Improving transparency in the use of health data for research: Draft recommendations for a data use registers standard

Green Paper for Consultation July 2021

UK Health Data Research Alliance

Green Paper v1.0 dated 09 July 2021

Table of Contents

UK Health Data Research Alliance

Executive Summary

A data use register (also known as a data release register or list of approved projects) offers the public a clear record of how their data is being used, by who and most importantly for what purpose. The April 2021 report 'Putting Good into Practice' by the National Data Guardian (NDG)[1] and Understanding Patient Data highlighted, *'transparency is required throughout the whole data life cycle, not just at the point of application'*.

There is also an expectation for the publication of *'clear statements of data users' credentials and sources of funding to protect against data manipulation, potentially non-altruistic motivations and hidden agendas.*' An appropriately designed and populated data user register that is openly available is therefore a necessary step towards meeting these expectations.

The UK Health Data Research Alliance brings together data custodians to establish best practice to enable the ethical use of UK health data for research and innovation at scale. It is therefore ideally placed to support the development and adoption of a data use register standard.
This Green Paper has been drafted following consultations and workshops with numerous organisations including health data research organisations and universities, data custodians and patient and public groups run between 25 March and 13 May 2021. We are very grateful for the high interest and participation to date (see Annex A for list of organisations). We are now seeking comments and contributions from all stakeholders on the recommendations set out below to help finalise the standards and support adoption. Our research on the current state of data use registers has provided valuable insight into the current practices of data controllers and custodians. The most significant finding was that only half of the data custodians analysed were found to publish registers. This was of biggest concern to the public contributors who highlighted that the priority must be to support organisations without a register to make this information public.

The feedback received to date by researchers, data custodians and the public has confirmed the need and value of data use registers in improving transparency in the use of health data for research, building public trust, as well as highlighting priorities and challenges likely to be faced in implementation.

- Recommendation 1: All data custodians and controllers responsible for the collection, storage and sharing of data for the purpose of research, innovation and service evaluation should publish a public record (data use register) of approved research studies, projects and other data uses.
- Recommendation 2: Data use registers should, as far as possible, be populated in near real time directly from information provided through the Data Access Request process to improve timeliness and accuracy of entries.
- Recommendation 3: Data use registers should be made available in both human readable and machine-readable formats.

- Recommendation 4: Data use registers should have a consistency of format and content based on the Five Safes framework to enable ease of understanding and aggregation of registers. A draft specification of recommended data fields to be included in data use registers is included in Table 1.
- Recommendation 5: Researchers, data custodians and funders should use data use registers to close the loop on the impact of data use by including links to research findings and other outputs as these become available – aligned with the #MakeItPublic strategy set out by Health Research Authority.

We welcome and encourage comments and contributions to help shape this standard and align it with other related work. We are also interested to hear suggestions as to how the adoption of standards for data use registers can be influenced and what barriers need to be addressed.

A **survey** has been created alongside this paper to establish level of agreement with the recommendations and capture feedback on the draft specification and implementation challenges.

This survey will remain open for 6 weeks over the summer.

Overview

Purpose

Transparency in the use of health data for research and innovation is the principal driver for this Green Paper. By establishing a core set of standards on who, how and why data is accessed, we hope to demonstrate the value and benefit of using health data, develop a culture of openness amongst data custodians, whilst also generating better insight into health data usage. Furthermore, publishing this information in an accessible and understandable format will help build public trust and, we hope, advocacy for data use. In addition to meeting transparency expectations, national bodies have a legal obligation to make this information publicly available. Research databases are also required to share a minimum dataset, as outlined by the ethics committee in their conditions for ethical approval[2].

An independent report published by the Centre for Data Ethics and Innovation[3] highlighted that the benefits of data access and use is not always felt equally among people. Data use registers can help public understanding while demonstrating the purpose of using confidential information for public benefit and the impact that might have been derived from it.

