BMJ Health & Care Informatics

Development of a data utility framework to support effective health data curation

Ben Gordon ,[1] Jake Barrett,[1] Clara Fennessy,[1] Caroline Cake,[1] Adam Milward,[2] Courtney Irwin,[2] Monica Jones,[1,3] Neil Sebire[1]

To cite: Gordon B, Barrett J, Fennessy C, *et al*. Development of a data utility framework to support effective health data curation. *BMJ Health Care Inform* 2021;**28**:e100303. doi:10.1136/bmjhci-2020-100303

► Additional online supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2020-100303).

Received 10 December 2020
Revised 12 April 2021
Accepted 16 April 2021

[1]Central Team, Health Data Research UK, London, UK
[2]Data Science, MetadataWorks, London, UK
[3]Health Data Research UK hub for cancer hosted by UCLPartners, DATA-CAN, Leeds, UK

Correspondence to
Ben Gordon;
ben.gordon@hdruk.ac.uk

ABSTRACT

Objectives The value of healthcare data is being increasingly recognised, including the need to improve health dataset utility. There is no established mechanism for evaluating healthcare dataset utility making it difficult to evaluate the effectiveness of activities improving the data. To describe the method for generating and involving the user community in developing a proposed framework for evaluation and communication of healthcare dataset utility for given research areas.

Methods An initial version of a matrix to review datasets across a range of dimensions was developed based on previous published findings regarding healthcare data. This was used to initiate a design process through interviews and surveys with data users representing a broad range of user types and use cases, to help develop a focused framework for characterising datasets.

Results Following 21 interviews, 31 survey responses and testing on 43 datasets, five major categories and 13 subcategories were identified as useful for a dataset, including Data Model, Completeness and Linkage. Each sub-category was graded to facilitate rapid and reproducible evaluation of dataset utility for specific use-cases. Testing of applicability to >40 existing datasets demonstrated potential usefulness for subsequent evaluation in real-world practice.

Discussion The research has developed an evidenced-based initial approach for a framework to understand the utility of a healthcare dataset. It is likely to require further refinement following wider application and additional categories may be required.

Conclusion The process has resulted in a user-centred designed framework for objectively evaluating the likely utility of specific healthcare datasets, and therefore, should be of value both for potential users of health data, and for data custodians to identify the areas to provide the optimal value for data curation investment.

Summary

What is already known?

- The concept of data quality is well established, but the overall usefulness of a dataset for a purpose may be impacted by numerous additional factors.
- One of the stated challenges in using UK health data for research is perceived lack of useful datasets.
- There is no currently available standard framework for evaluating the utility of a healthcare dataset for a particular purpose.

What does this paper add?

- We describe the process by which a framework for understanding the usefulness of a dataset was developed.
- Information on the key characteristics related to dataset utility, based on surveys, interviews and testing, were used to provide a standard framework for dataset evaluation according to purpose.

INTRODUCTION

Health Data Research UK (HDR UK) was established to unite the UK's health data to enable discoveries that improve people's lives.[1] By making health data available to researchers and innovators, it will be possible to more rapidly develop improved understanding of disease and approaches to prevent, treat and cure them. During the establishment of HDR UK an initial 'listening exercise' was carried out, collating responses across the landscape of health data users, which reported that the major perceived barriers to use of data for research and innovation were issues regarding data access and data quality.[2]

It is generally accepted that secondary use of health data for research and development has huge potential value but a significant amount of work to improve the data will be required to make such routine data useful.[3] One difficulty is that precisely which improvements that provide most value in this context remain unknown. For example, 'quality' of datasets for most users is usually composed of a view of technical 'data quality' dimensions such as completeness, in addition to subjective assessment of factors related to a specific use case.[4]

The recently published National Data Strategy highlights the importance of high-quality data for the UK, and specifically references the lack of standardised approaches

for assessing and managing data quality in this context.[5] With the potential for future widespread investment in the general area of 'improving data', it is important that this is evidence based and focused on the areas that will have the greatest impact. A widely adopted standard tool for assessing the broad usefulness of a dataset would help to inform this future development and investment.

Previous studies are available which address specific aspects of the broad area of 'data quality' in health, but none presents a similar framework as suggested here. For example, evaluation of data quality improvement programmes are described focusing on specific quality dimensions such as accuracy and precision.[6] There are approaches described for evaluating quality of medical device data,[7] use of rule based approaches for data quality evaluation and management,[8] and outputs of workshops focusing on health data quality issues.[9] There has been a suggestion that quality informatics may become a specific area of health informatics.[10] Despite such recognition of the importance of data quality for widespread uses, evaluation of data utility for specific purposes has remained difficult.[11]

The aim of this study was to develop a proposed framework for evaluating the usefulness of a health dataset, across a range of potential use cases, to rapidly identify those which are likely to be most applicable for the specific purpose, and to provide an objective method of evaluating or categorising a dataset in order to rationally deploy data curation resources. In addition to the needs across the health data community, HDR UK in particular requires a means of determining improvement across seven publicly funded HDR Hubs: consortia of organisations involved in improving data and providing access to it for research and innovation.[12] The Hubs present multiple, parallel experiments for improving data, and so understanding their effectiveness in improving datasets for particular use cases allows for effective evaluation and could focus future investment. This presented an opportunity and need to develop a data utility framework as a service development project for HDR UK.

METHODS

The initial framework was developed based on the broad areas relating to data utility, which had been identified from previously published evidence.[13] We adopted a user centred co-design approach which was designed to result in an understanding of the areas of user interest in data usefulness.[14] Given the absence of consensus on an existing approach, it was important to gain a broad understanding of the topic, so a combination of interviews, surveys and user testing was used to help achieve diversity of inputs into the development of the framework, in line with standard practice in user centred design.[15 16]

We, therefore, focused on the major issues relating to health dataset utility by interviewing a range of data users in the domain, followed by collation of additional views from key stakeholders across the community through a survey consultation process. This was done to create and discover the main areas of interest for such a framework across a range of user groups and use cases. These findings were then used to further refine and develop a proposed Data Utility Framework that could be subsequently used and iterated by the community.

The framework aimed to indicate the 'utility' of a dataset for researchers for a particular purpose based on a set of key attributes, which would be classified using predefined criteria into subsequent arbitrary qualitative categories (bronze, silver, gold and platinum). Principles for the framework included that the bronze-platinum categories should differentiate utility for any given dimension in a progressive manner that each dimension should have objective criteria that will allow users to determine a utility score (initially through self-evaluation) that a greater score can only be achieved if the dataset meets the criteria of the previous categories as well as an additional criterion for the new score and that there is no expectation that all datasets should achieve a particular classification since some use cases may only require a minimum standard, while others may require greater utility scores for their purposes.

For interviews, the interviewee's particular data requirements were used as a basis for discussion, and users articulated the ways by which they determine the utility of a dataset for their particular use case, that is, how they determined if a dataset was 'useful' for their purpose. They were specifically asked their views regarding various components of an initial proposed framework which was based on the previous HDR UK scoping information (online supplemental table 1). The number of times interviewees refer to specific components of the draft framework were quantified, to capture interest in the items as presented. This methodology is a common user-centred design approach, to identify features that a range of individuals would like to see represented.[17]

Interviewees were selected based on a segmented sample to ensure representation from multiple sectors, including artificial intelligence (AI)/tech firms, large pharmaceutical companies, National Health Service/data custodians and academics (figure 1). Forty individuals were contacted to request an interview. Of these, 8 did not respond, 10 declined to be interviewed, 3 were unable to schedule a time and 21 were finally interviewed. Interviews were held in April and May 2020 using an online meeting platform, and these were recorded to support transcription of the discussion.

