to ensure researchers are able to perform their analyses using their preferred tools whilst maintaining a maximum-security environment to protect patient data. The first improvement was the deployment of a CRAN-mirror for R software packages. The R-mirror allows researchers to download and install packages from an approved set of R packages to their individual project folders without the support of DaSH analysts. A researcher can submit a request to have new packages added to the R-mirror where a package has not yet been approved and DaSH staff assess the suitability of the package within the environment. This has improved researcher experience since they no longer need to wait for DaSH analyst staff to install packages and has eliminated the task from DaSH analysts' workload. The second improvement for particularly complex projects that requires the use of 'at risk software' (i.e. Python) has been to create a safe haven within the safe haven where specific parts of projects requiring researchers to access to Python are housed, thereby allowing them to conduct highly specialised, specific statistical analyses without exposing the standard safe haven to the risk that certain software poses. These projects are considered by

DaSH on a case-by-case basis, either based on the complexity of the project or where a project may be a multi-site project with researchers employing universal code not offered in the standard safe haven.

Discussion

In the 10 years since the Grampian Data Safe Haven was established, health and social care, and particularly the use of data-driven decision making within healthcare, has grown tremendously, as have the complexity of projects that DaSH supports. With increasing complexity of projects, we also acknowledge a number of lessons learned across governance, infrastructure, and operations.

One of the main challenges faced by researchers is the complicated project applications and approvals process, which can often have lengthy waiting periods whilst projects are being reviewed. DaSH has a team of Research Coordinators to assist researchers through the different permissions pathways depending on project requirements. However, a significant improvement to the challenge of governance that has been made for researchers wishing to access Grampian data only is the design and implementation of a streamlined permissions pathway. DaSH has worked together with NHS Grampian and the University of Aberdeen to establish the North Node Privacy Advisory Committee. The Committee provides project approvals to researchers with access to NHS Grampian patient (health) data for research purposes via a pathway that incorporates Sponsorship, Ethics, R&D and Caldicott for any data extracted, linked and hosted within DaSH. Project approvals take approximately one month for the NNPAC Committee to review, as compared to much longer waiting periods via PBPP. Finally, with the volume and complexity of projects growing with researchers and clinicians increasingly interested in using the DaSH facility, it was recognised by the DaSH Steering Committee that additional staff was required to help manage projects, workload and priorities across the DaSH team and to facilitate communication with researchers clearly to improve transparency, open research (insofar as possible with unconsented patient data) and provisioning of data [39]. The creation of a dedicated Operational Lead with project management qualifications within the DaSH team underscores the commitment to operational improvements within the team to improve efficiency and focus on agile, iterative and time-bound delivery of projects.

There are areas for further work that DaSH has begun or will begin to explore to continue to develop the range of research projects that can be supported within the safe haven.

First, DaSH is working to expand its provisioning of extracting and linking new and emerging data types. The first specification is the ingestion and expansion of anonymising modalities across imaging data to include x-ray, CT and MRI. The second provision is free-text data, where the substance of free text data is available to researchers after de-identification of any personal identifiable data in clinical free text. Another provision is to support projects linking genomic and tissue data to clinical datasets widening health determinant research.

Second, DaSH is expanding, improving and publishing metadata catalogues within the HDRUK Gateway and Research Data Scotland research portals to improve awareness of Grampian's datasets and a more comprehensive dictionary of variables and definitions to improve researcher experience when selecting extraction requirements. DaSH is committed to transparency and open research, and improvements to metadata will offer researchers confidence that their data extraction and linkage requests have been performed using the most appropriate relevant data fields. In parallel, DaSH is also exploring the creation of synthetic datasets to enhance transparency and open research whilst protecting patient confidentiality and privacy.