Data use registers also have the potential to improve the efficiency of research by highlighting past and present research projects and data uses to researchers and funders, prioritising funding for national data assets, and identifying under-served areas of data collections or research that could be prioritised for future

[2] QResearch. (2018, Dec 27). *Conditions of Ethical Approval*. https://www.qresearch.org/media/1155/257790-18em0400-qresearch-conditioins-of-ethical-approval-27122018.pdf

[3] Centre for Data Ethics and Innovation. Retrieved May 24, 2021, from: https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation

funding. These registers could also help close the loop on demonstrating public benefit by providing a much-needed link to the outputs and outcome of data use.

Data use registers are important for data custodians and the public, as well as to researchers and funders. Therefore, the question should not be whether this information is published, but rather to identify what the recommended standards should be and what represents best practice.

The standards will be useful for both data controllers and data custodians that operate access arrangements on behalf of data controllers. Data custodians that are in the process of developing a public register will benefit from a framework and guidance on the recommended approach, whilst those with established registers are recommended to strive for alignment to the agreed standards.

Background

The UK Health Data Research Alliance is an alliance of leading health, care and research organisations united to establish best practice to enable the ethical use of UK health data for research and innovation at scale.

With this work, the Alliance focuses on aligning approaches to improving and standardising the content, format, and frequency of publishing data use registers (also referred to as data release registers or list of approved studies). We define a data use register as a public record of data an organisation has provided approved individuals and organisations secure access to for the purpose of research, innovation, and service evaluation. It typically contains information about the type of data being accessed, the purpose, date of approval and name of organisation using (or receiving) the data.

The Alliance Letter of Intent[4] highlights transparency of governance and operations as one of the founding principles. As such, publishing a register of projects, studies and/or organisations that have used data under a member's custodianship is a vital demonstration of this.

Essential for success, is securing the widespread input of all key stakeholders. As the primary audience, the perspective of patients and the public is critical to identifying the minimum standard of a data use register. Nonetheless, the competitive nature of research is likely to influence the relationship between transparency and the link to research outputs. Therefore, researchers and innovators are also impacted by the publication of these registers. Finally, as the controllers and publishers of this information, data custodians will already have a good understanding of the need, value, and challenges to developing a standard.

The collective insight and experience of our stakeholders, combined with our understanding of the current state of data use registers and the shared learning on data transparency from other organisations has informed a preliminary set of recommendations. This initial phase of work has highlighted that variation in infrastructure and capability amongst data custodians may pose a risk to the adoption of 'high-level' or

[4] UK Health Data Research Alliance (2020, Mar). *Letter of Intent to Join the UK Health Data Research Alliance*. https://ukhealthdata.org/wp-content/uploads/2021/06/Alliance-Letter-of-Intent-Mar-20.pdf

resource dependent standards. In view of this, we suggest a minimum standard that can be adopted by most data custodians.

Insights from current state analysis

Our research on the current state of data use registers[5] has provided valuable insight into the practices of data controllers and custodians. Forty-six data custodians and controllers from the Alliance and seven non-Alliance members were included in this review. While half of the custodians analysed were found to publish registers, the others do not. As public representatives highlighted in discussions, the first important implementation step will be to support organisations without a register to make this information public.

A comparison of content highlighted greater prevalence of project titles, lay summaries, organisations, and applicant names, whilst information on the data sensitivity level, legal basis and location of data access was far less common.

When it came to format, most data use registers were published on a web page in either a list or table format; with the remainder published as either a link to a spreadsheet or a link to a pdf available for download. Information on publication frequency was not clear with most data custodians not specifying how often their registers were updated or when they were last updated.

This variability in the content, format, and frequency of publication amongst data custodians demonstrates the need for a standard and the great potential for improvement.

There is also an opportunity for behavioural changes amongst data custodians and researchers. Introducing the link to research outputs supported by assigning Digital Object Identifiers[6] (an identifier of an entity on digital networks, which provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks) to datasets has the potential to showcase the impact and benefit of data used for research, which should help build the confidence of patients and the public in sharing their data.