Interviews were semistructured (online supplemental table 2). Questions were sent at least 24 hours before the interview, in addition to the initial framework. Consent of the participants was taken from their initial agreement to participate in the interview process—all were informed of the nature of the project in developing the tool.[18] As the approach to develop the framework was service development work and did not include patients or staff in their clinical roles, ethical approval was not required—as confirmed by the Health Research Authority.[19]

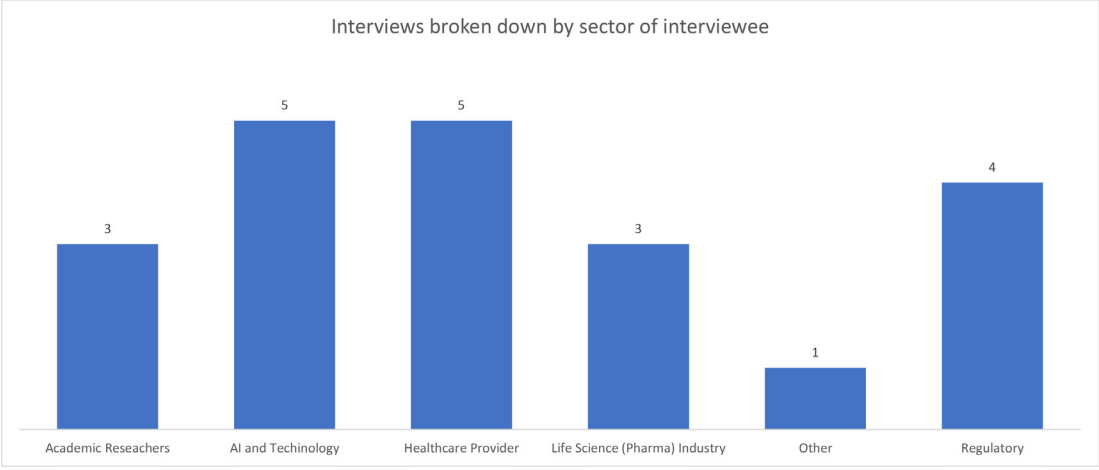

Figure 1 Bar chart showing breakdown of interviewees by sector.

The interviewees were asked to comment on the importance of dimensions within the proposed framework, and to suggest any dimensions which were not originally included, but were not otherwise directed and were free to discuss whichever aspects they felt most important.

Qualitative content analysis was used on the outputs of interviews to establish the relative interest in the various dimensions. This included an estimation of the categories for the proposed 'medallion' ratings (online supplemental table 3).

In June 2020, a survey was issued to the interviewee list, as well as HDR UK's Data Officer Community, with a request for all recipients to share with their own contacts. The survey (online supplemental table 4) requested input on the revised matrix. Responses from 30 individuals were received with some respondents spread across multiple sectors (figure 2). The content analysis was repeated to identify refinements and develop the second version of the matrix (online supplemental table 5).

Following the survey, the second version of the framework was included in a wider consultation, which was publicly open online from August to September 2020. In addition, the second version of the framework was applied to 43 existing datasets across seven HDR Hubs, with feedback provided from each team on the potential suitability and applicability of the framework. These datasets include routinely collected clinical data, genomic data, national datasets and imaging datasets. The feedback from this process was used to develop the final version of the framework but given that no existing 'gold standard' framework existed, formal testing of performance was not carried out as part of this process.

RESULTS

Participant characteristics

The interviewees represented a range of different sectors (figure 1). All were required to use health data for research, or support others in doing so, as part of their professional context.

Original framework comments

All interviews (21) emphasised the importance of a comprehensive dataset metadata, describing the nature

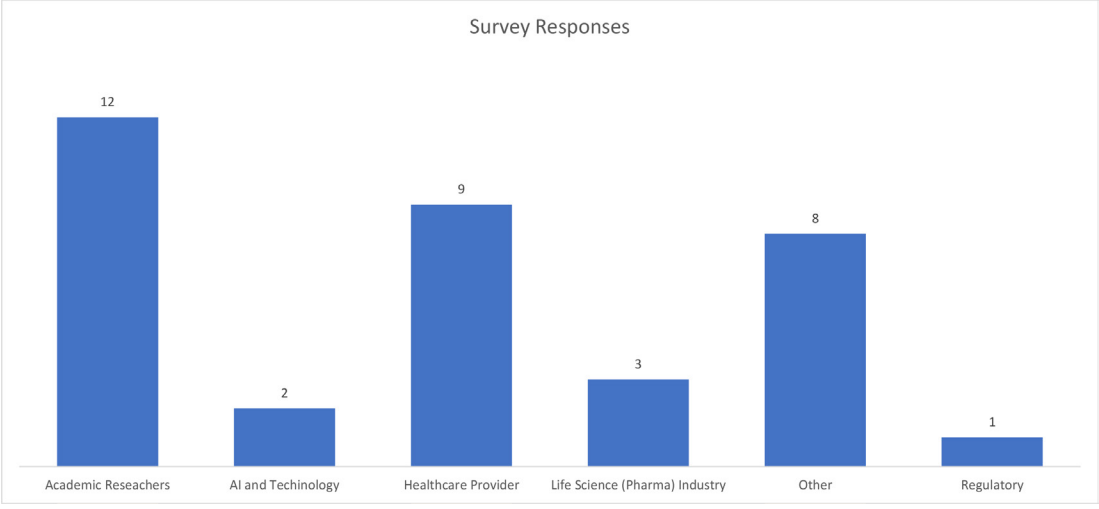

Figure 2 Bar chart showing breakdown of survey respondents by sector.

and scope of the data collection. This information enabled users to identify the utility and relevance of a dataset for specific use cases. A number of interviews (nine) emphasised the importance of a readily available data dictionary and ability to interrogate the dataset at a data element level. Beyond typical dataset descriptions, users emphasised the requirement to understand data provenance, especially in the case of data consolidated from multiple sources, where the provenance may differ across data elements. The number of mentions of each dimension of the original framework is provided in table 1 (note that these are a reflection of interest, rather than support, as they may be comments in support of the dimension or disagreement):

Description: characteristics and service

Interviewees drew little distinction between dimensions in the description and characteristics and service categories, as many of the elements relate to the information that is available about the dataset, known as the metadata. Several users, particularly from industry, noted the importance of upfront clarity regarding the uses applicable for the dataset, something that is not currently widely available. A number of interviewees described the available of additional resources and information on the dataset including previous academic publications, documentation of frequently asked questions and a contact for whom they could reach with questions as a key factor of data utility.

Beyond descriptive metadata, 'access' was a key consideration for data users sourcing data from outside their organisations. Users with commercial use cases (including companies developing and supplying data and machine learning products, and pharmaceutical companies) wanted indicators of which datasets were available for commercial use, in order to minimise enquiries into unsuitable data sources. Researchers noted that the amount of time required to gain approval for data access was a deciding factor in selecting datasets (and sometimes research questions) and commercial data users also highlighted this as a risk. The means of data access; (eg, direct download, use of a secure environment, via an internal analysis team) was noted as an important factor impacting time and costs for commercial organisations accessing data. This emphasis led to the creation of a separate category focusing on service.

The service category was subject to significant attention at the testing stage. Respondents noted that clarity was required on the role of the research environment, as well as refining the wording on the categories for timeliness.

Scale

The initial elements within the Scale category were perceived to be useful by interviewees. Some interviewees (3) mentioned the pathway coverage dimension and those from pharmaceutical background indicated that the longitudinal patient journey helped explain health outcomes. Many survey respondents (14 of the 20) who indicated their feelings deemed this element as either important or very important.

While nine interviewees specifically commented on the 'coverage' field, interview questions relating to the number of expected items did not yield meaningful answers due to the significant variability across use cases. Therefore, the number of entries element was excluded from the final matrix. Additionally, pharmaceutical companies wanted details of coverage on a specific number of patients meeting multiple requirements. Duration was excluded since as feedback suggested that 'length of patient follow-up' was an element of particular interest. Depth was also excluded since users valued the details of what was measured rather than the number of things being measured. Missing data and missing data handling was a topic of interest but was excluded from the final framework since it was difficult to agree the optimal method, but suggestions that this be included in the additional documentation and support section of the metadata.