Finally, increasing interest in multidisciplinary research by population scientists, researchers and clinician collaborators has shown the need for an expansion of datasets covering additional population data to understand connections between economic, demographic and social care with healthcare outcomes. To that end, DaSH is working on partnerships with local authorities in Grampian to link social care data with health records of vulnerable populations to improve healthcare outcomes and embed data-driven research methodologies within local authority teams.

Conclusions

Although situated within a relatively small population setting, DaSH holds a wealth of health and population data for North East Scotland that promotes diverse research opportunities whilst preserving individual privacy and confidentiality. Ongoing improvements to DaSH's infrastructure and operational processes continue to refine extraction and linkage techniques, expand the availability of new data and consider new opportunities for novel data linkages and real-world applications, supporting researchers and clinicians to improve outcomes in health and social care.

Most importantly, however, as a local TRE with extensive experience and research outputs, the services DaSH provides, its focus on innovation, collaboration and commitment to open research whilst ensuring patient privacy shows that bigger is not always better. The starting point of an effective national UK data strategy could be to focus on the needs of local communities, bringing together and engaging with the well-established, local TREs within the UK to information-share existing infrastructure and architecture, information governance processes, collaborations and partnerships (including commercial) to highlight innovations and transformative approaches already being undertaken within these data centres. Sharing local solutions to local challenges – in addition to local approaches to data curation, platform design and openness and re-use of code – could achieve alignment, harmonisation and shared innovation across the UK by facilitating collaborative opportunities encouraging data-driven solutions to local population needs. Perhaps a solution to better, broader and safer national health data research nationally is to harness the 'power and potential' (p. 5) of the work already underway locally.

Acknowledgements

Funding for DaSH is provided by the University of Aberdeen and NHS Grampian. Past and current grants from Research

Data Scotland also contribute to DaSH's funding, along with funding provided by DaSH researchers for use of the facility. The authors also acknowledge the DaSH Team (DaSH Clinical Lead Dr Shantini Paranjothy; DaSH Research Coordinators Joanne Lumsden, Vicky Munro and Diane Brown; DaSH Analysts Helen Rowlands, Adrian Martin, Jaroslaw Dymiter, and Michael Lackenby; and DaSH Information Security Manager Gary Cooper), without whom the work would not be possible.

Conflicts of interest

The authors declare they have no conflicts of interest.

References

1. Goldacre B, Morley J. Better, broader, safer: using health data for research and analysis [Internet]. 2022 [cited 15 October 2022]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067053/goldacre-review-using-health-data-for-research-and-analysis.pdf.

2. HDRUK. Trusted Research Environments and data management – Past, Present and Future [Internet]. 26 April 2022 [cited 15 October 2022]. Available from https://www.hdruk.ac.uk/wp-content/uploads/2021/04/Goldacre-Review-TRE-Response.pdf.

3. DAREUK. DARE UK welcomes findings of ministerial review into use of health data for research and analysis [Internet]. 11 April 2022 [cited 15 October 2022]. Available from: https://dareuk.org.uk/dare-uk-welcomes-findings-of-ministerial-review-into-use-of-health-data-for-research-and-analysis/.

4. National Assembly for Wales. The organisation of the NHS in the UK: comparing structures in the four countries [Internet]. 2015 [cited 23 October 2022]. Available from: https://senedd.wales/research%20documents/15-020%20-%20the%20organisation%20of%20the%20nhs%20in%20the%20uk%20comparing%20structures%20in%20the%20four%20countries/15-020.pdf.

5. Department of Health. Information: To share or not to share? The Information Governance Review [Internet]. 2013 [cited 15 October 2022]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf.

6. Scottish Government. Charter for Safe Havens in Scotland: Handling Unconsented Data from National Health Service Patient Records to Support Research and Statistics [Internet]. 2015 [cited 27 June 2022]. Available from: https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/.

7. About – ScottiH Informatics Programme [Internet]. 2011 [cited 27 June 2022]. Available from: http://www.scot-ship.ac.uk/about.html.