Learnings from the community

Workshops and interviews hosted on behalf of the Alliance stimulated discussion on the value of data use registers and provided better understanding of the needs and challenges to creating a standard. A mix of researchers, custodians, innovators, and members of the public contributed to the discussion. The following was highlighted for consideration:

- **Consistency in register names.** A standard approach to the naming of registers (or lists of approved projects) was highlighted as essential to public understanding and for demonstrating alignment in

[5] Nada Karrar, Kabir Khan, Sinduja Manohar, Paola Quattroni, David Seymour & Susheel Varma. (2021, May-26). *Analysis of Data Use Registers published by health data custodians in the UK*. https://www.medrxiv.org/content/10.1101/2021.05.25.21257785v1

[6] DataCite. *DOI Basics*. Retrieved June 08, 2021, from: https://support.datacite.org/docs/doi-basics

approach. 'Data use register' is considered preferable to a lay audience and more reflective of data use within a trusted research environment, rather than the traditional model of data 'release'.

- **Audience and impact on format.** Although, developing an easy-to-use platform could help meet the needs of the public, the information requirements of researchers could be more complex. In view of this, the group explored the possibility of offering a variation in the format and design of the data use register, supporting publication of data use registers on public web pages, whilst also ensuring that registers can be downloaded in a machine-readable format.
- **Accessibility through lay and standard terminology.** Clear lay summaries and consistency in the terminology used in data use registers was a top priority for all our stakeholders. Researchers may benefit from better signposting of guidance and standards already available on how to write a good lay summary. While offering recommendations on the content of a data use register, along with standard definitions of common data types would support an aligned approach to implementation, which will improve the public's understanding of how their data is being used.
- **Automation and impact on update frequency.** Members from our patient and public involvement panel, highlighted that data accuracy, authenticity and timeliness are essential to building their confidence and trust. It was suggested that automated processes could facilitate the transfer of information from data access management systems to data use registers in near real time, making it easier to keep data use registers up to date and ensure accurate information is publicly shared.
- **Improving the link to research outputs.** Demonstrating clear links between data use and research impact will help to create a culture of transparency and openness, as the Health Research Authority highlight in their 'make it public' strategy[7]; '*other than research for educational purposes and early phase trials, the findings, whether positive or negative, should be made accessible, with adequate consent and privacy safeguards, in a timely manner after they have finished.*' However, managing the relationship between approval and the resulting research outputs, is not always within the control of the data custodian, which means this information is not routinely captured and communicated. With this work we hope to inform best practice and highlight the relevance of data citation in data management practice, where alignment between researchers, data custodians and funders becomes necessary.

[7] Health Research Authority. (2020, Sep 03). *Make it Public: transparency and openness in health and social care research.* https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/research-transparency/make-it-public-transparency-and-openness-health-and-social-care-research/

UK Health Data Research Alliance

Emerging recommendations for data use registers

These recommendations have been informed by our research into the current state of data use registers, the shared learning from national bodies, as well as the knowledge, experience, and insight from key stakeholders.

Recommendation 1: All data custodians and controllers responsible for the collection, storage and sharing of data for the purpose of research, innovation and service evaluation should publish a public record (data use register) of approved research studies, projects and other data uses.

To support the accessibility and transparency of health care research, our recommendation is for data custodians to publish a register of any clinical or non-clinical research studies, projects and requests that have been approved for data use under their custodianship for research, innovation, and service evaluation. Data custodians who operate access arrangements on behalf of data controllers should also follow this recommendation.

As stated in the Alliance Letter of Intent, signed by new members, all Alliance member organisations are expected to *'publish a register of active projects accessing the data under their custodianship and new data access requests received. The register should be easy to find (as defined by a simple, quick and intuitive user journey) on a public facing website, and in an accessible format that includes at least the minimum dataset defined by the Alliance Board.'*

We would therefore expect Alliance members to lead by example for other UK organisations that provide access to health and social care data for research demonstrating the principle of transparency.