'Technical' quality dimensions

Throughout the surveys and interviews, the aspects of 'technical' data quality, defined here as those listed by the Data Management Association (DAMA), were seen as important, mentioned by 11 interviewees, but no indication could be given of specific required or expected levels for the dimensions. This led to the supplementation of the DAMA dimensions (Completeness, Uniqueness, Timeliness, Validity, Accuracy and Consistency) with an additional element relating to the data management process itself, which was supported by several (six) interviewees. This additional data management process element was considered by all but one survey respondent to be either important (14) or very important (7).

Added value

The additional information available about a dataset was generally considered to be of importance by interviewees and respondents. The majority of discussions in this area related to the ability to link the dataset with others, with this being mentioned 10 times by interviewees. Many survey respondents (12) felt that this element was very important. Respondents also commented that data enrichment was important (13) but there was no direct discussion on this dimension from the interviews. Current usage was relatively useful but was not mentioned directly by interviewees or in the respondent's comments. Data access requests were said to be a useful indicator of the level of utility of the dataset and more likely to be understandable with a functional and operational data governance process. There were mixed feelings among the respondents for this dimension deeming it low importance. For the provenance of access request dimension where there were only two specific comments in each of the interview and survey responses. The added value section was reduced in scope for the final matrix given this response.

Table 1 Table showing number of interview respondents commenting on each item from the original framework, to gain an understanding of an end users view of the initial framework categories

Category from initial framework	Dimension	Definition	Times mentioned in interviews
Description	Metadata completeness	Level of metadata completed	2
	Metadata quality	Richness of metadata completion—including within required formats and quality of qualitative fields	5
Characteristics and Service	Data source	The modality or source of data (eg, Electronic Health Record, study specific)	14
	Data model	The data model or schema used by the dataset (eg, Observational Medical Outcomes Partnership (OMOP), Informatics for Integrating Biology and the Bedside (i2b2))	16
	Data dictionary	Provided documented data dictionary and terminologies	9
	Provenance	The original source or jurisdiction of the dataset	
	Usage restrictions	The df to use the data for different purposes (eg, commercial licences, consent, expiry)	6
	Format	The technical presentation of the data format (eg, Digital Imaging and Communications in Medicine (DICOM) images vs Portable Graphics Format (PNG))	3
	Timeliness	How quickly the data can be provided—in a useful timescale	7
	Fairness of the data	Extent to which the data are findable, accessible, interoperable and reusable	0
	Phenome	Extent and description of included patients/ conditions (links with Phenome work re standards)	2
Scale	Coverage	No of individuals, data points, lab tests, images, etc included in the dataset	9
	Duration	Length of time to which the data relates	3
	Depth	Amount of information available per individual (eg, number of fields/records, types of data)	3
Quality	Completeness	The proportion of data entries that should be populated are populated (and inverse—proportion that should not be populated are not)	11
	Missing data handling	Description of missing value handling and default values	4
	Consistency/uniformity	Data are presented in the required format and a similar wayfor example, field types, date formats	1
	Uniqueness	Lack of duplication	3
	Validity	Data are valid based on acceptable 'rules' for example, age between 0 and 120, pregnancy in male patients, physiological readings within normal ranges	7
	Accuracy/verification	The extent to which the data reflects the 'real-world', for example, level of certainty that fields are accurate	6
	'Usefulness'	Qualitative, subjective measure by user (eg, Net Promoter Score/star rating)	12

Continued

Table 1 Continued

Category from initial framework	Dimension	Definition	Times mentioned in interviews
Added value	Linkage/mapping	Ability to link with other datasets	10
	Transformations/ derivations	Level of derived data and descriptions, manual versus Natural Language Processing, etc	1
	Accuracy/verification	Level of manual verification/sampling	0
	Annotation	Additional fields added to provide further information, including phenotyping	1

The final proposed version of the data utility framework following the user design feedback is provided in figure 3.

DISCUSSION

The findings of this study have provided an evidence-based initial approach for a proposed framework for identifying and understanding the main factors that are related to the utility of a healthcare dataset for secondary purposes across a range of industries and use cases. This has allowed development of a data quality matrix and classification system, which will be further refined through testing and implementation. The real-world usefulness of such a framework will be evaluated following implementation and feedback from users as well as usage statistics of particular datasets in relation to their framework score and classification. It is believed to be the first successful attempt to create a semi-structured framework for characterising health datasets on usability and allowing users, regardless of their specific use-case, to identify in advance whether a dataset would be useful for their purposes.

The main strengths of the study are that this is a practical and robust approach to a problem that has been theoretically reviewed but not addressed in an objective manner previously. The range of different stakeholders, through multiple cycles, and the repeated improvement and testing have allowed for the development of a framework that is able to be adapted to different data types and able to be implemented in real-life situations.

However, by necessity, this approach has had to use a non-random group of respondents, due to the process and potential selection bias of the survey respondents. The sampling strategy was developed based on a systematic but pragmatic approach to collect a range of views to be gathered from different organisations and sectors to ensure all main stakeholder groups were represented. Any potential bias could lead to a matrix that was incomplete, or had unnecessary emphasis on particular categories, and only large-scale feedback will determine these aspects. The continuous development of the matrix during the process was appropriate given the pragmatic nature of the project in the absence of an existing standard practice, but a more structured approach would have identified in advance the stages and cycles of development. This methodological limitation impacts the ability of the matrix in its present form to be used for other data contexts without further development and here we make no claims regarding cross-sector applicability. Given these limitations, future iterations of the framework are likely to evolve from the original version presented here based on real world usage and feedback. However, further research is required to identify the 'effectiveness' of the framework in terms of identifying areas to address through data improvement activities and then demonstrating that usability of the dataset increased subsequently. It is not known at this stage the power of the framework or its potential use in other circumstances.

The framework is likely to require further development as it is tested on an increasingly wide range of health datasets. This development will take several forms: the refinement of the existing categories, the addition of new categories, integration into tools and the creation of extensions. The refinement of the existing categories will continue as more feedback is received through the testing process. One key development is likely to be 'normalisation' of the categories—currently the breakdown is based on the surveys and interviews, as well as initial testing on >40 datasets. As the tool is applied to many more datasets, it may be necessary to adjust the categories to ensure an appropriate distribution across the existing categories in the UK health landscape.

Similarly, it is possible that further categories may be added. For example, during the development, a category on 'Research Environment' was proposed. However, an inability to reconcile the varying tensions between the current status of research environments across UK data custodians, user requests from particular sectors and the principles for Trusted Research Environments, this category was not included in the final version of the framework.