8. C, McGilchrist M, Mumtaz S, Hall C, Anderson LA, Zurowski J, Gordon S, Lumsden J, Munro V, Wozniak A, Sibley M, Banks C, Duncan C, Linksted P, Hume A, Stables CL, Mayor C, Caldwell J, Wilde K, Cole C, Jefferson E. A National Network of Safe Havens: Scottish Perspective. Journal of Medical Internet Research 2022 24(3): e31684. https://doi.org/10.2196/31684

9. Waind E. Trust, security and public interest: Striking the balance: A review of previous literature on public attitudes towards the sharing, linking and use of administrative data for research. International Journal of Population Data Science 2020 5(3):1368. https://doi.org/10.23889/ijpds.v5i3.1368

10. Aitken M, McAteer G, Davidson S, Frostick C, Cunningham-Burley S. Public Preferences Regarding Data Linkage for Health Research: A Discrete Choice Experiment, International Journal of Population Data Science. 2018 3(1):429. https://doi.org/10.23889/ijpds.v3i1.429.22

11. Stockdale J, Cassell J, Ford E. 'Giving something back': A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland (Revised). Wellcome Open Research. 2019 3(6). https://doi.org/10.12688/wellcomeopenres.13531.2.23

12. Tully MP, Hassan T, Oswald M, Ainsworth J. Commercial use of health data – A public 'trial' by citizens' jury. Learning Health Systems. 2019 3(4). https://doi.org/10.1002/lrh2.10200

13. Wellcome Trust. Summary Report of Qualitative Research into Public Attitudes to Personal Data and Linking Personal Data [Internet]. 2013 [cited 27 June 2022]. Available from: https://wellcomecollection.org/works/am3zpgqv.

14. University of Aberdeen Children of the 1950s [Internet]. ND [cited 27 June 2022]. Available from: https://www.abdn.ac.uk/birth-cohorts/1950s/.

15. University of Aberdeen, AMND [Internet]. ND [cited 27 June 2022]. Available from: https://www.abdn.ac.uk/iahs/research/obsgynae/amnd/index.php.

16. University of Aberdeen. PINE [Internet]. ND [cited 27 June 2022]. Available from: https://www.abdn.ac.uk/iahs/research/chronic-disease/pine-study-2066.php.

17. University of Aberdeen. PICC [Internet]. ND [cited 27 June 2022]. Available from: https://www.abdn.ac.uk/iahs/research/chronic-disease/picc-2071.php.

18. NHS Grampian. About NHS Grampian [Internet]. ND [cited 27 June 2022]. Available from: https://www.nhsgrampian.org/about-us/about-nhs-grampian/.

19. NRS Scotland. Aberdeen City Council Area Profile [Internet]. June 2021 [cited 27 June 2022]. Available from: https://www.nrscotland.gov.uk/files//statistics/council-area-data-sheets/aberdeen-city-council-profile.html.

20. Aberdeenshire Council. Aberdeenshire Profile 2022 [Internet]. April 2022 [cited 27 June 2022]. Available from: https://www.aberdeenshire.gov.uk/media/26182/aberdeenshireprofile.pdf.

21. Scottish Government. Scottish Index of Multiple Deprivation 2020 [Internet]. 2020 [cited 27 June 2022]. Available from: https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/.

22. Butler J., Deprivation in Grampian 2020 [Internet]. 2020 [cited 27 June 2022]. Available from: https://jessbutler.github.io/simd/.

23. Grampian Regional Equality Council. How Fair is North East Scotland? Integration & Community Cohesion in Aberdeen City, Aberdeenshire and Moray [Internet]. 2020 [cited 27 June 2022]. Available from: https://grec.co.uk/wp-content/uploads/HFINES_dec2021.pdf.

24. Pioneer: The European Network of Excellence for Big Data in Prostate Cancer [Internet]. ND [cited 15 October 2022]. Available from: https://prostate-pioneer.eu/.