Recommendation 2: Data use registers should, as far as possible, be populated in near real time directly from information provided through the Data Access Request process to improve timeliness and accuracy of entries.

Regular updates and maintenance of data use registers are crucial to delivering public confidence and trust in why and by whom data is being used and demonstrating reliability. In the spirit of transparency, we recommend organisations update data use registers as soon as a data access request is approved. Automation of systems would enable near real time updating of data use registers while minimising manual

processes for data custodians, as some organisations already demonstrate (CPRD[8]). Whilst organisations are at different levels of readiness to adopt automation, the Health Data Research Innovation Gateway (*the Gateway*) does offer the opportunity to upgrade data access request processes without the need for each custodian to invest in new systems. We would therefore encourage organisations to explore that as a route to improve both data access processes and publication of data use registers.

During the transition to near real time updating, we recommend that data use registers are no more than three months behind approvals.

Recommendation 3: Data use registers should be made available in both human readable and machine-readable formats.

The format of the register must be accessible for public and patients as our primary audiences. The format must enable data controllers to meet legal obligations under principle 1 of the Data Protection Act[9] and provide insight for funders, researchers and other system stakeholders interested in the use of health data for research and innovation.

The public are more likely to find and understand information about data use that is embedded on an organisation's web page. Researchers, funders, and other system stakeholders may benefit from a machine-readable format that enables aggregation and analysis, may have more detailed information included, and which can be seen as the canonical record at a point in time.

With varying qualities and types of processes of data management, those two formats may be produced in different order by custodians. The intention is that the data use register developed for the Gateway will be available in both formats.

[8] CPRD. *Approved studies using CPRD Data*. Retrieved June 03, 2021, from: https://www.cprd.com/protocol-list

[9] UK Legislation. *Data Protection Act 2018*. Retrieved June 03, 2021, from: https://www.legislation.gov.uk/ukpga/2018/12/part/4/chapter/2/crossheading/the-data-protection-principles/enacted

UK Health Data Research Alliance

Recommendation 4: Data use registers should have a consistency of format and content based on the Five Safes framework and an agreed specification to enable ease of understanding and aggregation of registers.

Following discussions with key stakeholders and extensive research on the information currently shared through data use registers, we have developed the following list of recommended fields for a data use register, in accordance with the Five Safes Framework[10] on data access (Table 1).

A tiered approach to the implementation of standards was identified as the most practical and sustainable method of adoption. To support this, we have identified the fields to be considered for a minimum standard (reflected below through an asterisk) that are not only essential to public and patient understanding of data use but also already captured by data custodians through their data access approval process, which may reduce the effort required in adoption.

All Alliance member organisations are expected to include the suggested minimum standard dataset in their current data use registers. Datasets listed in the Gateway and accessed for use will be included in a Gateway data use register complying with these standards as it becomes available.

5 Safes	Field	Definition
Safe People	Organisation Name*	The name of the legal entity that signs the contract to access the data
	Funders/Sponsors/Collaborators*	The name of any funders, sponsors or collaborators involved in the project
	Organisation Type*	The type of organisation that has signed a contract to access the data
	Applicant Name*	The name of the principal applicant that has been authorised to use the data (in format First Name, Last Name)
	Accredited Researcher Status[11]	The accreditation status of a researcher, as defined by the ONS Research Code of Practice and Accreditation criteria
	Organisation ID	A unique identifier for an organisation that is preferably an industry used standard such as Grid.ac[12]
	Applicant ID	A unique identifier for the applicant that is preferably an industry used standard such as Grid.ac
	Sub-licence arrangements (if any) *	Identifies whether there are any permissions for the applicant to share the data beyond the named parties. e.g., NHS Digital may approve a data release to the ONS, who then makes decisions about access to accredited researchers undertaking approved projects in their own trusted research environment.