The integration of the framework into tools will allow for its value to be realised in practice. In isolation, it provides a view of a given dataset, however, the power of such a framework approach comes with the ability for an individual to specify their requirements in a catalogue, such as the HDR Innovation Gateway, and use their requirements from the framework to remove the datasets which would not be fit for their purposes. The framework was deliberately designed to be able to be applied in a

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
Data Documentation	Documentation Completeness	Proportion of metadata (as in the current metadata specification) which is available in the expected format	This element will be calculated automatically based on the level of metadata available on the Gateway, and values set for each category			
	Availability of additional documentation and support	Available dataset documentation in addition to the data dictionary	Past journal articles demonstrate that knowledge of the data exists	Comprehensive ReadMe describing extracting and use of data, Dataset FAQS available, Visual data model provided	As Silver, plus dataset publication was supported with a journal article explaining the dataset in detail, or dataset training materials	As Gold, plus support personnel available to answer questions
	Data Model	Availability of clear, documented data model	Known and accepted data model but some key field un-coded or free text	Key fields codified using a local standard	Key fields codified using a national or international standard	Data Model conforms to a national standard and key fields codified using a national / international standard
	Data Dictionary	Provided documented data dictionary and terminologies	Data definitions available	Definitions compiled into local data dictionary which is available online	Dictionary relates to national definitions	Dictionary is based on international standards and includes mapping
	Provenance	Clear description of source and history of the dataset, providing a "transparent data pipeline"	Source of the dataset is documented	Source of the dataset and any transformations, rules and exclusions documented	All original data items listed, all transformations, rules and exclusion listed and impact of these	Ability to view earlier versions, including versions before any transformations have been applied data (in line with deidentification and information governance approval) and review the impact of each stage of data cleaning
Technical Quality	Data Quality Management Process	The level of maturity of the data quality management processes	A documented data management plan covering collection, auditing, and management is available for the dataset	Evidence that the data management plan has been implemented is available		Externally verified compliance with the data management plan, e.g. by International Organization for Standardization (ISO), Care Quality Commission (CQC), Information Commissioner's Office (ICO) or other body
	Data Management Association (DAMA) Quality Dimensions	Technical data quality dimensions: Completeness, Uniqueness, Accuracy, Validity, Timeliness and Consistency	These elements will be calculated with data profiling tools, and the category breakdown evaluated following further data collection			
Coverage	Pathway coverage	Representation of multi-disciplinary healthcare data	Contains data from a single speciality or area	Contains data from multiple specialties or services within a single tier of care	Contains multimodal data or data that is linked across two tiers (e.g. primary and secondary care)	Contains data across more than two tiers
	Length of follow up	Average timeframe in which a patient appears in a dataset (follow up period)	Between 1 - 6 months	Between 6 - 12 months	Between 1 - 10 years	More than 10 years
Access & Provision	Allowable uses	Allowable dataset usages as per the licencing agreement, following ethical and information governance approval	Available for specific academic research uses only	Available for academic and non-profit (e.g. charity, public sector) uses only	Available for limited commercial uses (e.g. relating to a specific domain), in addition to academic and other non-commercial uses	Available for wider commercial uses (in line with ethical and information governance approval), and addition to academic and other non-commercial uses
	Time Lag	Lag between the data being collected and added to the dataset	Approximately 1 year	Approximately 1 month	Approximately 1 week	Effectively real-time data
	Timeliness	Average data access request timeframe	Less than 6 months	Less than 3 months	Less than 1 month	Less than 2 weeks
Value & Interest	Linkages	Ability to link with other datasets	Identifiers to demonstrate ability to link to other datasets	Available linkages outlined and/or List of datasets previously successfully linked provided	List of restrictions on the type of linkages detailed. List of previously successful dataset linkages performed, with navigable links to linked datasets via a Digital Object Identifier (DOI) or Uniform Resource Locator (URL)	Existing linkage with reusable or downstream approvals
	Data Enrichments	Data sources enriched with annotations, image labels, phenomes, derivations, Natural Language Processing (NLP) derived data labels	The data include additional derived fields, or enriched data.	The data include additional derived fields, or enriched data used by other available data sources.	The derived fields or enriched data were generated from, or used by, a peer reviewed algorithm.	The data includes derived fields or enriched data from a national report.

Figure 3 Final version of the proposed data utility framework based on data user feedback.

general manner, however, it was solely tested on datasets and users relating to biomedical data and applications. Further work is required to explore whether it can be usefully applied to other data types, and if specific extensions are required for certain data modalities, such as imaging data.

CONCLUSION

In conclusion, we have proposed a codesigned and evidence-based health dataset utility framework, for potential widespread evaluation and use, which will be integrated into HDR UK's ambitions to make health data more useful for research and enable discoveries that improve people's

lives. The tool will be tested on datasets currently discoverable through the Innovation Gateway, and feedback from this process will be used to refine the framework as applicable to the range of potential use cases. Such as framework may also provide objective evidence to demonstrate the benefit of specific data improvement and 'curation' activities, including the potential for providing a return-on-investment type understanding of work in data.

Acknowledgements The authors gratefully acknowledge the contributions of all individuals who provided their input into this work through the interview, survey and testing phase.

Contributors BG, NS and CC were involved in the planning of the study. CI and AM conducted the user interviews and designed the survey. BG, NS and MJ oversaw the outputs of the user design and developed the iterations of the framework. BG, NS, JB and CF developed the project outputs into the paper. The guarantor of the content is NS.

Funding This work was supported by Medical Research Council capital funding (August 2019).

Competing interests MetadataWorks (AM and CI) were funded by HDR UK to conduct the initial information gathering. No competing interests are declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

ORCID iD

Ben Gordon http://orcid.org/0000-0002-3105-6774

REFERENCES

1 HDR UK One Institute Strategy. Health data research UK, 2020. Available: https://www.hdruk.ac.uk/wp-content/uploads/2019/11/191010-HDR-UK-One-Institute-Strategy-compressed-for-website.pdf
2 Health Data Research UK. Over 3,000 people, 30 locations and counting. Available: https://www.hdruk.org/news/over-3000-people-and-30-locations-and-counting/ [Accessed Oct 2020].
3 Ersnt & Young. Realising the value of health care data: a framework for the future EYGM; 2019.
4 Kahn MG, Callahan TJ, Barnard J, *et al*. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS* 2016;4:18.
5 Department for Digital, Culture, Media and Sport. National data strategy. Available: https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy [Accessed Oct 2020].
6 Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: a report of practical experience with the Canadian primary care sentinel surveillance network data. *Health Inf Manag* 2021;50:88-92.
7 Mavrogiorgou A, Kiourtis A, Perakis K, *et al*. Analyzing data and data sources towards a unified approach for ensuring end-to-end data and data sources quality in healthcare 4.0. *Comput Methods Programs Biomed* 2019;181:104967.
8 Wang Z, Dagtas S, Talburt J, *et al*. Rule-Based data quality assessment and monitoring system in healthcare facilities. *Stud Health Technol Inform* 2019;257:460–7.
9 Huser V, Kahn MG, Brown JS, *et al*. Methods for examining data quality in healthcare integrated data repositories. *Pac Symp Biocomput* 2018;23:628–33.
10 Coppersmith NA, Sarkar IN, Chen ES. Quality informatics: the convergence of healthcare data, analytics, and clinical excellence. *Appl Clin Inform* 2019;10:272–7.
11 Keller S, Korkmaz G, Orr M, *et al*. The evolution of data quality: understanding the Transdisciplinary origins of data quality concepts and approaches. *Annu Rev Stat Appl* 2017;4:85–108.
12 NHSX. Welcoming new health data research hubs. Available: https://digital.nhs.uk/blog/transformation-blog/2019/welcoming-new-health-data-research-hubs [Accessed Oct 2020].
13 Parkinson J. *The Data Quality Blueprint: A Practical and Holistic Approach: A Comprehensive Step by Step Guide to an Effective & Long Lasting Enterprise-Wide Data Quality Solution*. Sutton Coldfield: Holifast Limited, 2016.
14 Sandler I, Ostrom A, Bitner MJ, *et al*. Developing effective prevention services for the real world: a prevention service development model. *Am J Community Psychol* 2005;35:127–42.
15 . The field guide to Human-Centered design: design kit. San Francisco, Calif IDEO; 2015.
16 Marien S, Legrand D, Ramdoyal R, *et al*. A User-Centered design and usability testing of a web-based medication reconciliation application integrated in an eHealth network. *Int J Med Inform* 2019;126:138–46.
17 Stuij SM, Drossaert CHC, Labrie NHM, *et al*. Developing a digital training tool to support oncologists in the skill of information-provision: a user centred approach. *BMC Med Educ* 2020;20:135.
18 UK Data Service. Consent for data sharing. Available: https://ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing/surveys.aspx [Accessed Feb 2021].
19 HRA. Is my study research? 2019. Available: http://www.hra-decisiontools.org.uk/research/ [Accessed Feb 2021].