25. Ball, W, Black, C, Gordon, S, Ostrovska, B, Paranjothy, S, Rasalam, A, Ritchie, D, Rowlands, H, Rzewuska, M, Thompson, E, Wilde, K, Butler, J. Inequalities in children's mental health care: analysis of routinely collected data on prescribing and referrals to secondary care. BMC Psychiatry 23, 22 (2023). https://doi.org/10.1186/s12888-022-04438-5

26. Scottish Government. Technical Annex: Charter for Safe Havens in Scotland: Handling Unconsented Data from National Health Service Patient Records to Support Research and Statistics [Internet]. 2015 [cited 27 June 2022]. Available from: https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/pages/6/.

27. UK Caldicott Guardian Council. A Manual for Caldicott Guardians [Internet]. 2017 [cited 27 June 2022]. Available from: https://www.ukcgc.uk/manual/contents.

28. Taylor JA, Crowe S, Espuny Pujol F, et al. The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement. British Medical Journal Open 2021 (11): e047575. https://doi.org/10.1136/bmjopen-2020-047575.

29. Grampian Research Office (NHS Grampian and University of Aberdeen). NNPAC Data Studies [Internet]. ND [cited 27 June 2022]. Available from: https://www.abdn.ac.uk/clinicalresearchgovernance/nnpac-data-studies-134.php.

30. HM Government. UK Data Protection Act [Internet]. 2018 [cited 27 June 2022]. Available from: http://www.legislation.gov.uk/ukpga/2018/12/notes/division/6/index.htm.

31. European Union. General Data Protection Regulation [Internet]. 2016 [cited 27 June 2022]. Available from: https://gdpr-info.eu/.

32. UK Caldicott Guardian Council. A Manual for Caldicott Guardians: Common Law Duty of Confidentiality [Internet]. 2017 [cited 27 June 2022]. Available from: https://www.ukcgc.uk/duty-of-confidentiality.

33. Article 6(1) (e)of GDPR; Chapter 2(8), UK Data Protection Act 2018.

34. University of Aberdeen. DaSH-Projects [Internet]. ND [cited 27 June 2022]. Available from: https://www.abdn.ac.uk/research/digital-research/dash.

35. Sawhney S, Bell S, Black C, Fynbo Christiansen C, Heide-Jørgensen U, Kok Jensen S, Ronksley PE, Zhi Tan Z, Tonelli M, Walker H, James MT. Harmonization of epidemiology of acute kidney injury and acute kidney disease produces comparable findings across four geographic populations. Kidney International. 2022 101(6):1271–1281. https://doi.org/10.1016/j.kint.2022.02.033

36. Farrow L, Ashcroft GP, Zhong M, Anderson L. Using Artificial Intelligence to Revolutionise the Patient Care Pathway in Hip and Knee Arthroplasty (ARCHERY): Protocol for the Development of a Clinical Prediction Model. *JMIR Research Protocols*. 2022 11(5):e37092. https://doi.org/10.2196/37092

37. The Industrial Centre for Artificial Intelligence Research in Digital Diagnostics (iCAIRD). News [Internet]. ND [cited 27 June 2022]. Available from: https://icaird.com/news/.

38. Ball, W, Black, C, Gordon, S, Ostrovska, B, Paranjothy, S, Rasalam, A, Ritchie, D, Rowlands, H, Rzewuska, M, Thompson, E, Wilde, K, Butler, J. Inequalities in children's mental health care: analysis of routinely collected data on prescribing and referrals to secondary care. BMC Psychiatry 23, 22 (2023). https://doi.org/10.1186/s12888-022-04438-5

39. Chiocchio F, Rabbat F, Lebel P. Multi-Level Efficacy Evidence of a Combined Interprofessional Collaboration and Project Management Training Program for Healthcare Project Teams. Project Management Journal 2015 46(4):20–34. https://doi.org/10.1002/pmj.21507