[10] https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf

[11] Office for National Statistics. Accessing secure research data as an accredited. Retrieved May 24, 2021, from: https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme

[12] Global Research Identifier Database. Retrieved May 24, 2021, from: https://www.grid.ac

Safe Project	Project Title*	The title of the project/research study/request that the applicant is investigating using health data
	Public Benefit Statement*	A description in plain English of the anticipated outcomes, or impact of project on the public
	Approval Date*	The date the application was approved in ISO 8601 format
	Lay Summary*	A concise and clear description of the project, (e.g., as required by UK Research and Innovation[13] in funding applications). It should outline the problem, objectives and expected outcomes in language that is understandable to the public
	Legal Basis for Provision of Data[14]*	The legal basis that allows the applicant to lawfully process personally identifiable data, as specified by NHS Digital
	Common Law Duty of Confidentiality[15]*	In the application of the Common Law Duty of Confidentiality there are 2 options that enable a release: Consent (Reasonable Expectation) or Section 251 NHS Act 2006
	Release/Access Date*	The date the data was released in ISO 8601 format
	Project ID	A unique identifier for the project that is preferably an industry used standard, such as IRAS ID[16]. However, for non-research projects, a unique reference number created by the data custodian on receipt of the application is sufficient
	Technical Summary	A summary of the proposed research, in a manner that is suitable for a specialist reader
	Other Approval Committees	Reference to other decision-making bodies that the project has already been authorised by
	Project Start Date	The date the project is scheduled to start or actual start date in ISO 8601 format
	Project End Date	The date the project is scheduled to finish or actual end date in ISO 8601 format
Safe Data	Publisher Name	The name of the organisation responsible for running or supporting the data access request process
	Dataset(s) ID	A unique identifier for the dataset(s) being shared or accessed
	Dataset(s) Name*	The name of the dataset(s) being accessed, as determined by the data controller
	Data Sensitivity Level*	The level of identifiability of the data being accessed
	Request Category Type*	This categorises the 'purpose of the share' (i.e., research, policy development, etc)
	National Data Opt-Out[17] Applied? *	Specifies whether the preference for people to opt-out of their confidential patient information being used for secondary use has been applied to the data prior to release
	Request Frequency*	Determines whether this a 'one-off' request or a recurring dataset to be provided over a specific time period

[13] UK Research and Innovation. How to apply for research and innovation funding. Retrieved May 24, 2021, from: https://www.ukri.org/apply-for-funding/before-you-apply/how-to-apply-for-research-and-innovation-funding/

[14] NHS Digital. NHS Digital Statutory Legal Duties under the Health and Social Care Act 2012. Retrieved May 24, 2021, from: https://digital.nhs.uk/services/data-access-request-service-dars/dars-guidance/advice-supporting-data-access-from-nhs-digital-in-respect-of-gdpr#nhs-digital-statutory-legal-duties-under-the-health-and-social-care-act-2012

[15] Department of Health – Northern Ireland. *The Common Law Duty of Confidentiality.* Retrieved May 24, 2021, from: https://www.health-ni.gov.uk/articles/common-law-duty-confidentiality

[16] Integrated Research Application System (IRAS). Retrieved May 24, 2021, from: https://www.myresearchproject.org.uk

[17] NHS Digital. *National Data Opt-Out*. Retrieved May 24, 2021, from: https://digital.nhs.uk/services/national-data-opt-out

	Description of how the data will be used*	Details of how the data requested will be used (including how they will analyse the data or link it to other datasets)
	Description of the confidential data being used	A description of the specific patient identifiable fields that have been included in the dataset(s) being accessed
Safe Setting	Trusted Research Environment (TRE) or other specified location (s) where data use will take place[18]*	These are highly secure spaces for researchers to access sensitive data (also known as Data Safe Havens). A list of available TREs will be provided (see glossary. If another location has been used, it will need to be specified
	How has data been processed to enhance privacy?	Description of the tools or software used to reduce level of identifiable data being shared
Safe Outputs	Link to Research Outputs*	A link to any academic or non-academic research outputs, as they become available, including code used

*Table 1: Summary of the recommended fields for a data use register, grouped according to the Five Safes Framework (also accessible in GitHub). *Specifies the fields that have been proposed as the minimum standard for a data use register.*

Recommendation 5: Researchers, data custodians and funders should use data use registers to close the loop on the impact of data use by including links to research findings and other outputs as these become available.