Online Supplemental Table 1. Framework as at March 2020

Category	Dimension	Definition
Description	Metadata Completeness	Level of metadata completed
	Metadata Quality	Richness of metadata completion – including within required formats and quality of qualitative fields
Scale	Coverage	Number of individuals included in the dataset
	Duration	Length of time to which the data relates
	Depth	Number of information available per individual (e.g. number of fields)
Service	Format	The presentation of the data – to be presented in useful formats / interoperable standards
	Timeliness	How quickly the data can be provided – in a useful timescale
Quality	Completeness	The proportion of data entries that should be populated are populated (and inverse – proportion that should not be populated are not)
	Consistency / Uniformity	Data are presented in the required format and a similar way – e.g. field types, date formats
	Uniqueness	Lack of duplication
	Validity	Data are valid based on acceptable “rules” e.g. age between 0 and 120, pregnancy in male patients, physiological readings within normal ranges
	Accuracy / Verification	The extent to which the data reflects the “real-world”, e.g. level of certainty that fields are accurate
	“Usefulness”	Qualitative, subjective measure by user (e.g. NPS / star rating)
Added Value	Linkage / Mapping	Ability to link with other datasets
	Annotation	Additional fields added to provide further information, including phenotyping

Online Supplemental Table 2. Interview questions regarding the initial proposed framework

Category	Question
Main focus of the discussion	Can you tell us about an example of a particularly useful or high-quality dataset?
	How did you make use of this dataset?
	Why do you consider that it had these attributes?
	Alternatively, could you recall an experience with an un-useful or low-quality dataset; and the reasons you considered it to be of this nature?
Other points to discuss	As a data user, how do you imagine that a data utility framework/metrics/scores might be able to serve you, support you to make better use of datasets on the innovation gateway?
	Can you describe to us how you might imagine this information working?
	Based on your review of the framework, are there any dimensions that stand out to you as useful/not useful? Why/why not?
	Are there any dimensions that you have questions about, or aren't self-explanatory?
	For your own data needs and use cases, which dimensions would you consider most important/least important?
	For each dimension on the list, how would you rate their importance in terms of understanding data quality
	Are there any other utility or quality dimensions that you would add to this list?
	Data Users in your network: We are extending our engagements, and keen to speak with data users to get their feedback and ideas on data utility. Do you have 3-4 data users in your network you could put us in touch with to interview or survey?
Specific clarification points	**Format**: A particular element that HDR UK are keen to understand is whether your organisation would be able to comply with a requirement to provide data according to either a standard model and format (e.g. OMOP, or a requirement to make data available through a FHIR API) would the organisation be able to do this now? If not, what would be required for you to be able to do this? Is this already on organisational roadmaps? Is it something that could be feasible within a year or two? Or impossible without significant additional investment?
	Coverage: The suggestion here is "Number of individuals included in the dataset". What would be your requirements in terms of quickly understanding coverage (e.g. number of observations, sites etc)?
	Usefulness: In what format would subjective user feedback on the dataset be useful to you? Reviews? Five-star ratings?

	Validation or Transformation: Level of manual "cleaning" and Annotation: Additional fields added to provide further information, including phenotyping: General feedback on data quality often features statements such as the above, which doesn't specify the outcome. What specific indicators would be useful to you in relation to these statements?

Online Supplemental Table 3. Framework as at May 2020

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
Data Documentation	Documentation Quality	Weighted Data Documentation Score	< 66% of metadata specification fields meet format and content requirements	< 76% of metadata specification fields meet format and content requirements	< 86% of metadata specification fields meet format and content requirements	< 96% of metadata specification fields meet format and content requirements
	Availability of additional documentation and support	Available dataset documentation in addition to the data dictionary	Past journal articles demonstrate that knowledge of the data exists	Comprehensive ReadMe describing extracting and use of data, Dataset FAQS available, Visual data model provided	Dataset publication was supported with a journal article explaining the dataset in detail, or dataset training materials	Support personnel available to answer any questions
	Data Model	Availability of clear, documented data model	Known and accepted data model but some key field uncoded or free text	Key fields codified using a local standard	Key fields codified using a national or international standard	Data Model conforms to a national standard and key fields codified using a national / international standard
	Data Dictionary	Provided documented data dictionary and terminologies	Data definitions available	Definitions compiled into local data dictionary which is available online	Dictionary relates to national definitions	Dictionary is based on international standards and includes mapping(?)
	Provenance	Clear description of source and history of the dataset, providing a "transparent data pipeline"	Source of the dataset is documented	Source of the dataset and any transformations, rules and exclusions documented	All original data items listed, all transformations, rules and exclusion listed and impact of these	Ability to view earlier versions, including "raw" or "source" dataset, and review the impact of each stage/step
Technical Quality	Data Quality Management Process	The level of maturity of the data quality management processes	A documented data management plan covering collection, auditing, and management is available for the dataset	Evidence that the data management plan has been implemented is available	Demonstrated compliance with the data management plan	Externally verified compliance with the data management plan
	DAMA Data Quality Element - Completeness	The proportion of stored data against the potential of "100% complete"		*Indicators are under development by another project*		
	DAMA Data Quality Element - Uniqueness	No thing will be recorded more than once based upon how that thing is identified.				

	DAMA Data Quality Element - Timeliness	The degree to which data represent reality from the required point in time.				
	DAMA Data Quality Element - Validity	Data are valid if it conforms to the syntax (format, type, range) of its definition.				
	DAMA Data Quality Element - Accuracy	The degree to which data correctly describes the "real world" object or event being described.				
	DAMA Data Quality Element - Consistency	The absence of difference, when comparing two or more representations of a thing against a definition.				
Coverage	Pathway coverage	Representation of Multi disciplinary healthcare data	Contains data from a single speciality or area	Contains data from multiple specialties or services within a single tier of care	Contains multimodal data or data that is linked across two tiers (e.g. primary and secondary care)	Contains data across the whole pathway of care
	Length of follow up	Average timeframe in which a patient appears in a dataset (follow up period)	Between 1 - 6 months	Between 6 - 12 months	Between 1 - 10 years	More than 10 years
Access & Provision	Allowable uses	Allowable dataset usages as per the licencing agreement		Non-consented, aggregate data for specific academic uses	Aggregate data, for specific commercial use	Fully consented for commercial uses
	Research environment	Access, tooling and environment (once approved)	Requested analysis can be undertaken by internal teams and provided back in anonymised format to data requestors	The dataset can be used in a trusted research environment	The dataset can be used in a trusted research environment, but other data and tools can be brought in as required	The dataset can be used in requesting companies environment
	Format	The technical presentation of the data format (e.g. DICOM images vs PNG)	Format is explicitly defined	Format is explicitly defined and in widely readable (non-proprietary) format		Format is explicitly defined, in open and referenced format

	Access and approvals	IG Process for gaining approval to access data	Contact details for the relevant authority to request data access	Indicative timeframes for processing data access applications detailed	Details of application process and requirements	Detailed Information Governance Process described, outlining requirements for applications, basis of decisions, and anticipated timeframes
	Timeliness	Data time lag + Average data access request timeframe	More than 12 months	Less than 12 months	Less than 6 months	Less than 3 months
Value & Interest	Access Request	Number of data access inquires received	Some interest. Access inquires/requests >1	Some interest. Access inquires/requests >5	Some interest. Access inquires/requests >20	Some interest. Access inquires/requests >50
		Provenance of data access inquires received	Interest and inquires from colleagues and known associates	Interest and inquires from other national organisations	Interests and inquires from national commercial organisations	Interests and inquires from international bodies
	Current Usage	Number of active projects and tools using the dataset	> 1 active projects or tools using the dataset	> 5 active projects or tools using the dataset	> 10 active projects or tools using the dataset	> 30 active projects or tools using the dataset
	Linkages	Ability to link with other datasets	Identifiers to demonstrate ability to link to other datasets	Available linkages outlined and/or List of datasets previously successfully linked provided	List of restrictions on the type of linkages detailed. List of previously successful dataset linkages performed, with navigable links to linked datasets via at DOI/URL	Existing linkage with reusable or downstream approvals
	Data Enrichments	Data sources enriched with annotations, image labels, phenomes, derivations, NLP derived data labels	The data include additional derived fields, or enriched data.	The data include additional derived fields, or enriched data used by other available data sources.	The derived fields or enriched data were generated from, or used by, a peer reviewed algorithm.	The data includes derived fields or enriched data from a national report.