As outlined in the learnings from the community, demonstrating clear links between data use and impact (or lack of) will help to create a culture of transparency and openness. This is aligned with the #MakeItPublic strategy set out by Health Research Authority needs to be part of a system wide effort involving researchers, data custodians, funders, and regulators. It must also extend beyond traditional research outputs such as academic papers to include a wide range of potential impacts. We look forward to working with stakeholders to develop the levers to enable this to become routine practice.

Implementation - what would it take to change?

The readiness for change for each data custodian will vary and is likely to depend on the current depth and quality of their data use registers, as well as the infrastructure and resources available to support their transition to the recommended standard. This variability was evidenced through our research into the current state of data use registers, which highlighted a disparity in the content, functionality and timeliness of data currently published and more strikingly that almost half the organisations reviewed did not have a data use register publicly available.

The analysis of twenty-eight data custodians with known data use registers also revealed that there is a reasonable level of transparency on the people and projects that health data is being used by and for. However, there is very little information currently shared on the location (safe setting) for data usage and the precautions taken by data custodians to limit the risk of confidential data being accessed or disclosed. This demonstrates the need for greater transparency in this area and strengthens the case for the

[18] Tim Hubbard, Gerry Reilly, Susheel Varma, & David Seymour. (2020, July 21). *Trusted Research Environments (TRE) Green Paper (Version 2.0.0)*. Zenodo. http://doi.org/10.5281/zenodo.45947044

widespread use of Trusted Research Environments. The overall comparison of current state to the specification is summarised in Figure 1.

While we encourage all organisations to strive for adoption of the full standard, the priority is for each data custodian to meet the first recommendation and have a public data use register.

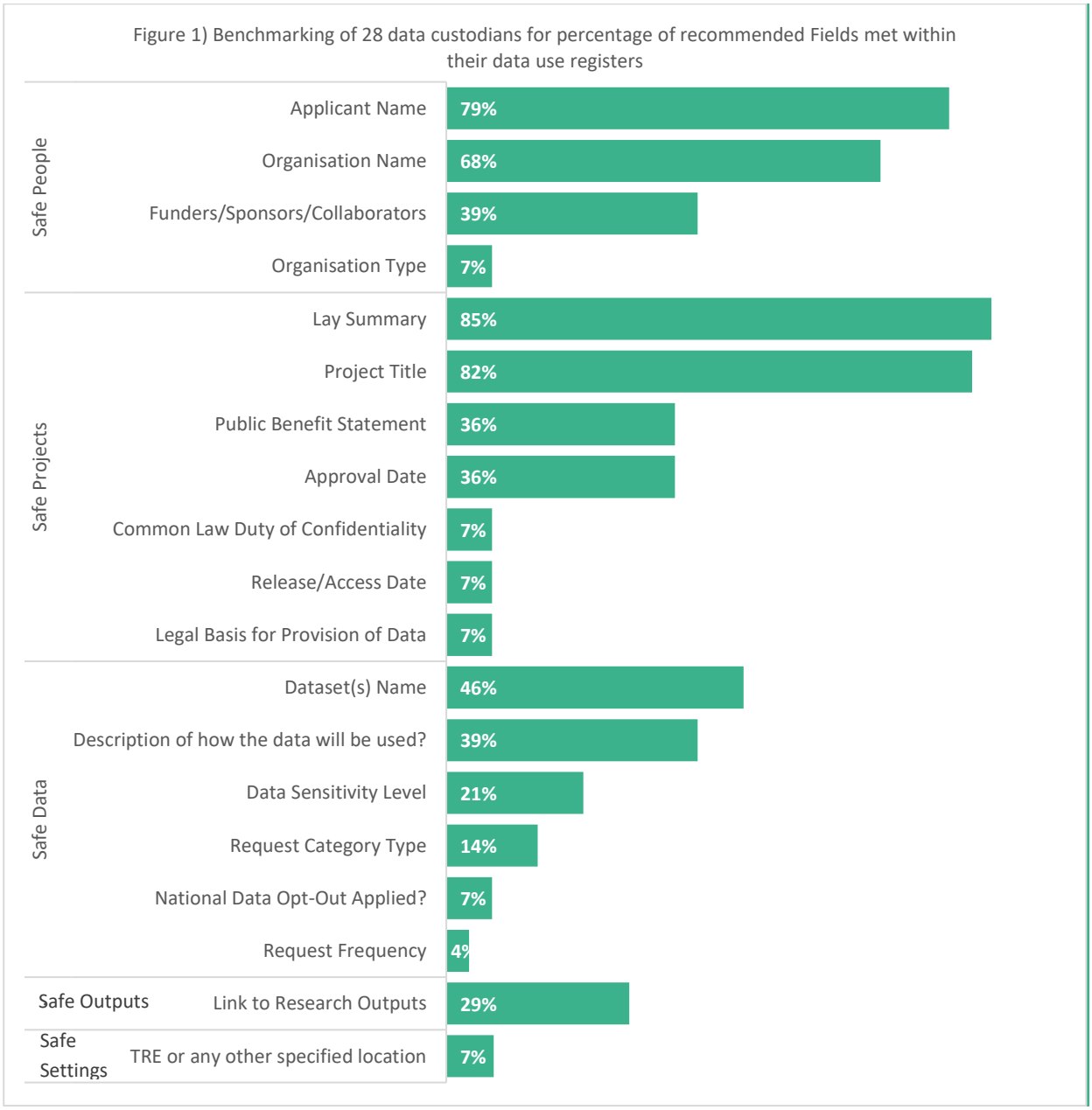

***Figure 1**: Summary of the fields that were included in the 28 Data Use Registers, by the percentage of data custodians meeting each requirement.*

UK Health Data Research Alliance

Next steps

We welcome and encourage comments and contributions to help shape this standard and align it with other related work. We are also interested to hear suggestions as to how the adoption of standards for data use registers can be influenced and what barriers need to be addressed.

A **survey** has been created alongside this paper to establish level of agreement with the recommendations and capture feedback on the draft specification and implementation challenges.

This survey will remain open for 6 weeks over the summer. All feedback and contributions will be incorporated into an updated white paper that will be shared with the Alliance Board at the October 2021 meeting with a view to publishing as soon as possible after that.

We would then intend to repeat the survey and analysis in early 2022 to understand the rate of progress.

UK Health Data Research Alliance

Appendix A: Organisations Consulted

Organisations in Attendance	Total Attendances
Anglia Ruskin University	1
Bristol University	2
Clinical Practice Research Datalink (CPRD)	5
DATA-Can	2
Edinburgh University	1
Flat Iron	1
Great Ormond Street Hospital NHS Trust	1
HIC Dundee	1
HQIP	3
Human Fertilisation & Embryology Authority	2
Imperial College London	2
Kings College London	1
Lancaster University	1
MedConfidential	2
National Neonatal Research Database (NNRD)	1
NHS Digital	4
NHSX	3
Nottingham University Hospitals NHS Trust	2
Nottinghamshire Healthcare NHS Foundation Trust	1
Our Future Health	2
Oxford Uni Hospital	1
Patient & Public Involvement Panel	15
Public Health England	2
Public Health Scotland	1
QResearch	7
Queen Mary University	1
Researcher	1
Roche	2
SAIL Databank	6
St. Bartholomew's Hospital	2
The Brain Tumour Charity	2
The Renal Registry	4
University College London	2
University Hospitals Birmingham NHS Trust	4
University of South Wales	1
Total	**89**

***Table 2**: List of organisations (including health data research organisations and universities, data custodians and patient and public groups) that were consulted through interviews and workshops held between 25 March and 13 May 2021.*