Online Supplemental Table 4. Survey questions with their responses.

<table>
<tr><th>Survey Question</th><th>Response Type/Options</th></tr>
<tr><td>1. Your name (optional)</td><td>[Free text]</td></tr>
<tr><td>2. The organisation where you work (optional)</td><td>[Free text]</td></tr>
<tr><td>3. The industry you work in</td><td>◯ HealthCare Provider
◯ Academia
◯ Technology & Artificial Intelligence
◯ Regulation
◯ Pharmaceutical Industry
◯ Other (please specify) [Free text]</td></tr>
<tr><td>4. How do you make use of health care data in your role?</td><td>[Free text]</td></tr>
<tr><td>5. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
<tr><td>6. Please indicate your thoughts on the importance of each dimension of the data documentation score</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Document Quality</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Data Dictionary</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Additional Documentation and Support</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Data Model</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Provenance</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>7. About the Weighted Data Documentation score:
The weighted data documentation score measures the completeness and quality of the metadata describing the dataset. The score, from 0-100, represent the percentage of metadata fields that are complete and meet the quality requirements. The completeness and quality score for each field of the metadata will be "weighted" based on the comparative utility or usefulness of each of the fields contained in the metadata.
If you are interested in finding out more about the weighted data documentation score and contributing your thoughts to the weightings, please follow the link to this short (approx 3 mins) supplementary survey: https://www.surveymonkey.co.uk/r/HDRUK_Data_Documentation_Score
General thoughts can be left below.</td><td>[Free text]</td></tr>
<tr><td>8. Above are other dimensions considered, but not included in this proposed framework. Please include any thoughts or comments below.</td><td>[Free text]</td></tr>
<tr><td>9. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
</table>

<table>
<tr><td>10. Please indicate the importance of each of the dimensions in the technical quality score</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Data Quality Management Process</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>11. In addition to this measure, another project within HDRUK is looking to develop tools to measure datasets against the six quality measures developed by the DAMA UK Working Group (as per above). These will be released at a later stage and incorporated into the overall measurements associated with the data utility framework.
Please include any thoughts or comments below.</td><td>[Free text]</td></tr>
<tr><td>12. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
<tr><td>13. Please indicate your thoughts on the importance of this dimension</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Pathway Coverage</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Length of Follow up</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>14. Above are other dimensions considered, but not included in this proposed framework. Please include any thoughts or comments below.</td><td>[Free text]</td></tr>
<tr><td>15. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
<tr><td>16. Please indicate your thoughts on the importance of these dimensions</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Allowable uses</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Research Environment</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Format</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Access and Approvals</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Timeliness</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>17. Above are other dimensions considered, but not included in this proposed framework. Please include any thoughts or comments below.</td><td>[Free text]</td></tr>
<tr><td>18. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
</table>

<table>
<tr><td>19. Please indicate your thoughts on the importance of these dimensions</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Number of Access Requests</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Provenance of Access Requests</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Current Usage</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Linkages</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Data Enrichments</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>20. Any final comments can be added below</td><td>[Free text]</td></tr>
<tr><td>21. OPTIONAL:
If you are interested in receiving feedback on the results of the survey, you can leave your email in the box below. Note: your email address will only be used for the purposes of sending through feedback once available.</td><td>[Free text]</td></tr>
</table>

Online Supplemental Table 5. Framework as at August 2020

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
Data Documentation	Documentation Quality	Weighted Data Documentation Score	< 66% of submitted metadata fields meet format and content requirements of the metadata specification	< 76% of submitted metadata fields meet format and content requirements of the metadata specification	< 86% of submitted metadata fields meet format and content requirements of the metadata specification	< 96% of submitted metadata fields meet format and content requirements of the metadata specification
	Data Model	Availability of clear, documented data model	Known and accepted data model but some key field un-coded or free text	Key fields codified using a local standard	Key fields codified using a national or international standard	Data Model conforms to a national standard and key fields codified using a national / international standard
	Data Dictionary	Provided documented data dictionary and terminologies	Data definitions available	Definitions compiled into local data dictionary which is available online	Dictionary available online, and relates to national definitions	Dictionary is based on international standards and includes mapping
	Provenance	Clear description of source and history of the dataset, providing a "transparent data pipeline"	Source of the dataset is documented	Source of the dataset and any transformations, rules and exclusions documented	All original data items listed, all transformations, rules and exclusion listed and impact of these	Ability to view earlier versions, including "raw" dataset, and review the impact of each stage of data cleaning
Technical Quality	Data Quality Management Process	The level of maturity of the data quality management processes	A documented data management plan covering collection, auditing, and management is available for the dataset	Evidence that the data management plan has been implemented is available		Externally verified compliance with the data management plan
	DAMA Data Quality Element - Completeness	The proportion of stored data against the potential of "100% complete"	*Set initial figures for this*			
	DAMA Data Quality Element - Validity	Data are valid if it conforms to the syntax (format, type, range) of its definition.				
Coverage	Pathway coverage	Representation of multi-disciplinary healthcare data	Contains data from a single speciality or area	Contains data from multiple specialties or services within a single tier of care	Contains multimodal data or data that is linked across two tiers (e.g. primary and secondary care)	Contains data across the whole pathway of care
	Length of follow up	Average timeframe in which a patient appears in a dataset (follow up period)	Between 1 - 6 months	Between 6 - 12 months	Between 1 - 10 years	More than 10 years
Access & Provision	Allowable uses	Allowable dataset usages as per the licencing agreement		Non-consented, aggregate data for specific academic uses (following IG approval)	Aggregate data, for academic and specific commercial uses (following IG approval)	Fully consented for commercial uses (following IG approval)

	Research environment	Access, tooling and environment (once approved)	Requested analysis can be undertaken by internal teams and provided back in anonymised format to data requestors	Users can access the dataset in a Trusted Research Environment		The dataset can be used in a Trusted Research Environment, and other data and tools can be securely brought in as needed
	Access and approvals	Information Governance Process for gaining approval to access data	Contact details for the relevant authority to request data access	Indicative timeframes for processing data access applications detailed	Details of application process and requirements	Detailed Information Governance process described, outlining requirements for applications, basis of decisions, and anticipated timeframes
	Time Lag	Lag between the data being collected and added to the dataset	Approximately 1 year	Approximately 1 month	Approximately 1 week	Effectively real time data
	Timeliness	Average data access request timeframe	More than 12 months	Less than 12 months	Less than 6 months	Less than 3 months
Value & Interest	Linkages	Ability to link with other datasets	Identifiers to demonstrate ability to link to other datasets	Available linkages outlined and/or List of datasets previously successfully linked provided	List of restrictions on the type of linkages detailed. List of previously successful dataset linkages performed, with navigable links to linked datasets via at DOI/URL	Existing linkage with reusable or downstream approvals
	Data Enrichments	Data sources enriched with annotations, image labels, phenomes, derivations, NLP derived data labels	The data include additional derived fields, or enriched data.	The data include additional derived fields, or enriched data used by other available data sources.	The derived fields or enriched data were generated from, or used by, a peer reviewed algorithm.	The data includes derived fields or enriched data from a national report.

Online Supplemental Table 1. Framework as at March 2020

Category	Dimension	Definition
Description	Metadata Completeness	Level of metadata completed
	Metadata Quality	Richness of metadata completion – including within required formats and quality of qualitative fields
Scale	Coverage	Number of individuals included in the dataset
	Duration	Length of time to which the data relates
	Depth	Number of information available per individual (e.g. number of fields)
Service	Format	The presentation of the data – to be presented in useful formats / interoperable standards
	Timeliness	How quickly the data can be provided – in a useful timescale
Quality	Completeness	The proportion of data entries that should be populated are populated (and inverse – proportion that should not be populated are not)
	Consistency / Uniformity	Data are presented in the required format and a similar way – e.g. field types, date formats
	Uniqueness	Lack of duplication
	Validity	Data are valid based on acceptable “rules” e.g. age between 0 and 120, pregnancy in male patients, physiological readings within normal ranges
	Accuracy / Verification	The extent to which the data reflects the “real-world”, e.g. level of certainty that fields are accurate
	“Usefulness”	Qualitative, subjective measure by user (e.g. NPS / star rating)
Added Value	Linkage / Mapping	Ability to link with other datasets
	Annotation	Additional fields added to provide further information, including phenotyping

Online Supplemental Table 2. Interview questions regarding the initial proposed framework

Category	Question
Main focus of the discussion	Can you tell us about an example of a particularly useful or high-quality dataset?
	How did you make use of this dataset?
	Why do you consider that it had these attributes?
	Alternatively, could you recall an experience with an un-useful or low-quality dataset; and the reasons you considered it to be of this nature?
Other points to discuss	As a data user, how do you imagine that a data utility framework/metrics/scores might be able to serve you, support you to make better use of datasets on the innovation gateway?
	Can you describe to us how you might imagine this information working?
	Based on your review of the framework, are there any dimensions that stand out to you as useful/not useful? Why/why not?
	Are there any dimensions that you have questions about, or aren't self-explanatory?
	For your own data needs and use cases, which dimensions would you consider most important/least important?
	For each dimension on the list, how would you rate their importance in terms of understanding data quality
	Are there any other utility or quality dimensions that you would add to this list?
	Data Users in your network: We are extending our engagements, and keen to speak with data users to get their feedback and ideas on data utility. Do you have 3-4 data users in your network you could put us in touch with to interview or survey?
Specific clarification points	**Format**: A particular element that HDR UK are keen to understand is whether your organisation would be able to comply with a requirement to provide data according to either a standard model and format (e.g. OMOP, or a requirement to make data available through a FHIR API) would the organisation be able to do this now? If not, what would be required for you to be able to do this? Is this already on organisational roadmaps? Is it something that could be feasible within a year or two? Or impossible without significant additional investment?
	Coverage: The suggestion here is "Number of individuals included in the dataset". What would be your requirements in terms of quickly understanding coverage (e.g. number of observations, sites etc)?
	Usefulness: In what format would subjective user feedback on the dataset be useful to you? Reviews? Five-star ratings?

	Validation or Transformation: Level of manual "cleaning" and Annotation: Additional fields added to provide further information, including phenotyping: General feedback on data quality often features statements such as the above, which doesn't specify the outcome. What specific indicators would be useful to you in relation to these statements?

Online Supplemental Table 3. Framework as at May 2020

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
Data Documentation	Documentation Quality	Weighted Data Documentation Score	< 66% of metadata specification fields meet format and content requirements	< 76% of metadata specification fields meet format and content requirements	< 86% of metadata specification fields meet format and content requirements	< 96% of metadata specification fields meet format and content requirements
	Availability of additional documentation and support	Available dataset documentation in addition to the data dictionary	Past journal articles demonstrate that knowledge of the data exists	Comprehensive ReadMe describing extracting and use of data, Dataset FAQS available, Visual data model provided	Dataset publication was supported with a journal article explaining the dataset in detail, or dataset training materials	Support personnel available to answer any questions
	Data Model	Availability of clear, documented data model	Known and accepted data model but some key field uncoded or free text	Key fields codified using a local standard	Key fields codified using a national or international standard	Data Model conforms to a national standard and key fields codified using a national / international standard
	Data Dictionary	Provided documented data dictionary and terminologies	Data definitions available	Definitions compiled into local data dictionary which is available online	Dictionary relates to national definitions	Dictionary is based on international standards and includes mapping(?)
	Provenance	Clear description of source and history of the dataset, providing a "transparent data pipeline"	Source of the dataset is documented	Source of the dataset and any transformations, rules and exclusions documented	All original data items listed, all transformations, rules and exclusion listed and impact of these	Ability to view earlier versions, including "raw" or "source" dataset, and review the impact of each stage/step
Technical Quality	Data Quality Management Process	The level of maturity of the data quality management processes	A documented data management plan covering collection, auditing, and management is available for the dataset	Evidence that the data management plan has been implemented is available	Demonstrated compliance with the data management plan	Externally verified compliance with the data management plan
	DAMA Data Quality Element - Completeness	The proportion of stored data against the potential of "100% complete"		*Indicators are under development by another project*		
	DAMA Data Quality Element - Uniqueness	No thing will be recorded more than once based upon how that thing is identified.				

	DAMA Data Quality Element - Timeliness	The degree to which data represent reality from the required point in time.				
	DAMA Data Quality Element - Validity	Data are valid if it conforms to the syntax (format, type, range) of its definition.				
	DAMA Data Quality Element - Accuracy	The degree to which data correctly describes the "real world" object or event being described.				
	DAMA Data Quality Element - Consistency	The absence of difference, when comparing two or more representations of a thing against a definition.				
Coverage	Pathway coverage	Representation of Multi disciplinary healthcare data	Contains data from a single speciality or area	Contains data from multiple specialties or services within a single tier of care	Contains multimodal data or data that is linked across two tiers (e.g. primary and secondary care)	Contains data across the whole pathway of care
	Length of follow up	Average timeframe in which a patient appears in a dataset (follow up period)	Between 1 - 6 months	Between 6 - 12 months	Between 1 - 10 years	More than 10 years
Access & Provision	Allowable uses	Allowable dataset usages as per the licencing agreement		Non-consented, aggregate data for specific academic uses	Aggregate data, for specific commercial use	Fully consented for commercial uses
	Research environment	Access, tooling and environment (once approved)	Requested analysis can be undertaken by internal teams and provided back in anonymised format to data requestors	The dataset can be used in a trusted research environment	The dataset can be used in a trusted research environment, but other data and tools can be brought in as required	The dataset can be used in requesting companies environment
	Format	The technical presentation of the data format (e.g. DICOM images vs PNG)	Format is explicitly defined	Format is explicitly defined and in widely readable (non-proprietary) format		Format is explicitly defined, in open and referenced format

	Access and approvals	IG Process for gaining approval to access data	Contact details for the relevant authority to request data access	Indicative timeframes for processing data access applications detailed	Details of application process and requirements	Detailed Information Governance Process described, outlining requirements for applications, basis of decisions, and anticipated timeframes
	Timeliness	Data time lag + Average data access request timeframe	More than 12 months	Less than 12 months	Less than 6 months	Less than 3 months
Value & Interest	Access Request	Number of data access inquires received	Some interest. Access inquires/requests >1	Some interest. Access inquires/requests >5	Some interest. Access inquires/requests >20	Some interest. Access inquires/requests >50
		Provenance of data access inquires received	Interest and inquires from colleagues and known associates	Interest and inquires from other national organisations	Interests and inquires from national commercial organisations	Interests and inquires from international bodies
	Current Usage	Number of active projects and tools using the dataset	> 1 active projects or tools using the dataset	> 5 active projects or tools using the dataset	> 10 active projects or tools using the dataset	> 30 active projects or tools using the dataset
	Linkages	Ability to link with other datasets	Identifiers to demonstrate ability to link to other datasets	Available linkages outlined and/or List of datasets previously successfully linked provided	List of restrictions on the type of linkages detailed. List of previously successful dataset linkages performed, with navigable links to linked datasets via at DOI/URL	Existing linkage with reusable or downstream approvals
	Data Enrichments	Data sources enriched with annotations, image labels, phenomes, derivations, NLP derived data labels	The data include additional derived fields, or enriched data.	The data include additional derived fields, or enriched data used by other available data sources.	The derived fields or enriched data were generated from, or used by, a peer reviewed algorithm.	The data includes derived fields or enriched data from a national report.

Online Supplemental Table 4. Survey questions with their responses.

<table>
<tr><th>Survey Question</th><th>Response Type/Options</th></tr>
<tr><td>1. Your name (optional)</td><td>[Free text]</td></tr>
<tr><td>2. The organisation where you work (optional)</td><td>[Free text]</td></tr>
<tr><td>3. The industry you work in</td><td>◯ HealthCare Provider
◯ Academia
◯ Technology & Artificial Intelligence
◯ Regulation
◯ Pharmaceutical Industry
◯ Other (please specify) [Free text]</td></tr>
<tr><td>4. How do you make use of health care data in your role?</td><td>[Free text]</td></tr>
<tr><td>5. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
<tr><td>6. Please indicate your thoughts on the importance of each dimension of the data documentation score</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Document Quality</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Data Dictionary</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Additional Documentation and Support</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Data Model</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Provenance</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>7. About the Weighted Data Documentation score:
The weighted data documentation score measures the completeness and quality of the metadata describing the dataset. The score, from 0-100, represent the percentage of metadata fields that are complete and meet the quality requirements. The completeness and quality score for each field of the metadata will be "weighted" based on the comparative utility or usefulness of each of the fields contained in the metadata.
If you are interested in finding out more about the weighted data documentation score and contributing your thoughts to the weightings, please follow the link to this short (approx 3 mins) supplementary survey: https://www.surveymonkey.co.uk/r/HDRUK_Data_Documentation_Score
General thoughts can be left below.</td><td>[Free text]</td></tr>
<tr><td>8. Above are other dimensions considered, but not included in this proposed framework. Please include any thoughts or comments below.</td><td>[Free text]</td></tr>
<tr><td>9. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
</table>

<table>
<tr><td>10. Please indicate the importance of each of the dimensions in the technical quality score</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Data Quality Management Process</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>11. In addition to this measure, another project within HDRUK is looking to develop tools to measure datasets against the six quality measures developed by the DAMA UK Working Group (as per above). These will be released at a later stage and incorporated into the overall measurements associated with the data utility framework.
Please include any thoughts or comments below.</td><td>[Free text]</td></tr>
<tr><td>12. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
<tr><td>13. Please indicate your thoughts on the importance of this dimension</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Pathway Coverage</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Length of Follow up</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>14. Above are other dimensions considered, but not included in this proposed framework. Please include any thoughts or comments below.</td><td>[Free text]</td></tr>
<tr><td>15. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
<tr><td>16. Please indicate your thoughts on the importance of these dimensions</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Allowable uses</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Research Environment</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Format</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Access and Approvals</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Timeliness</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>17. Above are other dimensions considered, but not included in this proposed framework. Please include any thoughts or comments below.</td><td>[Free text]</td></tr>
<tr><td>18. Do you believe these dimensions and rating represented above are useful measures of data utility?
Please leave any comments below.</td><td>[Free text]</td></tr>
</table>

<table>
<tr><td>19. Please indicate your thoughts on the importance of these dimensions</td><td>
<table>
<tr><th></th><th>Useless</th><th>Not Important</th><th>Unsure /Impartial</th><th>Important</th><th>Very Important</th></tr>
<tr><td>Number of Access Requests</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Provenance of Access Requests</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Current Usage</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Linkages</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Data Enrichments</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table>
</td></tr>
<tr><td>20. Any final comments can be added below</td><td>[Free text]</td></tr>
<tr><td>21. OPTIONAL:
If you are interested in receiving feedback on the results of the survey, you can leave your email in the box below. Note: your email address will only be used for the purposes of sending through feedback once available.</td><td>[Free text]</td></tr>
</table>

Online Supplemental Table 5. Framework as at August 2020

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
Data Documentation	Documentation Quality	Weighted Data Documentation Score	< 66% of submitted metadata fields meet format and content requirements of the metadata specification	< 76% of submitted metadata fields meet format and content requirements of the metadata specification	< 86% of submitted metadata fields meet format and content requirements of the metadata specification	< 96% of submitted metadata fields meet format and content requirements of the metadata specification
	Data Model	Availability of clear, documented data model	Known and accepted data model but some key field un-coded or free text	Key fields codified using a local standard	Key fields codified using a national or international standard	Data Model conforms to a national standard and key fields codified using a national / international standard
	Data Dictionary	Provided documented data dictionary and terminologies	Data definitions available	Definitions compiled into local data dictionary which is available online	Dictionary available online, and relates to national definitions	Dictionary is based on international standards and includes mapping
	Provenance	Clear description of source and history of the dataset, providing a "transparent data pipeline"	Source of the dataset is documented	Source of the dataset and any transformations, rules and exclusions documented	All original data items listed, all transformations, rules and exclusion listed and impact of these	Ability to view earlier versions, including "raw" dataset, and review the impact of each stage of data cleaning
Technical Quality	Data Quality Management Process	The level of maturity of the data quality management processes	A documented data management plan covering collection, auditing, and management is available for the dataset	Evidence that the data management plan has been implemented is available		Externally verified compliance with the data management plan
	DAMA Data Quality Element - Completeness	The proportion of stored data against the potential of "100% complete"	*Set initial figures for this*			
	DAMA Data Quality Element - Validity	Data are valid if it conforms to the syntax (format, type, range) of its definition.				
Coverage	Pathway coverage	Representation of multi-disciplinary healthcare data	Contains data from a single speciality or area	Contains data from multiple specialties or services within a single tier of care	Contains multimodal data or data that is linked across two tiers (e.g. primary and secondary care)	Contains data across the whole pathway of care
	Length of follow up	Average timeframe in which a patient appears in a dataset (follow up period)	Between 1 - 6 months	Between 6 - 12 months	Between 1 - 10 years	More than 10 years
Access & Provision	Allowable uses	Allowable dataset usages as per the licencing agreement		Non-consented, aggregate data for specific academic uses (following IG approval)	Aggregate data, for academic and specific commercial uses (following IG approval)	Fully consented for commercial uses (following IG approval)

	Research environment	Access, tooling and environment (once approved)	Requested analysis can be undertaken by internal teams and provided back in anonymised format to data requestors	Users can access the dataset in a Trusted Research Environment		The dataset can be used in a Trusted Research Environment, and other data and tools can be securely brought in as needed
	Access and approvals	Information Governance Process for gaining approval to access data	Contact details for the relevant authority to request data access	Indicative timeframes for processing data access applications detailed	Details of application process and requirements	Detailed Information Governance process described, outlining requirements for applications, basis of decisions, and anticipated timeframes
	Time Lag	Lag between the data being collected and added to the dataset	Approximately 1 year	Approximately 1 month	Approximately 1 week	Effectively real time data
	Timeliness	Average data access request timeframe	More than 12 months	Less than 12 months	Less than 6 months	Less than 3 months
Value & Interest	Linkages	Ability to link with other datasets	Identifiers to demonstrate ability to link to other datasets	Available linkages outlined and/or List of datasets previously successfully linked provided	List of restrictions on the type of linkages detailed. List of previously successful dataset linkages performed, with navigable links to linked datasets via at DOI/URL	Existing linkage with reusable or downstream approvals
	Data Enrichments	Data sources enriched with annotations, image labels, phenomes, derivations, NLP derived data labels	The data include additional derived fields, or enriched data.	The data include additional derived fields, or enriched data used by other available data sources.	The derived fields or enriched data were generated from, or used by, a peer reviewed algorithm.	The data includes derived fields or enriched data from a national report.